Information Technology & Tourism (2025) 27:739–766
https://doi.org/10.1007/s40558-025-00329-z

ORIGINAL RESEARCH

Improving trust in online reviews: a machine learning approach to detecting artificial intelligence-generated reviews

Ana Marta Santos[1] · **Nuno Antonio**[1,2]

Received: 10 June 2025 / Revised: 11 June 2025 / Accepted: 16 June 2025 /
Published online: 24 June 2025

Abstract

In the hotel industry, social reputation is critical. Consumers increasingly rely on online reviews for accommodation decisions, making Artificial Intelligence (AI) generated fraudulent reviews a significant threat. Distinguishing between genuine and AI-generated reviews is essential for hotels to maintain credibility. This study creates a unique dataset of AI-generated reviews and combines vectorization methods with text-based features to build a Machine Learning model for identifying non-genuine reviews. Results show that incorporating text-based features significantly improves detection accuracy, and simpler vectorization methods can be effective for simpler datasets. This study contributes to academia by providing a novel methodology and publicly available dataset for further research, and to the hotel industry by enhancing credibility and consumer trust through better review filtering.

Keywords Fraudulent reviews · AI-generated · Natural Language processing · Machine learning · Vectorization methods

✉ Nuno Antonio
nantonio@novaims.unl.pt

Ana Marta Santos
20221371@novaims.unl.pt

[1] NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Lisbon, Portugal

[2] CITUR, Universidade do Algarve, Faro, Portugal

1 Introduction

The internet's growing popularity has led to an increased use of e-commerce platforms for online transactions, accompanied by the prevalence of online reviews (Maurya et al. 2023). Customers often distrust sellers' opinions because of perceived bias and instead rely on detailed online reviews from previous consumers to make informed decisions (Baek et al. 2012). These reviews are considered a modern version of (Electronic) Word Of Mouth and play a crucial role in shaping consumers' purchase decisions (Baek et al. 2012). As a result, most consumers who read online reviews acknowledge their influence on their purchase decisions (Hajek et al. 2023; Zhang et al. 2014). These reviews also guide consumers and hold substantial importance for sellers, who use the insights to develop supplementary marketing strategies and shape future planning (Maurya et al. 2023).

In the hotel industry, a hotel's social reputation, which is heavily influenced by online reviews, is one of the most valuable assets (Antonio et al. 2018). The significant trust highlights this influence that nearly half of consumers place in online reviews, trusting them as much as personal recommendations (Hajek et al. 2023). Many fraudulent reviews have been reported across various platforms in the hospitality industry (Fong et al. 2022). Many companies hire skilled individuals, called spammers, to write deceptive opinion spam to promote their goods or services and gain financial benefits (Maurya et al. 2023).

Fraudulent reviews manipulate product rankings and damage customers' trust in the business (Hajek et al. 2023). Their presence has had a significant financial impact on online commerce, influencing consumers to spend $152 billion in recent years before this study (Gambetti and Han 2023). In response to the increasing issue of fraudulent reviews, companies like Amazon, Yelp, and ReviewMeta [1] have taken measures to monitor and manage their content. They use filters and transparent reports to identify and address fraudulent reviews (Moon et al. 2021). Machine Learning (ML) algorithms are currently used to detect these reviews. However, most studies typically focus on linguistic and stylistic features, which may not be sufficient for comprehensive detection. There is a need to explore alternative approaches (Luo et al., 2023). Distinguishing between genuine and fraudulent reviews remains challenging, as spammers are experts in crafting deceptive content. However, ongoing studies are fully dedicated to addressing this issue (Maurya et al. 2023). Some researchers underscore the need to identify unique characteristics of Artificial Intelligence (AI) generated reviews, as understanding these characteristics is essential for developing more effective detection methods (Jawahar et al. 2020; Luo et al. 2023; Salminen et al. 2022). Most existing research relies on linguistic and stylistic features - such as grammar usage, word frequencies, and n-gram distributions - to detect fraudulent reviews. In contrast, our study expands the feature set to include additional text-based features, such as readability indices, sentiment polarity, and part-of-speech distributions. These measures capture aspects of the review (e.g., emotional tone, ease of reading, presence of named entities) that extend beyond traditional stylistic analysis

[1] For more information, visit the official websites of Amazon, Yelp, and ReviewMeta.

Recent studies have focused on fraudulent reviews crafted by spammers. The rise of Large Language Models like ChatGPT has led to an increase in AI-generated fraudulent online reviews (Wang et al. 2023). The emergence of AI-generated reviews presents unique and escalating challenges that surpass those posed by traditional human-crafted fraudulent reviews. Unlike traditional fraudulent reviews, that may contain detectable errors or distinct stylistic signatures, AI-generated content can be produced at scale rapidly, with high linguistic sophistication, making it more difficult to detect due to its coherent, grammatically correct, and contextually plausible nature (Gambetti and Han 2023; Luo et al. 2023; Salminen et al. 2022). These reviews can also be produced at a large scale with unprecedented speed and low cost, enabling spammers to flood review systems more rapidly than ever before (Gambetti and Han 2023). AI-generated reviews also make it easy and fast to generate large volumes of fake reviews, when compared to traditional ones. They can create detailed, persuasive reviews at scale, making it harder for consumers and platforms to distinguish (Integral Ad Science 2024).

A significant limitation of the current studies on fraudulent reviews is the need for comprehensive and diverse datasets to identify AI-generated reviews. These limitations hinder the development of more accurate detection models (Jawahar et al. 2020; Knoedler et al. 2024; Luo et al. 2023; Salminen et al. 2022). It is essential to recognize that AI lacks a complete understanding of the everyday experiences of regular consumers, so even AI-generated reviews manipulated by spammers may show overly emotional language (Luo et al. 2023). Additionally, AI content is grammatically correct and easier to understand than human-generated content (Gambetti and Han 2023). Traditional linguistics elements can still be used to detect signs of AI-generated reviews (Luo et al. 2023). Despite the use of ML algorithms, marketers and e-commerce professionals have valid concerns. AI-generated reviews can go undetected by humans and even receive higher perceived usefulness scores than those written by humans (Gambetti and Han 2023). A reliable review platform is crucial for companies as it provides genuine product and service improvement feedback. Fraudulent reviews can significantly impact a product's ranking, as online marketplace algorithms use them as indicators (Salminen et al. 2022). This impact extends to both reputation and finances. For example, a one-star decrease on Yelp has been estimated to result in a five to nine% decline in revenue (Luca 2011).

Due to the importance of online reviews and the emerging threat of fraudulent reviews generated by AI, this study needs to explore the following Research Questions (RQs): RQ1)"What are the characteristics of AI-generated reviews compared to genuine reviews?" RQ2)"How can text-based features enhance the model's ability to identify AI-generated reviews?" and RQ3) "What methods are effective in detecting AI-generated reviews?"

To address these RQs, it is essential to thoroughly analyze the differences between authentic and AI-generated reviews (Salminen et al. 2022). To accomplish this, a dataset of reviews from different platforms was adapted and used as the foundation for this analysis. The initial phase of the analysis focused on evaluating various aspects of the reviews, such as language use, sentiment, and content structure. Extensive data preprocessing techniques were employed to identify AI-generated reviews, and vectorization methods were applied to prepare the data for analysis.

Subsequently, twelve different ML models were compared to identify the most suitable model for the data. Since fraudulent reviews can be produced on a larger scale and at a lower cost than human-generated fraudulent reviews (Salminen et al. 2022), there is an urgent need to find solutions to address this issue.

In summary, our study addresses a critical gap in the literature by focusing on the unique characteristics of AI-generated reviews, introducing a novel combination of text-based features and vectorization methods for detection. Beyond exploring these methods, we have also made our new, labeled dataset publicly available to facilitate further research. By enhancing detection accuracy, our approach offers both theoretical contributions - expanding on current fraudulent detection frameworks - and practical benefits for the hospitality industry, where the reliability of reviews directly influences consumer trust and revenue. In doing so, we provide a roadmap for future research that seeks to safeguard online platforms from deceptive AI-generated content.

2 Literature review

Whether written by humans or AI systems (Salminen et al. 2022), fraudulent reviews are deceptive testimonials lacking genuine experience (Lee et al. 2016). AI reviews, as described by Gambetti and Han (2023), are fraudulent testimonials created by AI systems trained on real reviews, aiming to mimic genuine ones. The development of advanced AI models like GPT poses significant challenges for social media platforms in detecting AI-generated reviews (Gambetti and Han 2023).

Distinguishing genuine customer feedback from fake reviews is a critical challenge that managers must address. Concerns about online reviews and academic integrity underscore the need to detect AI-generated content, especially with the rise of tools like ChatGPT (Corizzo and Leal-Arenas 2023). The development of AI tools such as Toolbaz[2] and Junia[3], which generate fraudulent reviews, highlights the importance of understanding the underlying patterns in text generation methods (Shah et al. 2023).

