Deconstructing Review Deception: A Study on Counterfactual Explanation and XAI in Detecting Fake and GPT-Generated Reviews

Olga Chernyaeva
Pusan National University
misslelka@pusan.ac.kr

Taeho Hong
Pusan National University
hongth@pusan.ac.kr

One-Ki Daniel Lee
University of Massachusetts Boston
daniel.lee@umb.edu

Abstract

Our models not only deliver high-performing predictions but also illuminate the decision-making processes underlying these predictions. By experimenting with five datasets, we have showcased our framework's prowess in generating diverse and specific counterfactuals, thereby enhancing deception detection capabilities and supporting review authenticity assessments. The results demonstrate the significant contribution of our research in furthering the understanding of AI-generated review detection and, more broadly, AI interpretability. Experimentation on five datasets reveals our framework's ability to produce diverse and specific counterfactuals, significantly enriching deception detection capabilities and facilitating the evaluation of review authenticity. Our robust model offers a novel contribution to the understanding of AI applications, marking a significant step forward in both the detection of deceptive reviews and the broader field of AI interpretability.

Keywords: Fake review detection, XAI, Counterfactual explanation, generated reviews, GPT.

1. Introduction

Online product reviews, representing a form of electronic word-of-mouth (eWOM), play a pivotal role in shaping consumer purchasing decisions (Tran and Strutton, 2020). According to a survey, more than 80% of American consumers consult online reviews before purchasing (Smith and Anderson, 2016). This significant influence has created a lucrative opportunity for producing fraudulent or manipulated reviews with the intent of promoting products or services or tarnishing competitors' reputations. According to the study by Salminen et al, 2022, fake reviews can be primarily produced in two ways: (a) through human generation, where individuals are compensated to craft seemingly authentic but misleading reviews about products they have never encountered, and (b) through computer generation, where text-generation algorithms are employed to automate the creation of fraudulent reviews.

In the past, human-generated fake reviews were commodified in a "market of fakes" (He et al., 2022) where one could order reviews online, and human authors would execute the task. However, advancements in natural language processing (NLP) and machine learning (ML) have spurred the automation of fake review generation. The recent advancements in language models, particularly OpenAI's Generative Pre-trained Transformer (GPT) series, have stirred significant academic and industry interest due to their unprecedented capabilities in generating highly coherent and contextually accurate text (Brown et al., 2020). Given its proficiency in mimicking human-like text generation, GPT offers both opportunities and challenges for review platforms, as it could be potentially leveraged to generate indistinguishable fake reviews, thereby complicating the detection landscape (Wang et al., 2022). Utilizing generative language models, fraudulent reviews can now be produced at a substantial scale and at a fraction of the cost compared to human-generated fake reviews. Prior research primarily focused on discerning fake reviews, leaving a conspicuous absence in studies addressing the interpretability of detection models and their application to AI-generated reviews (Liu and Lang, 2019). For instance, a model's prediction of a review as fake lacks value without a sufficient explanation for its decision. This is especially relevant in AI-generated reviews where the model's complexity often obscures the decision-making process.

While AI's increasing role in review generation is undeniable, research exploring counterfactual explanations to illuminate detection model results remains scarce. Previous studies have made significant strides toward fake review detection (Ott et al., 2013; Yu et al., 2022), however, the interpretability of these models is not sufficiently addressed. Specifically, counterfactual explanations, instrumental for understanding the precise conditions influencing model predictions, remain under-examined (Wachter et al., 2017). This illustrates a substantial research gap

URI: https://hdl.handle.net/10125/106431
978-0-9981331-7-1

concerning the application of explainable AI (XAI) and counterfactual explanations in fake review detection.

Addressing the identified research gaps, our study seeks to comprehensively detect and interpret both fake and GPT-generated reviews. This objective is pursued via a four-stage methodology. Initially, in Phase 1, we collected both genuine and deceptive reviews of premier New York restaurants from Yelp.com, an influential online review platform. Subsequently, Phase 2 utilized OpenAI's GPT series to synthesize three categories of reviews: Real-based Generated Reviews (RbGR), Fake-based Generated Reviews (FbGR), and User-guided Generated Reviews (UGGR). In Phase 3, six diverse models, specifically Logistic Regression, Decision Tree, Random Forest, XGBoost, Artificial Neural Network (ANN), and Support Vector Machine (SVM), are applied to discern fake and fabricated reviews across five separate datasets. Ultimately, Phase 4 endeavors to illuminate the detection outcomes using explainable AI (XAI) techniques, notably SHAP, supplemented by counterfactual explanations.

This study signifies a crucial stride in the sustained effort against deceptive reviews, introducing an innovative strategy that amalgamates detection, interpretation, and counterfactual explanation. Academically, our research constitutes a substantial addition to the extant literature, specifically within the XAI realm. Practically, our study furnishes an efficient framework for platforms and enterprises to promptly identify fake reviews, thereby bolstering transparency and trust. Additionally, the methodologies and insights derived from this study establish a robust groundwork for future explorations into AI interpretability across various disciplines

2. Literature Review

2.1. AI Text Generation: Ethical Landscape

The domain of text generation has observed significant advancements with the advent of machine learning, particularly deep learning methodologies. The purpose of text generation systems is to create coherent and contextually appropriate text that matches human-level fluency and relevance. Techniques applied in text generation typically fall under either template-based, retrieval-based, or generative models, the latter of which has gained substantial attention due to its ability to produce diverse and novel outputs (Gao et al., 2019). Generative Pre-training (GPT) models, introduced by OpenAI, mark a significant milestone in text generation, particularly in terms of their language understanding capabilities and the diversity of the generated text (Radford et al., 2018).

However, the rise of AI agents for Automated Text Generation (ATG) brings up new ethical challenges. These AI agents can generate large volumes of high-quality content that sounds very human-like, and their usage is on the rise. Traditional automated bots are often easy to spot as non-human (Floridi and Chiriatti, 2020), but there have been instances where users on platforms like Reddit took more than a week to realize they were interacting with an AI agent (Heaven, 2020). This AI-powered text generation can be misused to spread disinformation, which means creating fake news, reviews, letters, or impersonating others online. Disinformation means making up information to mislead or present a biased view of something (Tandoc et al., 2018). This spreading of disinformation to manipulate how the public thinks is becoming more common and is a threat to businesses across different industries (Petratos, 2021). As a result, Illia et al. (2023) have outlined new ethical issues arising from the use of AI agents for automated text generation. These include AI agents being used for mass manipulation and spreading disinformation, AI generating low-quality but believable content, and a decrease in direct communication between human. Since manipulated text generation is more common in areas like review generation in e-commerce, it's crucial to figure out how to detect the content generated by AI. But before that, it's important to understand what fake reviews are and how we detect them.

2.2. Fake Reviews Detection

The proliferation of online user-generated content has been paralleled by a corresponding rise in fake reviews, thereby presenting significant challenges alongside new research avenues. Fake reviews have achieved attention as an academic inquiry in terms of their impact on consumer behavior and trust, their defining characteristics, and their detection employing various machine learning methodologies (Liu and Lang, 2019; Mohawesh et al., 2021). Mohawesh et al. (2021) made a substantial contribution to understanding the concept drift in fake review detection and illustrating the temporal shifts in the patterns of these reviews. The primary methodologies for fake review detection draw on the linguistic traits of reviews (through Natural Language Processing or NLP), the reviewer's attributes, or a hybrid of both approaches. Researchers concur that there are no specific words that can reliably distinguish a fake review from a truthful one (He et al., 2022), and hence, the focus is often on a multitude of micro-linguistic characteristics. For instance, Shojaee et al. (2013) utilized syntactic and lexical features to differentiate between authentic and fake reviews. To enhance the performance of the detection algorithm, particularly with respect to accurately identifying deceptive reviews, researchers frequently advocate for the use of an amalgamation of features drawn from the text and context of the review (Ott et al., 2011; Li et al., 2014c).

Further, psycholinguistic analyses have been deployed using the Linguistic Inquiry and Word Count (LIWC) software (Ott et al., 2011; Li et al., 2014c), and tools such as Coh-Metrix have been used to assess review texts on dimensions such as cohesion, language, and readability (Plotkina et al., 2020). Banerjee et al. (2017) identified specific linguistic traits indicative of the cognitive challenges associated with crafting fake reviews such as negligence. Essential textual features highlighted include self-referencing, uncertainty cues, specific fillers, and tentative words. Li et al. (2014c) and Yoo and Gretzel (2009) observed that fraudulent reviews tend to mention brand names or product names excessively, often written from a second-person perspective with minimal use of self-reference.

After the rising AI text generation, previous studies applied AI text generation to achieve various purposes, including the generation of fake reviews, content augmentation, and automated customer service responses. Salminen et al. (2022) utilized the GPT-2 model, a significant advancement in text-generation techniques at that time, in the experiment of creating an automated detection model. Chernyaeva and Hong (2022) applied GPT-3 to generate fake reviews and solve the imbalanced problem of the review dataset. Both studies, however, excluded reviews shorter than five words. This exclusion raises questions about the detection system's efficacy on brief reviews, which are commonplace in online review platforms, thereby suggesting that the performance and generalizability of the model on short reviews necessitate further scrutiny. Another limitation pertaining to the research suffers from a lack of a comprehensive explanation of the results. While machine learning techniques have been applied, a detailed interpretation of the findings is not provided. This omission inhibits the study's potential to contribute to the growing call for more transparent and interpretable machine-learning outcomes (Doshi-Velez and Kim, 2017).

2.3. XAI and Counterfactual Explanation

Explainable Artificial Intelligence (XAI) has emerged as a pivotal field of research aiming to create transparency in AI model predictions, thereby improving their interpretability. In the context of text classification, interpretability involves understanding why certain texts are classified into specific categories by the model. This understanding aids in uncovering the model's decision-making process and helps in trust-building among its users (Adadi and Berrada, 2018). One of the prevalent tools for achieving explainability in AI is the Shapley Additive exPlanations (SHAP) framework. It is a unified approach to interpreting the output of any machine learning model and uses the game-theoretic concept of Shapley values to distribute the contribution of each feature to the prediction for each instance (Lundberg and Lee, 2017). SHAP has been utilized in several studies to provide transparency in text classification. For instance, Weitz et al. (2021) explored the potential of virtual agents in XAI designs on the perceived trust of end-users. They conducted a user study based on a simple speech recognition system for keyword classification and found that the integration of virtual agents led to increased user trust in the XAI system.