2.1 Characteristics of fraudulent reviews

Spammers have become experts in writing fraudulent reviews, making it difficult to verify their authenticity (Maurya et al. 2023). Understanding the characteristics of these reviews is crucial in online reviews due to their significant impact on consumer trust and purchasing decisions in online marketplaces (Kumar et al. 2023). To gain a comprehensive understanding of fraudulent reviews, three types of reviews will be compared: genuine reviews, human-generated fraudulent reviews, and AI-generated reviews.

Fraudulent reviews often exhibit characteristics that reveal the authors' intent to deceive (Lee et al. 2016). According to Bathla et al. (2022), genuine reviews tend to contain contextual information, while fraudulent reviews often use more emotional

[2] Further insights on Toolbaz can be explored through Toolbaz.com.

[3] For additional details refer to Junia.ai.

language. Spammers may manipulate certain words or phrases because they lack in-depth product or service knowledge. Studies have also suggested that the word distribution in fraudulent and genuine reviews can differ significantly. For instance, the exact words may be used in both types of reviews, but they will be grouped into different topics (Lee et al. 2016).

Reviews generated by spammers often use emotional exaggeration, exclamation marks, and power-dominant language, maintaining a more negative emotional tone than genuine reviews. They typically favor first-person pronouns, avoid third-person usage, lack purchase details, rely on quotes, emphasize work-related content, simplify cognitive effort for persuasion, and are usually written in present and future tenses (Moon et al. 2021).

The emergence of GPT-2, a powerful language model, has further complicated the task of detecting fraudulent reviews. Although it enables the creation of grammatically correct and persuasive text (Salminen et al. 2022), careful analysis can uncover distinctive characteristics (Knoedler et al. 2024). AI texts often exhibit a monotonous writing style due to their tendency to repeat similar words and phrases (Corizzo and Leal-Arenas 2023). They also tend to overuse superlatives, employ fewer verbs, and be shorter and less detailed (Salminen et al. 2022). In contrast, human-generated text tends to incorporate words with greater emotional intensity. (Corizzo and Leal-Arenas 2023). A significant aspect of fraudulent reviews is the presence of common patterns in vocabulary and sentence structure. Researchers have identified specific keywords, repetitive phrases, and, as stated before, unnatural or overly emotional language patterns that can indicate a high likelihood of fraud (Knoedler et al. 2024). Additionally, both human-generated and AI-generated reviews may exhibit distinct patterns in emotional language usage, involving the use of exaggerated positive or negative sentiments that deviate from typical human expression (Jawahar et al. 2020).

Another challenge AI-generated reviews pose is their tendency to receive higher average ratings than their human-generated counterparts. The users posting these fraudulent reviews often exhibit specific behavioral patterns, such as limited activity on Yelp with fewer previously posted reviews, friends, and photos. Interestingly, AI content is generally easier to understand than human-generated content (Gambetti and Han 2023).

While humans face difficulty detecting fraudulent reviews, ML algorithms can effectively identify them (Salminen et al. 2022). However, more research is needed to identify the unique characteristics of AI-generated reviews and advance the development of effective models. Current AI/ML models lack awareness, motives, or understanding of human language to grasp the nuances of fraudulent reviews fully (Salminen et al. 2022).

2.2 Existing methods for detecting fraudulent reviews

The development of effective methods to detect fraudulent reviews has evolved, with researchers identifying diverse approaches to address this prevalent issue. Vidanagama et al. (2020) have categorized these methods into three main groups: ML-based, network-based, and pattern-mining approaches. Additionally, researchers have classified these methods into content-based, behavior-based, information-based,

and spammer group detection strategies. These classifications can provide a comprehensive framework for understanding the various techniques used in identifying fraudulent reviews (Salminen et al. 2022).

Content-based detection involves analyzing the textual content of reviews using Natural Language Processing (NLP) techniques (Salminen et al. 2022). Both Bathla et al. (2022) and Duma et al. (2023) used Deep Learning (DL) techniques, specifically a hybrid model combining Convolutional Neural Networks and Long Short-Term Memory networks, to accurately identify fraudulent reviews. While both studies used NLP techniques, they differed in the scope of features considered. Bathla et al. (2022) focused only on the review text, while Duma et al. (2023) included aspect-specific ratings and overall ratings, demonstrating the effectiveness of a broader feature set. Moon et al. (2021) introduced the "All-Terms" model, which aims to capture a wider range of linguistic features and patterns in review texts. While this approach achieved impressive results, incorporating textual and non-textual information is generally considered more effective (Salminen et al. 2022).

Behavior-based detection effectively analyzes non-textual features such as user IDs, location, review count, and other suspicious behaviors (Salminen et al. 2022). Ruan et al. (2020) developed a Geolocation-based Account Detection Model that effectively combined account-related features (number of reviews and responses) with geolocation information (offline activity locations) to identify fraudulent reviews. Similarly, Wang and Chen (2020) proposed FraudGuard, a model designed to detect emerging Reputation Fraud Campaigns by analyzing collusive behaviors and interactions among reviewers, reviews, and products. While the findings are promising in both studies, acquiring features such as IP or similarly exclusive information may not be practical for everyone.

Hajek et al. (2023) highlighted the importance of combining aspect-based sentiment analysis (ABSA) with behavioral and linguistic features using ML methods to detect fraudulent reviews. In a similar vein, Budhi et al. (2021) proposed a hybrid approach that merged content-based and behavior-based features using ML classifiers, introducing a set of 133 features. Integrating these extensive features with the ABSA model proposed by Hajek et al. (2023) could further improve the robustness of the models. The studies related to human-generated fraudulent detection that have been mentioned are reflected in Table 1.

The rise of artificially generated content presents a new challenge for researchers studying fraudulent reviews crafted by spammers. In 2022, TripAdvisor removed over 20,000 AI reviews, highlighting the seriousness of this issue (Collinson 2023). These reviews pose significant consumer risks, making it crucial to use ML techniques to identify them. While some studies have focused on developing automation techniques to detect AI-generated content (Luo et al. 2023), few have specifically addressed the problem in the context of online reviews.

Different studies have taken distinct positions and used varied techniques to address this challenge. For example, Salminen et al. (2022) and Gambetti and Han (2023) adopted existing models, including OpenAI's GPT model, to generate fraudulent reviews. Salminen et al. (2022) fine-tuned the RoBERTa model, while Gambetti and Han (2023) successfully used a GPT output detector, surpassing existing solutions. Despite the success of Gambetti and Han (2023), their study, which focused

Table 1 Summary of human-generated fraudulent detection techniques

Study	Technique Used	Evaluation Metrics	Domain-specific	Background/Context
Ruan et al. (2020)	GADM (Adaptive Boosting, Long Short-Term Memory)	Accuracy, F1 score	O2O platforms	Proposed a model for detecting manual fraudulent reviews, leveraging account and geolocation features.
Wang and Chen (2020)	Conditional Random Fields (CRFs) for labeling reviews in a MARS	F1 score, emphasizing a tradeoff between precision and recall	Online product reviews	Introduced FraudGuard, a model designed to detect emerging Reputation Fraud Campaigns.
Budhi et al. (2021)	Hybrid content-based and behavior-based feature extraction approach, ML classifiers	Binary accuracy, precision, recall, F1 score, area under the curve (AUC)	E-commerce platforms	Proposed a hybrid content and behavior-based operating in a parallel environment to detect fraudulent reviews.
Moon et al. (2021)	Survey-based text categorization, "All Terms" model	Predictive accuracy, goodness-of-fit compared to benchmark methods	Hotel Services	Created a survey-based text categorization approach to identify fraudulent consumer reviews.
Bathla et al. (2022)	Aspect extraction and sentiment analysis using Conventional Neural Networks and Long Short-Term Memory hybrid model	Precision, F-measure, and Accuracy	E-commerce platforms	Researchers employed Deep Learning techniques to analyze aspects of reviews.
Duma et al. (2023)	Deep Hybrid Model	Accuracy, Precision, Recall, F1 Score	Hospitality Industry	Developed a deep hybrid model for fraud detection, considering both aspect-specific and overall ratings.
Hajek et al. (2023)	ABSA and ML learning methods	Accuracy, AUC, and F1-score	E-commerce platforms	Proposed a fraudulent review detection model using ABSA customized for diverse product categories.

on data from 2021 to 2022 during the COVID-19 pandemic, would benefit from a broader timeframe to fully understand the impact of fraudulent reviews on consumer behavior. Conversely, Gambetti and Han (2023) used high-quality elite restaurant reviews, leading to greater credibility in their dataset.