Feature attribution methods (e.g., SHAP) and counterfactual explanations aim to provide interpretability to AI model predictions, however through different means (Wang and Chen, 2023). SHAP distributes the contribution of each feature to the prediction for each instance, thereby explaining the output of the model. On the other hand, counterfactual explanations provide insights into the model's decision-making process by showing how the outcome would change if the input were different. Diverse Counterfactual Explanations (DiCE) is a framework for generating and evaluating a diverse set of counterfactual explanations, which are hypothetical examples that show how to obtain a different prediction from a machine learning model. DiCE is based on determinantal point processes and is designed to generate counterfactuals that are diverse and approximate local decision boundaries well, outperforming prior approaches to generating diverse counterfactuals (Mothilal et al., 2019). The existing research on fake and generated reviews detection reveals a notable gap: the unexplored application of both SHAP and DICE methodologies. Both methods have the potential to significantly enhance model interpretability and transparency. Our research intends to bridge this gap, pioneering the application of both SHAP and DICE in this domain.

3. Research Framework and Analysis

As shown in Figure 1, our research framework comprises four stages: data collection, review generation, detection of fake and generated reviews, and explanation of the models.

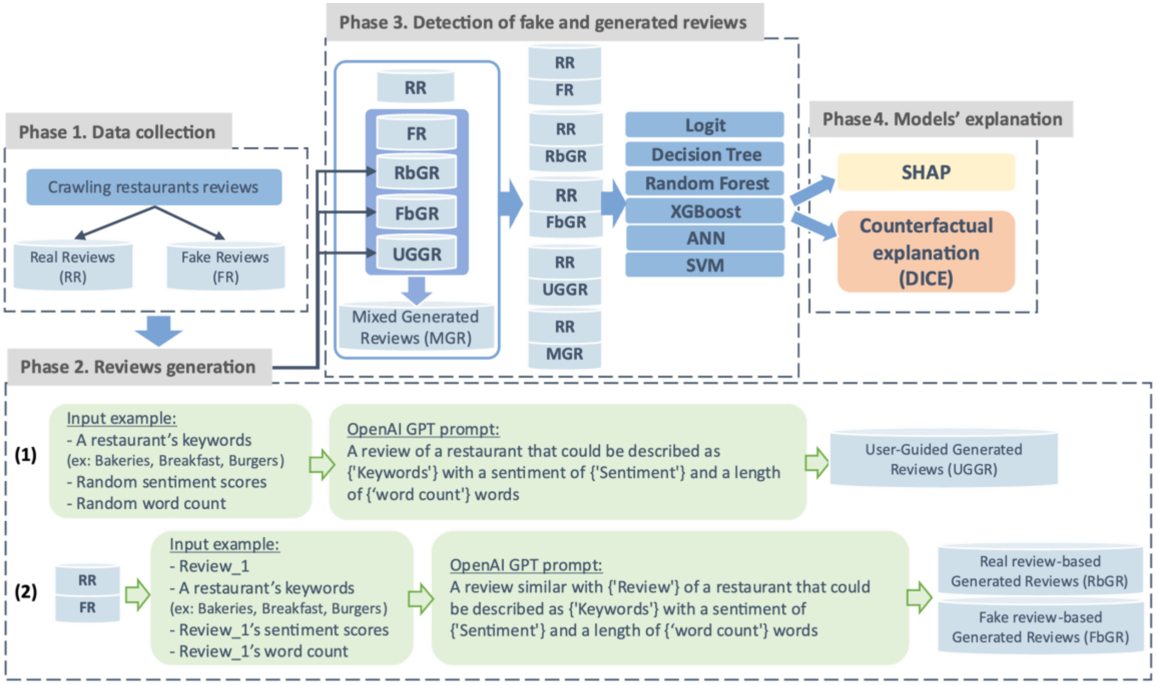
Phase 3. Detection of fake and generated reviews
RR
FR
RbGR
FbGR
UGGR
Mixed Generated Reviews (MGR)
RR
FR
RR
RbGR
RR
FbGR
RR
UGGR
RR
MGR
Logit
Decision Tree
Random Forest
XGBoost
ANN
SVM
Phase4. Models' explanation
SHAP
Counterfactual explanation (DICE)
Phase 1. Data collection
Crawling restaurants reviews
Real Reviews (RR)
Fake Reviews (FR)
Phase 2. Reviews generation
(1)
Input example:
- A restaurant's keywords
(ex: Bakeries, Breakfast, Burgers)
- Random sentiment scores
- Random word count
OpenAI GPT prompt:
A review of a restaurant that could be described as {'Keywords'} with a sentiment of {'Sentiment'} and a length of {'word count'} words
User-Guided Generated Reviews (UGGR)
(2)
RR
FR
Input example:
- Review_1
- A restaurant's keywords
(ex: Bakeries, Breakfast, Burgers)
- Review_1's sentiment scores
- Review_1's word count
OpenAI GPT prompt:
A review similar with {'Review'} of a restaurant that could be described as {'Keywords'} with a sentiment of {'Sentiment'} and a length of {'word count'} words
Real review-based Generated Reviews (RbGR)
Fake review-based Generated Reviews (FbGR)

Figure 1. The research framework of fake and generated review detection

Phase 1. Data Collection: We procured both authentic and fake reviews of top restaurants in New York from Yelp.com. Yelp's proprietary automated recommendation software filters reviews (recommended and non-recommended) based on detailed reviewer information exclusive to platform owners such as IP addresses, posting frequency, geolocation, and additional seller information (Salminen et al., 2022). Consequently, we classified the recommended reviews as 'real' and the non-recommended ones as 'fake.' For our experiment, we collected reviews from the top 10 restaurants in New York, yielding 68,644 real reviews and 10,567 fake ones. To construct an accurate model, we performed pair-matching on our dataset, resulting in a final dataset of 10,000 real and 10,000 fake reviews.

Phase 2. Review Generation: To generate reviews we applied two approaches: 1) synthetic review generation, which constructs fake reviews based on pre-existing real reviews (Salminen et al., 2022; Crawford et al., 2015); 2) user-guided approach, which is based on prompt engineering of the user. The first approach contrasts with the traditional methods, which distinguish fake and real reviews and label them accordingly (Jindal and Liu, 2008) or employ human authors for the purpose of crafting counterfeit content (Ott et al., 2011). The second user-guided approach applied prompt engineering, a technique that emphasizes the prominence of prompt engineering in the realm of linguistic model interaction. White et al. (2023) have emphasized the growing importance of this approach, especially when dealing with sophisticated models such as ChatGPT. Within this framework, prompts are delineated as specific directives bestowed upon expansive language models. In the context of generating reviews, prompts serve as directives provided to expansive language models, directing them to establish regulations, streamline operations, and guarantee the presence of specific attributes in the generated content. To facilitate the generation of reviews, we incorporated prompts with input components such as restaurant-related keywords, intended sentiments for the review, and review lengths. Applying both approaches is deemed important, as fake review creators are all the time searching for more efficient ways to generate fake reviews at scale, with minimum human involvement (Cheng and Ho, 2015).

We used OpenAI's GPT-3.5 to generate reviews via two different methods: User-Guided Review Generation (UGGR), where the input features in the generation prompt included the keyword of the top-10 restaurants in New York (provided in the Yelp.com restaurant page), random sentiment scores ranging from -1 to 1, and a random review length ranging from 1 to 936 words (matching the maximum review length in the total review dataset); and Existent Review-Based Generation, which involved creating two generated

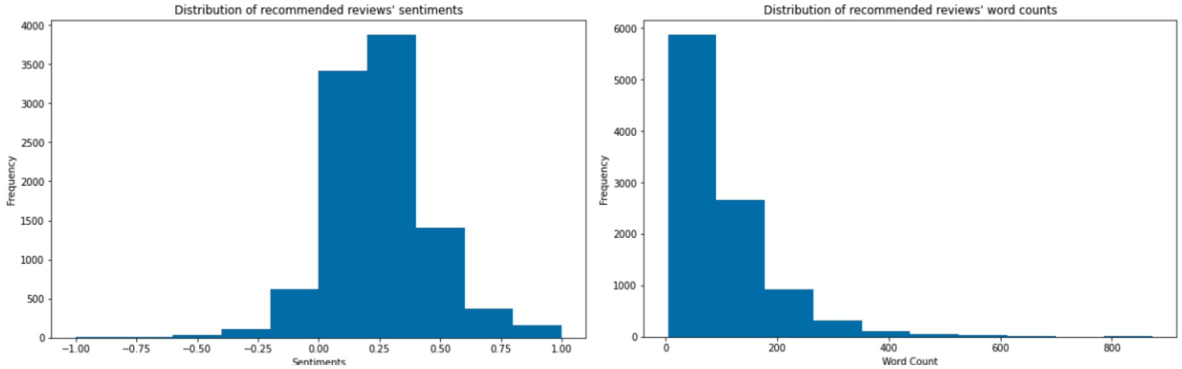

Figure 2. The distribution of sentiment scores and word count of real reviews

review datasets, each with 10,000 reviews: Real Review-based Generated Reviews (RbGR) and Fake Review-based Generated Reviews (FbGR). RbGR was designed to develop a detection model capable of identifying generated reviews that closely resemble real ones, while FbGR aimed to develop a model capable of detecting generated reviews that retained unique features of the original manipulated reviews. The input features in the generation prompt included a review (real or fake), the keywords of the review's restaurant, the sentiment scores of the review, and the review length. The distribution of sentiment scores and word count of real reviews is shown in Figure 2.

Phase 3. Detection of Fake and Generated Reviews: The third phase of our research is concerned with the detection of fake and generated reviews. To begin, a novel dataset, dubbed Mixed Generated Reviews (MGR), was assembled. The MGR comprises a balanced assortment of 2,500 each of fake reviews, User-Guided Generated Reviews (UGGR), Real Review-based Generated Reviews (RbGR), and Fake Review-based Generated Reviews (FbGR). The compilation of this diverse dataset is critical as it provides a rich assortment of review types, enhancing the robustness of the subsequent detection process. To perform a detailed linguistic analysis of each dataset, we employed the linguistic inquiry and word count (LIWC-22), an extensively validated tool for psychometric text analysis. This approach enabled us to extract 120 distinct and explainable numerical features from the review text within each dataset, thus providing a rich feature set for further analysis (Pennebaker et al., 2015). As such, our study incorporated a sophisticated linguistic lens better to understand the nuances of fake and generated reviews, further bolstering the robustness of our analysis and results. For the detection, six varied models were chosen for their proven effectiveness in classification tasks - Logistic Regression, Decision Tree, Random Forest, XGBoost, Artificial Neural Network (ANN), and Support Vector Machine (SVM). Each of these models was independently applied to detect fake and generated reviews across the five separate datasets. The performance of each model was evaluated using standard metrics like accuracy, precision, recall, and F1-score, ensuring a comprehensive appraisal of the model's effectiveness in fake and generated review detection.