Luo et al. (2023) introduced crucial categories of variables for identifying AI-generated reviews: traditional linguistic features, innovative linguistic characteristics, and AI probability. Their "AI-Generated Review Detection with Cumulative Probability" method effectively combined linguistic features and AI probability to distinguish genuine and AI-generated reviews. While the study did not include behavioral features due to data limitations, it represented a notable advancement in AI review detection techniques.

Knoedler et al. (2024) employed Copyleaks[4] to systematically analyze human-written and AI-generated patient reviews in plastic and aesthetic surgery. Although

[4] For more information, visit their website: copyleaks.com.

their study did not introduce a novel model, it offered valuable insights into the distinguishing characteristics of AI-generated reviews. Studies exploring AI text identification can offer valuable insights into potential detection methods. However, there is still a limited number of studies specifically focused on creating new methods for detecting AI-generated reviews.

Islam et al. (2023) conducted a study on using ML algorithms to identify texts generated by AI. They evaluated eleven different ML and DL algorithms and discovered that Extremely Randomized Trees Classification achieved the highest accuracy of 77%. This suggests that a similar approach could be promising for AI-generated reviews. In contrast, Maloyan et al. (2022) focused on developing a pipeline that utilizes fine-tuned language models, specifically Bidirectional Encoder Representations from Transformers (BERT) models, and an ensemble method. Their approach achieved first place in the binary classification for the DIALOG-22 RuATD challenge. However, it is important to note that their solution demands considerable computational power and is unsuitable for real-time applications.

The studies related to AI-generated reviews mentioned in this study are summarized in Table 2. Despite the increasing sophistication of AI reviews, which can appear indistinguishable from those composed by humans, certain linguistic patterns still indicate their AI origin. This leads us to contemplate conventional linguistic elements to identify distinctive traits associated with AI reviews (Luo et al. 2023).

Table 2 Summary of AI-generated fraudulent detection techniques

Study	Technique Used	Evaluation Metrics	Domain-Specific	Background/Context
Salminen et al. (2022)	Language models (GPT-2, ULMFiT, RoBERTa) for fraudulent review generation and detection	Accuracy, Kappa metric, Sensitivity, Specificity, AUC	Online product reviews	Investigated the use of fine-tuned RoBERTa and ML models to detect fraudulent reviews. Also evaluates the performance of human annotators in detecting fraudulent reviews compared to the ML models.
Gambetti and Han (2023)	Fine-tuned a pre-trained GPT-2 model to classify fraudulent versus real reviews	Accuracy, F1-score, precision, recall, and AUC-ROC	Restaurant Reviews	Proposed a fine-tuning of a GPT output detector, using high-quality elite restaurant reviews to generate AI-generated fraudulent reviews.
Luo et al. (2023)	AI-Generated Review Detection with Cumulative Probability method	Accuracy, precision, recall, F-Measure, and Area Under Curve (AUC) score	-	Proposed a novel method for AI-generated review using cumulative probability, traditional linguistic features, and evaluation metrics across diverse real-world datasets.
Knoedler et al. (2024)	AI Content Detection Software (Copyleaks AI) and Human Participants	Classification Performance, Emotional Tone Analysis, Review Length Analysis, Use of Discipline-Specific Vocabulary	Plastic and Aesthetic Surgery	Explored challenges distinguishing between authentic and AI-generated patient reviews in plastic and aesthetic surgery.

In summary, recent works on review fraud detection often rely on linguistic and stylistic features, such as word frequencies, n-gram patterns, and syntactic structures, to discriminate between deceptive and genuine content (Moon et al. 2021; Maurya et al. 2023). While these features have yielded promising results against human-generated fraudulent reviews, they may not sufficiently capture the nuanced patterns in AI-generated text, which increasingly mimics genuine writing (Salminen et al. 2022). Furthermore, studies focusing on AI-generated reviews remain limited in number and scope. For instance, Salminen et al. (2022) and Luo et al. (2023) highlight the value of Machine Learning tools but rely largely on smaller or specialized datasets, which hampers the generalizability of their findings. Behavior-based detection approaches provide valuable non-textual signals (e.g., user posting history, IP data) but are less viable when such metadata is unavailable for open platforms (Ruan et al. 2020; Wang and Chen 2020). At the same time, most research blends either purely content-based or purely behavior-based strategies, leaving an opportunity to investigate how a richer set of text-centric features -readability metrics, sentiment polarity, named-entity densities - might enhance detection accuracy.

Taken together, the existing literature demonstrates steady progress in detecting human-generated fraudulent reviews, but it underscores two clear limitations: first, an overemphasis on traditional linguistic and stylistic cues that may not detect advanced AI-generated reviews effectively, and second, an overall lack of comprehensive, diverse datasets that capture a broad range of AI-generated and genuine content. These gaps indicate the need for research that (1) expands beyond standard stylistic markers into more holistic text-based features and (2) assembles a balanced corpus of human-written and AI-generated reviews suitable for robust Machine Learning experiments.

More recent research has made significant strides in the detection of AI-generated reviews. Theoretical foundations and zero-shot detection methods have been emphasized, with studies demonstrating that reliable detection is feasible, even as AI-generated text approaches human-quality, provided sufficient data and appropriate model architectures are employed (Chakraborty et al. 2024; Hans et al. 2024). Other work has advanced domain-agnostic detection techniques for diverse, real-world applications (Li et al., n.d.). Additionally, several studies have addressed robustness and language-specific challenges, broadening detection capabilities across languages and enhancing resilience against adversarial inputs (Bao et al. 2024).

3 Methodology

This study aimed to achieve two primary objectives: to develop a robust model to detect whether a review is AI-generated or genuine, explicitly focusing on the hotel industry, and to identify the characteristics of AI-generated reviews.

This study adopted a systematic approach, outlined in Fig. 1. The initial step involved data preprocessing, focusing on cleaning the textual data and extracting relevant features, which could be either numerical or textual. Various vectorization methods were evaluated and implemented to convert textual features into numerical representations suitable for the ML algorithms. Finally, twelve ML models were

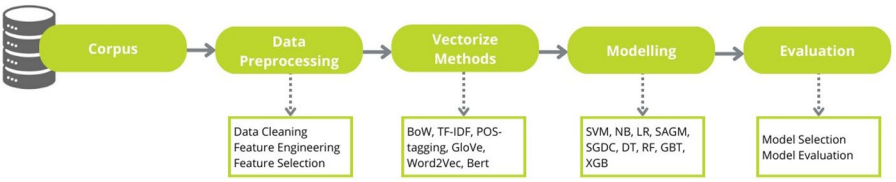

Fig. 1 Study approach

assessed, and key metrics were used to assess the performance of the trained models. The primary tool used to develop the model was Jupyter Notebook, using Python as the programming and analysis language.

3.1 Data Understanding

Selecting the appropriate dataset is vital as it significantly impacts the effectiveness and generalization of the detection models (Maurya et al. 2023). However, acquiring labeled datasets for AI-generated review detection can be challenging. No existing public datasets for this specific purpose were found during the data collection phase. For that reason, a new dataset was developed. While meta-data or behavioral features such as review posting time and star rating given by the reviewers could improve model performance (Maurya et al. 2023), for the reasons mentioned above, that type of meta-data was not available. Therefore, we focused on the text content of the reviews.

Researchers often collect genuine reviews from trustworthy platforms to create AI-generated reviews using tools like OpenAI's GPT-2 (Gambetti and Han 2023; Knoedler et al. 2024; Salminen et al. 2022). Following this approach, a public dataset containing genuine and deceptive reviews of 20 hotels in Chicago was obtained (Ott et al. 2011, 2013). The dataset included 800 genuine reviews and 800 deceptive reviews from Mechanical Turk[5], which were excluded since they were not AI-generated reviews.

Several AI tools, namely ChatGPT[6], Gemini[7], ToolBaz,[8] and Chatsonic[9], were used to create the AI-generated reviews. Specific prompts were crafted to instruct the AI tools to mimic human written styles to make the reviews similar to genuine ones. This approach reflects the tactics spammers use in the real world, who deliberately attempt to make AI-generated reviews indistinguishable from genuine ones. Details regarding the sources and target distributions of the final dataset are shown in Fig. 2.

To ensure the complexity of the generated reviews, a study (see Appendix A) was conducted involving 20 reviews and 10 participants to compare their ability to detect AI-generated reviews against the final models' performance. The final dataset

[5] www.mturk.com.

[6] chat.openai.com.

[7] gemini.google.com.

[8] toolbaz.com/writer/review-generator.

[9] writesonic.com.