Phase 4. Model Explanation: The final phase of our research addresses model interpretability, a critical aspect in the application of machine learning, particularly in contexts requiring transparency and trust. The models' decision-making processes were elucidated using Explainable AI (XAI) techniques, with a primary focus on SHapley Additive exPlanations (SHAP) values. SHAP values provide a measure of the importance of each feature in making a prediction, thereby providing a clearer picture of what factors the model is considering in distinguishing between fake and generated reviews from real ones. In addition to SHAP, we also employed counterfactual explanations to enhance the interpretability of the models. Counterfactual explanations provide insights into how the outcome of a prediction could change with different input values, thus offering a more intuitive understanding of the model's behavior (Delaney et al., 2023). The combination of SHAP and counterfactual explanations strengthens the interpretability of our models, leading to more robust and trustworthy outcomes (Fernández-Loría et al., 2022).

4. Results

The results from the detection of fake and generated reviews are detailed in Table 1. To offer a holistic understanding of the performance of the detection models, we report the evaluation metrics Accuracy, Precision, Recall, and F1 score. Each of these metrics provides unique insights into the model's performance in the context of fake and generated reviews detection.

- *Accuracy* reflects the proportion of total predictions that are correct.
- *Precision* is the proportion of positive predictions that are actually correct.
- *Recall* is the proportion of actual positive cases that are correctly identified.
- The *F1* score is the harmonic mean of Precision and Recall and provides a balance between the two.

For example, for Dataset 1, which detected fake reviews, the Artificial Neural Network model demonstrated superior performance in detecting fake reviews, achieving an Accuracy of 0.763, Precision of 0.791, Recall of 0.728, and F1 score of 0.757. This suggests that the ANN model effectively discerned between real and fake reviews, with most of its predictions being both accurate and precise.

In Phase 4's results, we provide XAI and a counterfactual explanation of Phase 3 detection. The results of SHAP (shown in Table 2) offer a quantitative illustration of how individual features contribute to the predictive output. Positive SHAP values suggest a feature that propels the model's prediction upwards (toward class '1'), whereas negative SHAP values are indicative of a feature that lowers the model's prediction (toward class '0').

For instance, in the context of Dataset 1, with the Artificial Neural Network (ANN) model as a case study, a few salient features appear to play a critical role in the classification task. The 'Word Count (WC)' feature exhibits the highest mean absolute SHAP value, signifying its pronounced influence on the model's decision-making. Additional influential features are the presence of an apostrophe and the positive tone of the text. A comprehensive exploration of all linguistic features is given in Table 3. However, SHAP results do not explicitly elaborate on the differences between fake and real reviews. To bridge this gap, we provide the analysis with counterfactual explanations via DICE, with the results presented in Tables 4,5,6,7,8.

Table 1. Results of Detection of Fake and Generated Reviews (test set)

		RR&FR (1)	RR&RbGR (2)	RR&FbGR (3)	RR&UGGR (4)	RR&MGR (5)
Logit	**Accuracy**	0.735	0.74	0.846	0.981	0.771
	Precision	0.744	0.796	0.846	0.978	0.778
	Recall	0.717	0.69	0.833	0.984	0.746
	F1	0.73	0.732	0.834	0.981	0.762
Decision Trees	**Accuracy**	0.692	0.737	0.785	0.94	0.715
	Precision	0.691	0.736	0.765	0.936	0.701
	Recall	0.695	0.746	0.801	0.944	0.73
	F1	0.693	0.741	0.782	0.94	0.715
Random Forest	**Accuracy**	0.761	0.837	0.865	0.978	0.808
	Precision	0.786	0.902	0.879	0.986	0.847
	Recall	0.716	0.76	0.837	0.97	0.744
	F1	0.75	0.825	0.857	0.979	0.792
XGBoost	**Accuracy**	0.751	0.837	**0.871**	**0.986**	**0.808**
	Precision	0.772	0.881	**0.888**	**0.987**	**0.858**
	Recall	0.712	0.784	**0.840**	**0.985**	**0.744**
	F1	0.741	0.829	**0.863**	**0.986**	**0.8**
Artificial Neural Networks	**Accuracy**	**0.763**	**0.845**	0.853	0.985	0.803
	Precision	**0.791**	**0.919**	0.847	0.983	0.85
	Recall	**0.728**	**0.77**	0.85	0.986	0.727
	F1	**0.757**	**0.842**	0.848	0.985	0.784
Support Vector Machines	**Accuracy**	0.738	0.842	0.849	0.981	0.778
	Precision	0.729	0.901	0.859	0.977	0.814
	Recall	0.756	0.771	0.822	0.984	0.709
	F1	0.742	0.831	0.84	0.981	0.758

Table 2. The Results of SHAP

	FR	RbGR	FbGR	UGGR	MGR
Top 5 features (mean SHAP value)	WC (0.06)	Comma (0.07)	Period (0.08)	article (0.04)	Comma (0.04)
	Apostro (0.03)	Period (0.06)	Comma (0.07)	time (0.04)	Period (0.04)
	tone_pos (0.03)	WPS (0.05)	WPS (0.06)	leisure (0.03)	AllPunc (0.04)
	Tone (0.03)	focuspresent (0.05)	AllPunc (0.06)	function (0.03)	Authentic (0.03)
	Dic (0.03)	article (0.04)	focuspresent (0.05)	auxverb (0.03)	space (0.03)

Table 3. A comprehensive exploration of linguistic features (from LIWC)

Feature name	Description/Most frequently used exemplars
Affect	Affect/ good, well, new, love
allnone	All-or-none/ all, no, never, always
AllPunc	All punctuation
Apostro	Apostrophes
attention	Attention/look, look for, watch, check
Authentic	Perceived honesty, genuineness
Comma	Comma
Conv	Conversational/yeah, oh, yes, okay
Det	Determiners/ the, at, that, my
Dic	Dictionary words
emo_pos	Positive emotion/good, love, happy, hope
Exclam	Exclamation points
focuspresent	Present focus /is, are, I'm, can
function	Total function words / the, to, and, I
leisure	Leisure/ game, fun, play, party
number	Numbers/one, two, first, once
Period	Periods (punctuation)
polite	Politeness/thank, please, thanks, good morning
ppron	Personal pronouns/I, you, my, me
	Prepositions/ to, of, in, for
Social	Social processes/you, we, he, she
tone	Emotional tone / Degree of positive (negative) tone
they	3rd person plural
time	Time/when, now, then, day
WC	Total word count
WPS	Average words per sentence

Table 4. Counterfactual explanations of fake review detection

	Modified features				
	WPS	Dic	Affect	polite	Conv
Orig. instance: FR (1)	5.5	100.0	18.1	0.0	0.0
Counterfactuals: RR (0)	687.0	85.9	38.7	27.9	16.9
	687.0	85.9	38.7	27.9	16.9
	687.0	85.9	38.7	27.9	16.9

Table 5. Counterfactual explanations of RbGR detection

	Modified features				
	emo_pos	attention	Exclam	number	Social
Orig.instance: RbGR (1)	4.35	0.0	4.35	0.0	4.35
Counterfactuals: RR (0)	72.8	-	-	-	-
	-	14.4	48.5	-	-
	-	-	-	37.2	37.2

Table 6. Counterfactual explanations of FbGR detection

	Modified features				
	WPS	Dic	ppron	number	Social
Orig. instance: FbGR (1)	6.0	100.0	11.11	0.0	0.0
Counterfactuals: RR (0)	132.2	16.8	10.1	-	-
	-	-	-	89.1	1.3
	-	-	-	89.1	1.3

Table 7. Counterfactual explanations of UGGR detection

	Modified features				
	allnone	Exclam	they	Comma	number
Orig. instance: UGGR (1)	2.5	0.0	0.0	7.5	0.0
Counterfactuals: RR (0)	5.4	149.3	-	-	-
	5.4	149.3	-	-	-
	-	137.7	11.0	21.1	15.9

Table 8. Counterfactual explanations of MGR detection

	Modified features				
	WC	Exclam	Period	time	AllPunc
Orig. instance: MGR (1)	36	0.0	8.33	0.0	22.22
Counterfactuals: RR (0)	219.0	94.9	-	-	-
	-	-	179.4	26.9	412.5
	-	-	179.4	-	338.1

Table 9 summarizes what we found from the above results regarding the distinct ways in which fake and generated reviews differ from real reviews.

Table 9. The Interpretation of counterfactual explanations of fake and generated review detection

	Interpretation of counterfactual explanations
Fake review detection	(1) fake reviews have fewer words per sentence than real reviews; (2) real reviews use fewer dictionary words than fake reviews; (3) fake reviews use less affect words such as 'good, well, new, love, etc' than real reviews; (4) real reviews are more polite than fake reviews; (5) real review uses more conversational words such as 'yeah, oh, yes, okay' than a fake one.
RbGR detection	(1) RbGR are less emotionally positive than some real reviews; (2) RbGR utilize fewer attention-drawing words such as 'look, look for, watch, check' than some real reviews; (3) RbGR use fewer exclamatory phrases than real; (4) RbGR contain fewer numerical representations than real reviews; (5) RbGR employ fewer words associated with social processes (e.g., 'you', 'we', 'he', 'she') than real reviews;

FbGR detection	(1) FbGR has fewer words per sentence than a real reviews; (2) FbGR use more dictionary words than a real reviews; (3) FbGR use more personal pronouns words such as 'I, you, my, me' than real reviews; (4) FbGR use fewer numbers than real reviews; (5) FbGR use fewer words related to social processes such as 'you, we, he, she' than real reviews.
UGGR detection	(1) UGGR incorporates more all-or-none words, such as 'all, no, never, always' compared to a real review; (2) UGGR employs fewer exclamation points compared to a real review; (3) UGGR utilizes fewer instances of 3rd person plural words compared to real reviews; (4) UGGR uses fewer commas than real reviews; (5) UGGR contains fewer numerical representations than real reviews.
MGR detection	(1) MGR shorter than a real reviews; (2) MGR employs fewer exclamation points compared to a real reviews; (3) MGR have fewer periods than real reviews; (4) MGR uses fewer words related to time than real reviews; (5) MGR contains fewer punctuations than real reviews.