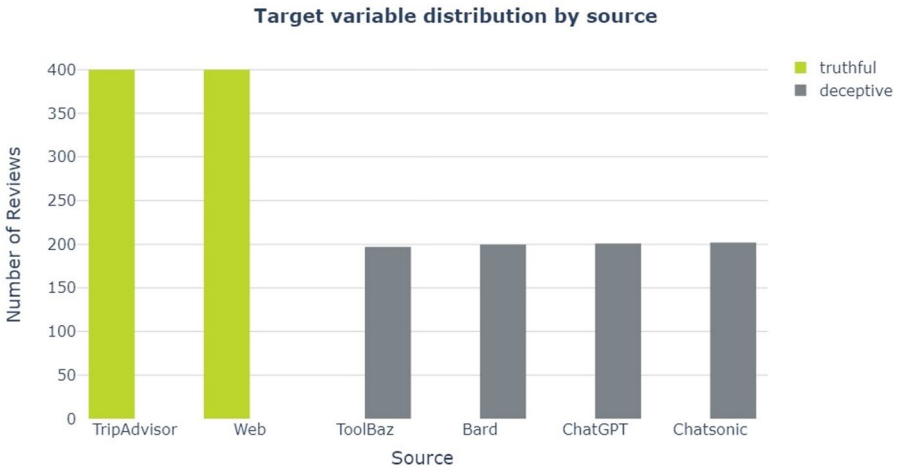

Fig. 2 Target variable distribution by source

Table 3 Summary statistics

Feature	Unique	Top	Frequency
Deceptive	2	Truthful	800
Hotel	20	Conrad	80
Polarity	2	Positive	800
Source	6	TripAdvisor	400
Text	1594	My daughter and I woke…	2

included an equal balance of positive and negative sentiment within the 800 deceptive reviews, maintaining a comprehensive range of reviews and enriching dataset diversity. As detailed in Table 3, it contained five features, each associated with 1,594 unique reviews. Among these features, only "text" (review corpus) and "deceptive" (target variable) were identified as beneficial for the model. The remaining features were only retained in the first analysis for exploratory data analysis.

3.2 Data Preparation

The dataset needed to be transformed into its final form for model input to address the challenges associated with raw data. This involved performing data cleaning to remove inconsistencies and errors. Subsequently, new features were created from the review content, and feature selection to reduce data dimensionality was conducted. Finally, five vectorization methods were used to ensure the review content was processed correctly and suitable for the model.

3.2.1 Data cleaning

Using the Python pandas library (McKinney 2010), duplicate entries were removed, resulting in a dataset free of duplicates. Since there were no missing values, no corrective actions were necessary.

Text preprocessing techniques were implemented to standardize the format and improve the texts' suitability for NLP tasks. These techniques included HTML tag removal with BeautifulSoup (Richardson 2019), converting text to lowercase, eliminating punctuation and special characters using Python regex (Rossum 2020), removing stop words, and applying lemmatization using the NLTK Library (Bird et al. 2009). Lemmatization was chosen over stemming because it produces actual words while stemming truncates word endings.

3.2.2 Feature engineering

In the feature engineering process, various features were created from the review content to enhance the model performance while managing complexity (Maurya et al. 2023). Initially, statistical features such as words, number count, characters count, syllable count, word density, uppercase letters, stop words, punctuation marks, contractions, and sentence count were generated. Although less explored, this approach provided valuable insights into the text data. Libraries such as Pandas (McKinney 2010) and NLTK (Bird et al. 2009) facilitated data manipulations and linguistic tasks, along with Python built-in functions for specific calculations. Readability features were incorporated (Luo et al. 2023) using the Textstat library (Bansal and Aggarwal 2019). These features included the Flesch-Kincaid grade level, Coleman-Liau Index, SMOG index, Flesch Reading Ease score, Fog Scale, and Dale Chall Readability score, offering valuable insights into the text's complexity and understandability, as explained in DuBay (2004) study. Additionally, a feature that calculates the Reading Time was included. For sentiment analysis, polarity and subjectivity scores were obtained using the TextBlob library (Loria 2018). Part of Speech (POS) tagging (Vanroose 2004) and Named Entity Recognition (NER) (Au et al. 2022) techniques were employed using NLTK (Bird et al. 2009) and SpaCy libraries (Honnibal and Montani 2017) to identify grammatical elements and named entities. The final set of numerical features (before the feature selection process) is represented in Table 4.

The dataset's limited size restricted the number of features that could be introduced to avoid overfitting. A combination of simple and more complex features was created to capture different aspects of the review content. However, the scope of feature engineering was restricted to the review content itself, as the inclusion of numerical features such as author information or star ratings was not available in the dataset.

Normalization was performed using Min-Max Normalization with a feature range of (-1, 1) to ensure a consistent range for all features. The implementation was facilitated by the Scikit-learn library (Pedregosa et al. 2012).

3.2.3 Vectorize methods

For ML algorithms to understand text reviews, it is necessary to transform the preprocessed text into numerical representations (Sueno 2020). Using vectorization methods becomes imperative in this context. By exploring various techniques, the objective was to find the most effective approach for capturing the meaning within the reviews. Prior to implementing this process, the dataset was divided into 80% for

Table 4 Numerical features created

Feature	Feature Description	Feature Name	Feature Description
syllable	Number of syllables.	**jj_count**	Number of adjectives.
words	Number of words.	**SMOG_Index**	SMOG Index.
num_numbers	Number of numbers.	**Flesch_Reading_Ease**	Flesch Reading Ease.
unique_words	Number of unique words.	**Flesch_Kincaid_Grade**	Flesch-Kincaid grade level.
characters	Number of characters.	**Fog_Scale**	Fog Scale Readability Index.
stopwords	Number of stopwords.	**Dale_Chall_Readability**	Dale-Chall Readability Score.
word_density	Ratio of total words to total characters.	**Reading_Time**	Estimated time to read.
punctuation	Number of punctuation marks.	**Coleman_Liau_Index**	Coleman-Liau Readability Index.
uppercase	Number of uppercase letters.	**polarity_text**	Polarity score of the text.
sentences	Number of sentences.	**subjectivity_text**	Subjectivity score of the text.
contractions	Number of contractions.	**person_count**	Number of person entities.
nn_count	Number of nouns.	**org_count**	Number of organization entities.
pr_count	Number of pronouns.	**loc_count**	Number of location entities.
vb_count	Number of verbs.		

training and 20% for testing to prevent data leakage in the test dataset. The initial approach used Bag of Words (BoW) and Term Frequency Inverse Document Frequency (TF-IDF) techniques (Abubakar et al. 2022). Scikit-learn's (Pedregosa et al. 2012) facilitated the implementation of these methods. They focused on the importance of words within a review and their relative frequency across the entire dataset, which is particularly useful for smaller datasets.

However, pre-trained word embeddings were explored to go beyond basic word counts and capture deeper semantic relationships within the text. These embeddings included Global Vectors for word representation (GloVe) (Pennington et al. 2014) and Word2Vec (Abubakar et al. 2022). The GloVe embeddings were sourced from the "glove.6B.100d.txt"[10] file, while Word2Vec utilized both Continuous Bag of Words (CBOW) and Skip-gram architectures via the Gensim Library (Rehurek and Sojka 2011). Both embeddings represent words based on their co-occurrence patterns, essentially capturing how words relate to each other. This enables the model to understand the meaning behind the words and not just their presence (Turing et al. 2023).

Finally, BERT embeddings (Masala et al. 2020) obtained with the transformers library were also considered, as they provide the potential for further optimization. However, given the potential for overfitting with smaller datasets, the focus remained on the abovementioned methods. The more complex vectorized methods were con-

[10] To acess the GloVe file used, you can visit the following link: https://nlp.stanford.edu/projects/glove/.

verted into 2-dimensional vectors and data frames to match the previously mentioned numerical features.

3.2.4 Feature selection

Identifying the most relevant features is crucial for optimizing model performance and reducing computational complexity. Spearman's rank correlation coefficients were used to visualize the relationships between features and the target variable. An analysis of the correlations between features was also done. Further feature analysis was done by applying three statistical techniques: (1) Recursive Feature Elimination (RFE) (Guyon et al. 2002) to iteratively remove features with minimal contribution to the model's learning and prediction capabilities; (2) Analysis of Variance (ANOVA) to ensure the retention of features with a statistically significant influence on the target variable; (3) Lasso Regression (Tibshirani 1996) helped create a more concise and informative set of features for model training (Muthukrishnan and Rohini 2016). The decision on how to keep or remove features based on these methods' results is presented in Table 5.