5. Conclusion and Discussion

In conclusion, the experimental results showcased the varying efficiencies of different models in detecting fake and generated reviews across multiple datasets. Among the models, ANN demonstrated proficiency in detecting fake reviews and real review-based generated reviews, while XGBoost exhibited superior performance in recognizing fake review-based generated reviews, user-guided generated reviews, and mixed generated reviews. These findings elucidate the critical role of model selection, reinforcing that different algorithms may excel under diverse scenarios, thereby requiring tailored applications based on the specific context and requirements. Further, the study underscores the efficacy of contemporary machine learning algorithms such as ANN and XGBoost in advancing the detection and classification of fake and various kinds of generated reviews. This research, therefore, contributes significantly to the domain of fake review detection, providing valuable insights for future works and practical implications for online platforms striving to enhance the integrity and reliability of user-generated reviews.

The findings presented in Phase 4 provide a valuable integration of Explainable AI (XAI) and counterfactual explanations, casting light on the performance of Phase 3 detection mechanisms. Using SHAP, we have quantitatively gauged the degree of contribution from individual features to the predictive output, thereby unearthing the distinguishing elements between fake and real reviews across different datasets. In the case of Dataset 1, our analysis demonstrated that the Word Count feature, the existence of an apostrophe, and the positive tone were crucial factors. Notably, though, SHAP values alone do not suffice to demarcate the difference between fake and real reviews explicitly. Consequently, we utilized DICE to present a counterfactual explanation, identifying five distinctive features that set fake reviews apart from their real counterparts. This dual methodology—SHAP for quantification and DICE for detailed explanations—was similarly effective when applied to Dataset 2 and Dataset 4. For instance, in Dataset 2, while the presence of commas, periods, and the count of words per sentence emerged as significant features from SHAP analysis, DICE underscored the differences between real review-based generated reviews and real reviews, including factors such as emotional positivity and the usage of attention-drawing words, exclamation phrases, numerical representations, and socially-related words. Moreover, in the case of Dataset 4, while the SHAP method indicated that the presence of the article was the most consequential feature for User-Guided Generated Reviews, the counterfactual explanations through DICE outlined five key distinctions between UGGR and real reviews, further enriching our understanding of linguistic variances.

Finally, a similar approach was implemented for Dataset 5, focusing on the prediction results of mixed-generated reviews. Here, SHAP values emphasized the role of features like commas, periods, and various punctuations, whereas DICE revealed nuances such as the shorter length of mixed-generated reviews and their reduced use of exclamation points, periods, time-related words, and punctuations. Overall, the integration of XAI and counterfactual explanation strategies has elucidated the intricate differences between real and various types of generated reviews, underscoring the complexity of the classification task. Our findings thereby contribute to the ongoing conversation on reliable fake review detection, demonstrating the value of an interpretative, feature-focused approach in enhancing the sophistication and accuracy of predictive models.

This research significantly contributes to the academic field of fake review detection, particularly in the context of generated reviews. A key aspect of this contribution is the extension of existing literature to include the detection of generated reviews. With the increasing popularity of open chat AI and GPT technologies, the generation of reviews is becoming more prevalent. This shift in focus addresses a new and emerging challenge in the field of fake review detection, thereby expanding the scope of academic discourse in this area. The study further enhances the understanding of fake review detection by providing comprehensive insights into the workings of the detection models. Unlike previous studies that often present the results of their fake review detection models without detailed explanations, this research employs two popular methods, SHAP (SHapley Additive exPlanations) and DICE (DIverse Counterfactual Explanations), to elucidate how individual features contribute to the predictive output of the models. This approach not only deepens the understanding of the detection models but also contributes to the development of more reliable and robust models for fake review detection.

Finally, the research provides a benchmark for future studies in the field. The experimental results demonstrate the effectiveness of contemporary machine learning algorithms in classifying and detecting fake and different types of generated reviews, providing a valuable reference point for future studies aiming to develop and improve fake review detection models. Additionally, by integrating Explainable AI (XAI) into the study, it provides a pathway for other researchers to understand and interpret the black-box models commonly used in fake review detection. This contribution is significant as it encourages transparency and trust in machine learning models, a growing concern in the field of artificial intelligence. In conclusion, this research not only advances academic understanding in the field of fake review detection but also paves the way for future studies by providing a comprehensive approach to understanding and explaining the results of fake and generated review detection models.

From a practical perspective, the research outcomes can have significant implications for various stakeholders in online platforms, particularly those reliant on user-generated reviews. Understanding the nuances of fake and generated reviews can help develop more robust and efficient algorithms for their detection, thereby enhancing the credibility of online platforms. The findings offer valuable insights to platform administrators, suggesting tailored application of detection models based on specific context and requirements. Moreover, the specific features identified as critical in distinguishing between real and generated reviews can guide the design of detection algorithms and moderation tools. By incorporating these insights, the detection tools can be made more accurate and reliable. The research, therefore, not only contributes to the academic discourse on fake review detection but also provides practical implications for improving the integrity and reliability of user-generated reviews on online platforms.