Table 5 Feature selection summary

Features	Spearman Correlation	ANOVA	RFE	Lasso	Conclusion
syllable	Keep	Keep	Remove	Keep	Keep
words	Remove	Keep	Remove	Keep	Keep
num_numbers	Keep	Keep	Remove	Remove	Keep
unique_words	Remove	Keep	Remove	Remove	Remove
characters	Remove	Keep	Remove	Remove	Remove
stopwords	Remove	Keep	Remove	Remove	Remove
word_density	Keep	Keep	Remove	Keep	Keep
punctuation	Remove	Keep	Remove	Remove	Remove
uppercase	Keep	Keep	Remove	Remove	Keep
sentences	Keep	Keep	Remove	Remove	Keep
contractions	Keep	Keep	Remove	Keep	Keep
nn_count	Remove	Keep	Remove	Keep	Remove
pr_count	Remove	Keep	Remove	Keep	Remove
vb_count	Remove	Keep	Remove	Keep	Remove
jj_count	Keep	Keep	Remove	Keep	Keep
SMOG_Index	Keep	Keep	Keep	Keep	Keep
Flesch_Reading_Ease	Keep	Keep	Remove	Keep	Keep
Flesch_Kincaid_Grade	Remove	Keep	Remove	Keep	Remove
Fog_Scale	Remove	Keep	Remove	Keep	Remove
Dale_Chall_Readability	Remove	Keep	Keep	Keep	Keep
Reading_Time	Remove	Keep	Remove	Remove	Remove
Coleman_Liau_Index	Remove	Keep	Keep	Keep	Keep
polarity_text	Keep	Keep	Remove	Keep	Keep
subjectivity_text	Keep	Remove	Remove	Keep	Keep
person_count	Keep	Keep	Remove	Remove	Remove
org_count	Keep	Keep	Remove	Remove	Remove
loc_count	Keep	Remove	Remove	Remove	Remove

3.3 Explanatory data analysis

An Explanatory Data Analysis (EDA) was conducted to investigate the characteristics of AI-generated reviews. This analysis employed text-based techniques, including text statistics, n-grams, word clouds (Oesper et al. 2011), sentiment analysis, and NER and POS tagging. Visualizations were generated using libraries like Matplotlib (Hunter 2007) and Plotly (Inc., P. T., 2015).

The analysis of text statistics and n-gram aimed to reveal phrasing patterns within the reviews. Histograms were generated to illustrate the distribution of text statistics features, top-10-word frequencies, and top-10 2-gram frequencies.

Word clouds were employed with the WordCloud library (Oesper et al. 2011) to visualize the frequency and distribution of words within positive and negative reviews, further categorized by their origin: AI-generated or Genuine reviews. This facilitated the identification of linguistic patterns and vocabulary characteristics specific to each category. The results of this analysis are presented in Fig. 3.

Sentiment analysis explored the polarity and subjectivity scores of the reviews. While the polarity score indicates the positive or negative sentiment of the text, the subjectivity score represents the factual or emotional nature of the text (Satapathy et al. 2022). The distributions of these scores can be observed in Figs. 4 and 5. In the positive polarity plot, values above 0 indicate positive sentiments, while values below 0 indicate negative sentiments. Conversely, in the subjectivity scores plot, values closer to 0 represent a more factual text, while values closer to 1 represent a more subjective and emotional text. These visualizations provided insights into the reviews' overall sentiment and emotional tone from both AI and genuine sources.

A detailed NER and POS tagging analysis revealed insights into the structure and content of AI-generated and genuine reviews. Figures 6 and 7 visualize each review type's top 10 frequencies of NER and POS tagging. Analyzing these patterns helped identify how AI-generated and genuine reviews differ in their references to specific entities and how they construct sentences grammatically.

Fig. 3 WordClouds

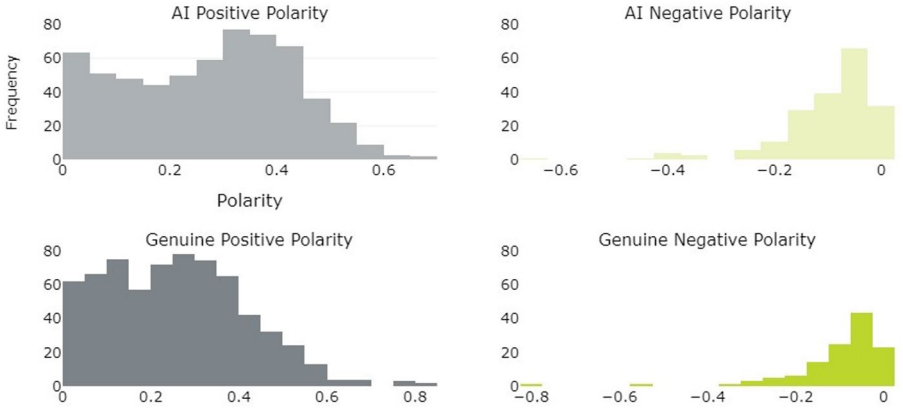

Fig. 4 Distribution of polarity scores

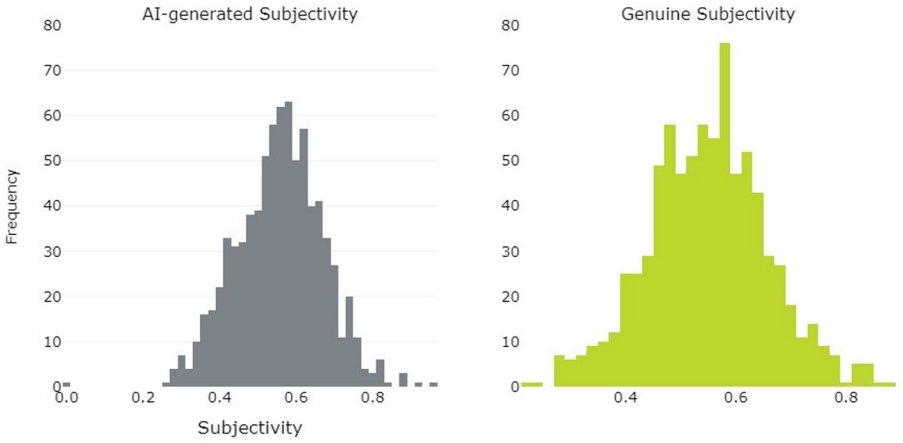

Fig. 5 Distribution of subjectivity scores

3.4 Modeling and evaluation

The selection of twelve ML models was driven by two key considerations: their established effectiveness in text classification tasks and their potential for identifying fraudulent behavior. Research by Islam et al. (2023) and Maurya et al. (2023) informed the selection of models like Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), and Gradient Boosting, all known for their strengths in text analysis. Additionally, models like Bagging Classifier, Multi-Layers Perceptron Model Classifier (MLP), Adaptive Boosting (Hastie et al. 2009), Stochastic Gradient Descent (SGD), Decision Tree (Robertson 2004), eXtreme Gradient Boosting (Chen and Guestrin 2016), and Extra Tree Classifiers (Geurts et al. 2006) were included due to their ability to learn com-

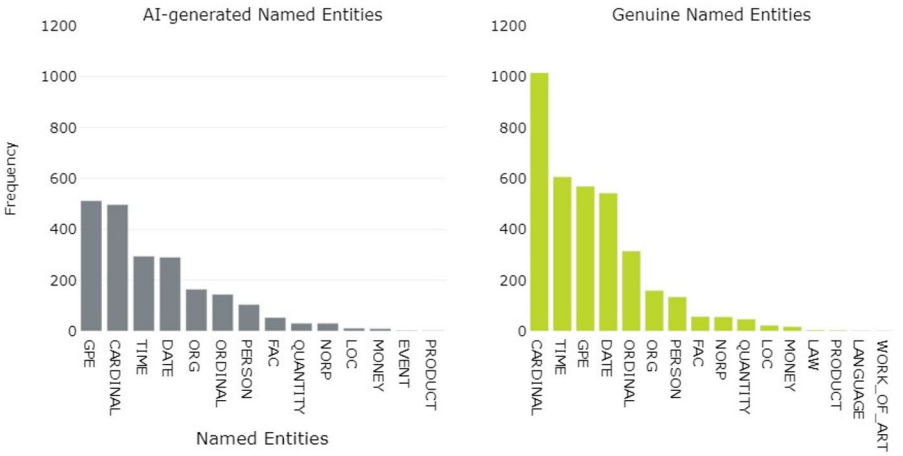

Fig. 6 Top named entity frequencies

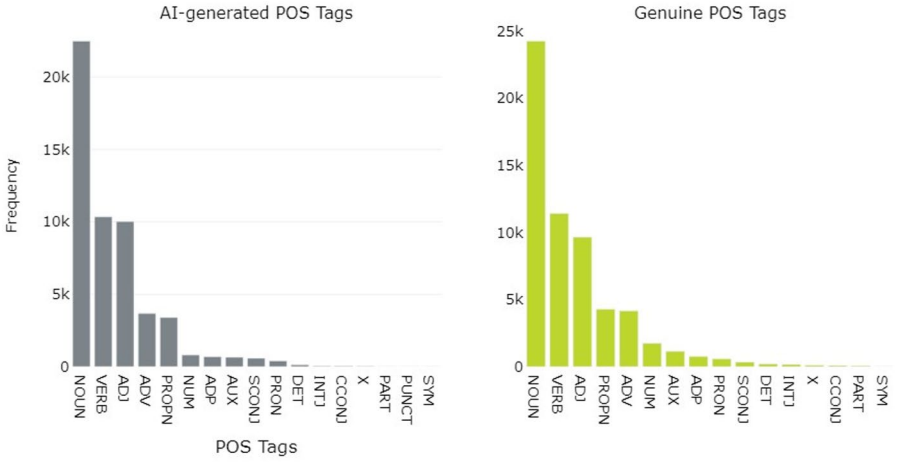

Fig. 7 Top part-of-speech tag frequencies

plex patterns within data, potentially helpful in identifying linguistic characteristics of AI-generated reviews.

Each model was initially trained using default hyperparameters on the training data with two distinct feature sets: one using only vectorize techniques and another combining vectorized and text-based features. This approach allowed for a comparison of the effectiveness of each feature set in model performance. A Random Search approach was employed for the top 3 performing models. This involved defining a broad range of potential hyperparameter configurations using the Scikit-learn (Pedregosa et al. 2012) "RandomSearchCV" function. Following this automated search, the best-performing hyperparameter combination for each model was further

fine-tuned manually to enhance predictive capabilities and address potential issues like underfitting or overfitting.

Overfitting occurs when a model memorizes the noise and patterns within the training data at the expense of generalizability. This leads to poor performance on unseen data. Underfitting happens when a model is too simplistic and cannot capture the essential relationships within the data, resulting in inaccurate predictions (Pothuganti 2018).

The effectiveness of the trained models was assessed by a combination of metrics commonly employed for balanced datasets: F1-Score, Accuracy, Precision, and Recall (Powers 2008). These metrics were calculated using the Scikit-learn Library (Pedregosa et al. 2012) and provided a comprehensive evaluation of model performance across various aspects of prediction accuracy. A 5-fold Cross-Validation (CV) (Berrar 2019) approach was implemented to ensure robust evaluation and generalization. This involved training the model on a portion of the data and testing on the remaining portion, repeated k times. Scikit-learn (Pedregosa et al. 2012) facilitated this process, helping avoid overfitting and providing a more reliable performance estimate. Finally, the CV results were compared against the independent test dataset's performance to assess potential overfitting issues.

4 Results and discussion

The analysis focused on the characteristics of AI-generated reviews to better understand this data type and to evaluate the effectiveness of the different vectorization methods. Additionally, insights into the additional experiment are provided and compared with other relevant studies.

4.1 AI-reviews characteristics

A descriptive analysis was conducted to address the RQ regarding the characteristics of AI-generated reviews. The results showed that AI reviews are generally shorter, with less varied sentence and word length and higher word density, indicating the use of repetitive phrases and shorter words. This aligns with Corizzo and Leal-Arenas (2023) observation that AI-generated text often exhibits a monotonous writing style.

In contrast, genuine reviews use richer language with varied punctuation and POS. They also contain more contextual information and specific vocabulary than AI-generated reviews, which rely on generic sentiment terms. While some studies noted that AI-generated reviews tend to have more emotional language (Bathla et al. 2022; Jawahar et al. 2020), this study found that genuine reviews express stronger feelings. In contrast, AI reviews remain generally neutral with attempts to mimic emotional language.

AI-generated reviews often use function words like "despite" and "overall", indicating more even word usage but lacking factual information. On the other hand, genuine reviews contain more factual details such as locations, dates, and times, reflecting real experiences. While the grammatical structure of AI-generated reviews

may resemble human writing, genuine reviews have a slightly richer structure, using a wider variety of nouns, verbs, adjectives, and proper nouns.

4.2 Experiments results

Two sets of experiments were conducted to assess the impact of text-based features on detecting AI-generated reviews. In the first set, models were evaluated using only vectorized features with default parameters. In the second set, these features were tested along with text-based features. The results of the top three models are presented in Table 6.

Adding text-based features to vectorized methods generally improved model performance, indicating that these features provide valuable complementary information. However, the effectiveness varied. Word2Vec embeddings initially showed the lowest accuracy at around 53% but significantly improved with text-based features, with an increase of over 34%. However, the final accuracy was still lower than other methods. In contrast, TF-IDF vectorization demonstrated the best performance, with moderate gains from text-based features ranging from 1.88 to 3.14% increase. This

Table 6 Experiment results in test dataset

BoW Vectorize Method		
Model	**Accuracy (set 1)**	**Accuracy (set 2)**
LR	92.48%	94.36%
SVM	91.22%	94.04%
MLP	91.84%	93.73%
TF-IDF Vectorize Method		
Model	***Accuracy (set 1)***	***Accuracy (set 2)***
SGD	91.84%	94.98%
MLP	91.84%	94.67%
LR	92.79%	94.67%
GloVe Embeddings Vectorize Method		
LR	91.85%	92.48%
RF	90.28%	91.85%
SVM	91.54%	91.22%
Word2Vec CBOW Vectorize Method		
Model	***Accuracy (set 1)***	***Accuracy (set 2)***
LR	53.29%	89.97%
SVM	53.29%	89.66%
KNN	53.29%	87.78%
Word2Vec Skip-Gram Vectorize Method		
Model	***Accuracy (set 1)***	***Accuracy (set 2)***
MLP	53.29%	90.90%
LR	53.29%	90.60%
SGD	53.29%	89.97%
Bert Vectorize Method		
Model	***Accuracy (set 1)***	***Accuracy (set 2)***
LR	92.48%	93.42%
MLP	91.54%	92.80%
SVM	91.22%	92.48%

outcome aligns with Dessı et al. (2021), which suggested that simpler methods might outperform complex embeddings with limited training data.

4.3 Hyper tunning results

Despite achieving good results with general vectorization methods, there was significant overfitting, as indicated by the performance gap between test and train data. This overfitting likely occurred because the models were not able to generalize well with the limited size of the dataset. To address this issue, Random Search and manual tunning were used. Random Search aimed to find optimal hyperparameter configuration, while manual tuning was applied to models with other 3% overfitting. The final models' performance of the models is detailed in Table 7.

Table 7 Final models' performance

BoW Vectorize Method				
Model	**Accuracy (train)**	**Accuracy (test)**	**Recall (train)**	**Recall (test)**
LR	96.55%	94.04%	94.76%	91.95%
SVC	95.14%	93.73%	91.99%	89.93%
MLP	95.61%	93.73%	94.49%	93.289%
TF-IDF Vectorize Method				
Model	***Accuracy (train)***	***Accuracy (test)***	***Recall (train)***	***Recall (test)***
SGD	95.92%	94.67%	93.99%	93.29%
MLP	97.57%	94.98%	96.61%	94.63%
LR	92.86%	94.36%	90.60%	93.96%
GloVe Embeddings Vectorize Method				
Model	***Accuracy (train)***	***Accuracy (test)***	***Recall (train)***	***Recall (test)***
LR	90.98%	92.48%	88.60%	92.62%
RF	93.65%	90.60%	89.06%	87.25%
SVC	93.02%	92.16%	90.14%	91.28%
Word2Vec CBOW Vectorize Method				
Model	***Accuracy (train)***	***Accuracy (test)***	***Recall (train)***	***Recall (test)***
LR	86.59%	90.28%	84.90%	87.25%
SVC	87.92%	89.66%	82.74%	85.52%
KNN	88.62%	87.78%	88.29%	84.56%
Word2Vec Skip-Gram Vectorize Method				
Model	***Accuracy (train)***	***Accuracy (test)***	***Recall (train)***	***Recall (test)***
LR	86.8%	90.60%	85.21%	91.28%
SGD	86.6%	90.28%	83.05%	88.59%
MLP	89.33%	88.09%	85.67%	87.92%
Bert Vectorize Method				
Model	***Accuracy (train)***	***Accuracy (test)***	***Recall (train)***	***Recall (test)***
LR	94.75%	93.42%	93.37%	91.28%
MLP	93.49%	92.16%	91.37%	89.26%
SVC	93.02%	91.85%	90.45%	89.26%

The TF-IDF and BoW methods performed better than word embeddings like GloVe and Word2vec, suggesting that simpler methods capture essential characteristics better for this specific dataset. BERT embeddings showed competitive performance, indicating their potential effectiveness, but they might not outperform simpler methods. LR achieved strong results across most vectorization methods, highlighting its effectiveness for this classification task. MLP also showed promised results in recall but needed to be more consistent with LR. It also had lower accuracy for training and test data, indicating that these models needed help correctly identifying AI-generated reviews. The best-performing combination was TF-IDF with SGD, achieving an F1-Score of 94.63%, Accuracy of 94.67%, Precision of 95.21% and Recall of 93.29% in the test dataset.