The limitations of this study provide avenues for further research. Firstly, the exclusive reliance on ANN and XGBoost models for detecting fake and generated reviews may impede the comprehensive applicability of our findings. Future investigations would benefit from incorporating a wider range of machine learning models such as BERT to optimize detection precision and establish a more distinct connection between theory and methodology of fake and generated reviews detection. Secondly, by focusing on restaurant reviews from a singular dataset, the study might not encapsulate the full spectrum of linguistic intricacies characterizing fake and real reviews across various platforms or sectors. Expanding to more diverse datasets in future research could ensure a more holistic understanding. Thirdly, the study's dependence on Yelp's recommender system for classifying reviews as 'real' or 'fake' poses a potential limitation. Such a method might inadvertently perpetuate biases inherent to Yelp's system. A more independent and objective mechanism for review classification would be advantageous for subsequent studies. Lastly, given the enigmatic nature of deep learning models, the results derived from interpretability tools like SHAP and DiCE need to be interpreted with prudence. The complexity of their outputs invites thorough academic examination, highlighting the need for meticulous validation of their application.

6. References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138-52160.

Banerjee, S., Bhattacharyya, S., & Bose, I. (2017). Whose online reviews to trust? Understanding reviewer trustworthiness and its impact on business. Decision Support Systems, 96, 17-26.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Cheng, Y. H., & Ho, H. Y. (2015). Social influence's impact on reader perceptions of online reviews. Journal of Business Research, 68(4), 883-887.

Chernyaeva, O., & Hong, T. (2022). The Detection of Online Manipulated Reviews Using Machine Learning and GPT-3. Journal of Intelligence and Information Systems, 28(4), 347-364.

Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. Journal of Big Data, 2(1), 1-24.

Delaney, E., Pakrashi, A., Greene, D., & Keane, M. T. (2023). Counterfactual explanations for misclassified images: How human and machine explanations differ. Artificial Intelligence, 103995.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Fernández-Loría, C., Provost, F., & Han, X. (2022). Explaining data-driven decisions made by AI systems: the counterfactual approach. MIS Quarterly, 46(3), 1635-1660.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30, 681-694.

Gao, J., Galley, M., Li, L., & Dolan, B. (2019). Neural approaches to conversational AI. Foundations and Trends® in Information Retrieval, 13(2-3), 127-298.

He, S., Hollenbeck, B., & Proserpio, D. (2022). The market for fake reviews. Marketing Science.

Heaven, W. D. (2020). A GPT-3 bot posted comments on Reddit for a week and no one noticed. MIT Technology Review. Retrieved November, 24, 2020.

Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. Decision support systems, 52(3), 674-684.

Illia, L., Colleoni, E., & Zyglidopoulos, S. (2023). Ethical implications of text generation in the age of artificial intelligence. Business Ethics, the Environment & Responsibility, 32(1), 201-210.

Jindal, N., & Liu, B. (2007, October). Analyzing and detecting review spam. In Seventh IEEE international conference on data mining (ICDM 2007) (pp. 547-552). IEEE.

Li, F., Huang, M., Yang, Y., & Zhu, X. (2014c). Learning to identify review spam. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (pp. 2488-2493).

Liu, H., & Lang, B. (2019). Machine learning and deep learning methods for intrusion detection systems: A survey. applied sciences, 9(20), 4396.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765-4774).

Mohawesh, R., Xu, S., Tran, S. N., Ollington, R., Springer, M., Jararweh, Y., & Maqsood, S. (2021). Fake reviews detection: A survey. IEEE Access, 9, 65771-65802.

Mothilal, R. K., Sharma, A., & Tan, C. (2019). Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 607-617).

Ott, M., Cardie, C., & Hancock, J. T. (2013, June). Negative deceptive opinion spam. In Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies (pp. 497-501).

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. arXiv preprint arXiv:1107.4557.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. The University of Texas at Austin.

Petratos, P. N. (2021). Misinformation, disinformation, and fake news: Cyber risks to business. Business Horizons, 64(6), 763-774.

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

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

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

Shojaee, S., Murad, M. A. A., Azman, A. B., Sharef, N. M., & Nadali, S. (2013, December). Detecting deceptive reviews using lexical and syntactic features. In 2013 13th International Conference on Intellient Systems Design and Applications (pp. 53-58). IEEE.

Smith, A., & Anderson, M. (2016). Online shopping and e-commerce. Pew Research Center.

Tandoc Jr, E. C., Lim, Z. W., & Ling, R. (2018). Defining “fake news” A typology of scholarly definitions. Digital journalism, 6(2), 137-153.

Tran, G. A., & Strutton, D. (2020). Comparing email and SNS users: Investigating e-servicescape, customer reviews, trust, loyalty and E-WOM. Journal of Retailing and Consumer Services, 53, 101782.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841.

Wang, B., Zubiaga, A., Liakata, M., & Procter, R. (2015). Making the most of tweet-inherent features for social spam detection on Twitter. arXiv preprint arXiv:1503.07405.

Wang, Y. C., & Chen, T. (2023). Adapted techniques of explainable artificial intelligence for explaining genetic algorithms on the example of job scheduling. Expert Systems with Applications, 121369.

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2021). “Let me explain!”: exploring the potential of virtual agents in explainable AI interaction design. Journal on Multimodal User Interfaces, 15(2), 87-98.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., ... & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint arXiv:2302.11382.

Yoo, K. H., & Gretzel, U. (2009). Comparison of deceptive and truthful travel reviews. In Information and communication technologies in tourism 2009 (pp. 37-47).

Yu, S., Ren, J., Li, S., Naseriparsa, M., & Xia, F. (2022). Graph learning for fake review detection. Frontiers in Artificial Intelligence, 5, 922589.