4.4 Reviews' complexity

A human evaluation was conducted to assess the complexity and readability of AI-generated reviews. A group of 10 participants was asked to categorize 20 reviews as "human" or "AI-generated". The average human accuracy was 60.50%, while the machine's final combination accuracy was 95%. This indicates that distinguishing AI-generated reviews from genuine ones was challenging for the participants. The AI-generated reviews in this study exhibited a complexity that made it difficult for an average person to distinguish from human writing.

This finding supports the notion that humans find AI-generated reviews challenging to detect. The participants' difficulty in accurately identifying AI-generated reviews highlights the potential value of automated detection methods. The superior performance of the final ML model, significantly better than the average human accuracy, highlights its value for tasks like filtering out AI-generated reviews from online platforms.

4.5 Performance comparison

Several studies have explored detecting fraudulent reviews, but only a few focus on identifying AI-generated reviews. Salminen et al. (2022) and Gambetti and Han (2023) achieved promising results using fine-tuned RoBERTa and GPT-Neo models without performing any additional features or vectorization techniques. Meanwhile, Luo et al. (2023) claimed to be pioneers in using ML algorithms for AI-generated review detection, but they have only used features without vectorization techniques and achieved an 85.57% accuracy. In contrast to existing studies, this research implemented text-based features with a combination of vectorization methods.

In addition to AI-generated reviews, studies on fraudulent content detection written by spammers also offer valuable insights. Bathla et al. (2022) used a complex CNN-LSTM model for detecting fraudulent reviews based on specific aspects, but their accuracy was lower than in the current study. Budhi et al. (2021) employed 113 features without using vectorization techniques on an unbalanced dataset, employing both under and over-sampling, which limits comparability. On the other hand, Duma et al. (2023) achieved higher performance using a public dataset with a larger size

Table 8 Performance comparison

Author	Model	Accuracy	Precision	Recall	F-Scores
Luo et al. (2023)	Ada-Boost	88.57%	87.17%	83.47%	85.28%
Salminen et al. (2022)	fakeRoBERTa	96.64%	97%	97%	97%
Gambetti and Han (2023)	GPT-Neo fine-tune	95.51%	95.80%	95.15%	95.48%
Bathla et al. (2022)	LSTM	89.5%	88.7%	-	87.2%
Duma et al. (2023)	CNN & LSTM	99.5%	97.7%	96.1%	96.5%
My approach	TF-IDF with SGD	94.67%	95.21%	93.29%	94.63%

and a more complex CNN-LSTM model. The comparison results with other studies can be seen in Table 8.

5 Conclusions and future works

This study addressed a critical challenge faced by online review platforms, particularly those related to the hospitality industry: distinguishing between genuine and AI-generated reviews. This is important for maintaining user trust and promoting the reputation of hotels, which is a valuable asset for the industry (Antonio et al. 2018). To address this problem, a model capable of accurately identifying AI-generated reviews was developed by researching a combination of vectorization techniques and ML algorithms.

Significant progress was achieved in detecting AI-generated reviews. The final model, which combined TF-IDF vectorization and SGD model, achieved an impressive F1 Score of 94.63%, surpassing previous studies (Luo et al. 2023). This performance can be attributed to the exploration of various vectorization techniques, the implementation of text-based features, and a dual approach to hyperparameter tuning using both random search and manual tuning. Additionally, the study analyzed the linguistic characteristics of AI-generated reviews, a less explored area in existing studies.

This study has made valuable contributions to both academia and practical applications. For the academic community, it provided novel insights into the linguistic characteristics of AI-generated reviews. Unlike existing studies focusing on fine-tuning existing models, this study presented a novel approach by combining vectorization methods with ML algorithms and text-based features. Our study shows

we should move beyond purely linguistic indicators and include additional features (e.g., readability metrics, sentiment scores, and contextual variables). Our findings demonstrate that specific linguistic attributes can still be valuable, particularly when combined with a broader set of text-centric or behavioral factors. In other words, our results neither dismiss the usefulness of linguistic insights nor minimize their relevance. Instead, they highlight the need for an expanded, multifaceted approach capable of detecting increasingly sophisticated AI-generated reviews.

Additionally, the study focused on the hotel industry, a new application, as most studies focus on product reviews. Another significant contribution of this study is creating and making publicly available a dataset with genuine and AI-generated reviews (available for download at <not show in this version for anonymity>), a valuable resource for future research on the topic. This study established a framework for future research on AI-generated review detection for businesses and practitioners. It contributed to combatting fraudulent online reviews, as integrating these detection methods into online platforms or hotel management systems (e.g., social reputation monitoring systems) would offer a broader impact and protect users directly.

Despite the results achieved, there is room for improvement, and there are limitations that future studies can address. Although this study provides valuable insights by examining descriptive and visual differences between AI-generated and genuine reviews, we acknowledge that we did not conduct formal statistical significance tests (e.g., t-tests or non-parametric tests) to validate these observed discrepancies. Further studies refine and confirm the robustness of our detection approaches with these tests. Also, the reliance on a smaller dataset may have influenced the results, with simpler techniques achieving better performance compared to more complex ones. Additionally, small datasets can lead to overfitting, especially in more complex or varied real-world scenarios. This is particularly important given that state-of-the-art models, such as large transformers, typically require tens of thousands of diverse examples to achieve robust and generalizable performance (Chakraborty et al. 2024). Furthermore, reliance on a single domain may bias the model towards domain-specific features, reducing its effectiveness when applied elsewhere. To mitigate these limitations, future research should prioritize collecting and using a larger, more diverse dataset that spans multiple domains, and even combine the dataset created in this paper with others from multiple sectors.

Spammers may try to make AI-generated reviews appear more human by manually modifying them. Therefore, exploring combined detection methods for spotting both AI-generated and human-written fraudulent reviews would be beneficial. Additionally, a growing challenge lies in the detection of partially AI-assisted reviews, where genuine users employ AI tools to enhance or rewrite their own content. This hybrid form complicates the binary distinction between human and AI-generated text and should be taken into account in future research. It would also be helpful to analyze datasets in languages other than English to improve the effectiveness of these detection methods. Future research should not rely solely on vectorization methods for detecting AI-generated content. Incorporating text-based features into models, as well as using feature selection methods, could enhance model performance.

One other limitation is that the model could mistakenly flag legitimate reviews that have merely been refined with AI tools for grammar or style. While our study focuses

on AI-generated content that intentionally aims to deceive, we acknowledge the need to refine future models to distinguish between partial AI assistance and fraudulent, fully AI-produced reviews. Implementing more nuanced detection criteria or transparency mechanisms - such as user disclosures regarding AI use - could mitigate the risk of penalizing well-intentioned reviewers.

Our evaluation centered on widely adopted metrics (accuracy, precision, recall, and F1), which facilitate comparability with prior studies. However, future work could explore false positive and false negative rates to assess the real-world consequences of misclassification. These additional analyses are particularly relevant if AI-assisted reviews - intended primarily for text refinement - are inadvertently flagged as fraudulent. Moreover, classical Machine Learning and improved feature engineering may not be enough to capture the full breadth of AI-generated patterns that modern deep learning architectures like transformers could uncover. For this reason, future research might incorporate models such as BERT or GPT-based detectors, which inherently learn text representations and may further enhance detection accuracy. This line of research could also refine our understanding of how hand-engineered and automatically extracted features contribute to distinguishing AI-generated content from genuine user reviews.

From an applied standpoint, review platforms face the challenge of distinguishing genuinely deceptive AI-generated content from AI-assisted text that is not intended to mislead. Over-regulation risks discouraging legitimate reviewers who rely on AI for grammar checks or language enhancement. Therefore, any practical implementation of our detection model should incorporate transparent user policies and, where feasible, user disclosures indicating partial AI assistance. Platforms can maintain trust without unduly penalizing honest contributors by clarifying acceptable AI usage and focusing on overtly deceptive practices. Future research might explore methods to automate this differentiation, balancing business needs, user satisfaction, and the imperative to curb fraudulent reviews.

By addressing these limitations and future research directions, advancements can be made in AI-generated review detection, providing more robust solutions for online platforms.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40558-025-00329-z.

Author contributions A.S. created the dataset, performed data analysis, modeled it, and wrote the first draft of the manuscript. N.A. provided conceptualization, validation, and supervision. Both authors contributed to the literature review, interpretation of results, manuscript revision, and final manuscript approval.

Funding Open access funding provided by FCT|FCCN (b-on). This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under the project - UIDB/04152 - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS.

Data availability The dataset used in the research is going to be shared publicly here: https://doi.org/10.17605/OSF.IO/G6QZD.

Declarations

Competing interests The authors declare no competing interests.

References

Abubakar HD, Umar M, Bakale MA (2022) Sentiment classification: review of text vectorization methods: bag of words, Tf-Idf, Word2vec and Doc2vec. SLU J Sci Technol 4(1):27–33. https://doi.org/10.56471/slujst.v4i.266

Antonio N, De Almeida AM, Nunes L, Batista F, Ribeiro R (2018) Hotel online reviews: creating a multi-source aggregated index. Int J Contemp Hospitality Manage 30(12):3574–3591. https://doi.org/10.1108/IJCHM-05-2017-0302

Au TWT, Lampos V, Cox I (2022) E-NER — An annotated named entity recognition Corpus of legal text. Proc Nat Legal Lang Process Workshop 2022 246–255. https://doi.org/10.18653/v1/2022.nllp-1.22

Baek H, Ahn J, Choi Y (2012) Helpfulness of online consumer reviews: readers' objectives and review cues. Int J Electron Commer 17(2):99–126. https://doi.org/10.2753/JEC1086-4415170204

Bansal S, Aggarwal C (2019) *textstat: Calculate statistical features from text* (0.7.3) [Python]. https://github.com/shivam5992/textstat

Bao G, Zhao Y, Teng Z, Yang L, Zhang Y (2024) *Fast-DetectGPT: efficient Zero-Shot detection of Machine-Generated text via conditional probability curvature*. OpenReview. https://openreview.net/forum?id=Bpcgcr8E8Z

Bathla G, Singh P, Singh RK, Cambria E, Tiwari R (2022) Intelligent fake reviews detection based on aspect extraction and analysis using deep learning. Neural Comput Appl 34(22):20213–20229. https://doi.org/10.1007/s00521-022-07531-8

Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. J Mach Learn Res 13:281–305. https://doi.org/10.5555/2188385.2188395

Berrar D (2019) Cross-Validation. In *Encyclopedia of Bioinformatics and Computational Biology* (pp. 542–545). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Bird S, Klein E, Loper E (2009) Natural Language processing with python: analyzing text with natural Language toolkit. O'Reilly Media, Inc. https://www.oreilly.com/library/view/natural-language-processing/9780596803346/

Chakraborty S, Bedi A, Zhu S, An B, Manocha D, Huang F (2024) *Position: on the Possibilities of AI-Generated Text Detection*. PMLR. https://proceedings.mlr.press/v235/chakraborty24a.html

Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Collinson P (2023), July 15 Fake reviews: Can we trust what we read online as use of AI explodes? *The Guardian*. https://www.theguardian.com/money/2023/jul/15/fake-reviews-ai-artificial-intelligence-hotels-restaurants-products

Corizzo R, Leal-Arenas S (2023) One-Class learning for AI-Generated essay detection. Appl Sci 13(13):7901. https://doi.org/10.3390/app13137901

Dessı D, Helaoui R, Recupero DR, Riboni D (2021) *TF-IDF vs Word Embeddings for Morbidity Identification in Clinical Notes: An Initial Study*. https://doi.org/10.48550/arXiv.2105.09632

DuBay WH (2004) The principles of readability. ERIC Clgh. http://files.eric.ed.gov/fulltext/ED490073.pdf

Duma RA, Niu Z, Nyamawe AS, Tchaye-Kondi J, Yusuf AA (2023) A deep hybrid model for fake review detection by jointly leveraging review text, overall ratings, and aspect ratings. Soft Comput 27(10):6281–6296. https://doi.org/10.1007/s00500-023-07897-4

Fong LHN, Ye BH, Leung D, Leung XY (2022) Unmasking the imposter: do fake hotel reviewers show their faces in profile pictures? Annals Tourism Res 93:103321. https://doi.org/10.1016/j.annals.2021.103321

Gambetti A, Han Q (2023) *Combat AI With AI: Counteract Machine-Generated Fake Restaurant Reviews on Social Media* (arXiv:2302.07731). arXiv. http://arxiv.org/abs/2302.07731

Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63(1):3–42. https://doi.org/10.1007/s10994-006-6226-1

Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for Cancer classification using support vector machines. Mach Learn 46(1–3):389–422. https://doi.org/10.1023/A:1012487302797

Hajek P, Hikkerova L, Sahut J-M (2023) Fake review detection in e-Commerce platforms using aspect-based sentiment analysis. J Bus Res 167:114143. https://doi.org/10.1016/j.jbusres.2023.114143

Hans A, Schwarzschild A, Cherepanova V, Kazemi H, Saha A, Goldblum M, Geiping J, Goldstein T (2024) *Spotting LLMs with Binoculars: Zero-Shot Detection of Machine-Generated Text*. arXiv.org. https://arxiv.org/abs/2401.12070

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York. https://doi.org/10.1007/978-0-387-84858-7

Honnibal, Montani (2017), January 1 *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. Sentometrics Research. https://sentometrics-research.com/publication/72/

Hunter JD (2007) Matplotlib: A 2D graphics environment. Comput Sci Eng 9(3):90–95. https://doi.org/10.1109/MCSE.2007.55

Inc. PT (2015) *Collaborative data science. Montreal, QC: Plotly Technologies Inc. Retrieved* from https://plotly.com/

Integral Ad Science (2024) *Fraud in Generative AI: A deep dive into how Gen AI affects marketers*. Retrieved from https://integralads.com/insider/fraud-generative-ai-marketers/

Islam N, Sutradhar D, Noor H, Raya JT, Maisha MT, Farid DM (2023) *Distinguishing Human Generated Text From ChatGPT Generated Text Using Machine Learning* (arXiv:2306.01761). arXiv. http://arxiv.org/abs/2306.01761

Jawahar G, Abdul-Mageed M, Lakshmanan VS, L (2020) Automatic detection of machine generated text: A critical survey. Proc 28th Int Conf Comput Linguistics 2296–2309. https://doi.org/10.18653/v1/2020.coling-main.208

Knoedler S, Sofo G, Kern B, Frank K, Cotofana S, Von Isenburg S, Könneker S, Mazzarone F, Dorafshar AH, Knoedler L, Alfertshofer M (2024) Modern machiavelli?? The illusion of ChatGPT-generated patient reviews in plastic and aesthetic surgery based on 9000 review classifications. J Plast Reconstr Aesthetic Surg 88:99–108. https://doi.org/10.1016/j.bjps.2023.10.119

Kumar R, Mukherjee S, Rana NP (2023) Exploring latent characteristics of fake reviews and their intermediary role in persuading buying decisions. Inform Syst Front. https://doi.org/10.1007/s10796-023-10401-w

Lee KD, Han K, Myaeng S-H (2016) Capturing Word Choice Patterns with LDA for Fake Review Detection in Sentiment Analysis. *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics*, 1–7. https://doi.org/10.1145/2912845.2912868

Li Y, Li Q, Cui L, Bi W, Wang Z, Wang L, Yang L, Shi S, Zhang Y, University Z Westlake University, The University of Hong Kong, Jilin University, & Tencent AI lab. (n.d.). MAGE: Machine-generated Text Detection in the Wild. *Abstract*. https://aclanthology.org/2024.acl-long.3.pdf

Lomonosov MSU, Maloyan N, Nutfullin B, Lomonosov MSU, Ilyshin E, Lomonosov MSU (2022) DIALOG-22 RuATD generated text detection. Comput Linguistics Intellect Technol 394–401. https://doi.org/10.28995/2075-7182-2022-21-394-401

Loria S (2018) Textblob Documentation. *Release 0.15*, *2*

Luca M (2011) Reviews, reputation, and revenue: the case of yelp.com. SSRN Electron J. https://doi.org/10.2139/ssrn.1928601

Luo J, Nan G, Li D, Tan Y (2023) AI-Generated review detection. SSRN Electron J. https://doi.org/10.2139/ssrn.4610727

Masala M, Ruseti S, Dascalu M (2020) RoBERT– A Romanian BERT model. Proc 28th Int Conf Comput Linguistics 6626–6637. https://doi.org/10.18653/v1/2020.coling-main.581

Maurya SK, Singh D, Maurya AK (2023) Deceptive opinion spam detection approaches: A literature survey. Appl Intell 53(2):2189–2234. https://doi.org/10.1007/s10489-022-03427-1
McKinney W (2010) Data structures for statistical computing in Python. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a
Moon S, Kim M-Y, Iacobucci D (2021) Content analysis of fake consumer reviews by survey-based text categorization. Int J Res Mark 38(2):343–364. https://doi.org/10.1016/j.ijresmar.2020.08.001
Muthukrishnan R, Rohini R (2016) LASSO: A feature selection technique in predictive modeling for machine learning. 2016 IEEE Int Conf Adv Comput Appl (ICACA) 18–20. https://doi.org/10.1109/ICACA.2016.7887916
Oesper L, Merico D, Isserlin R, Bader GD (2011) WordCloud: A cytoscape plugin to create a visual semantic summary of networks. Source Code Biol Med 6(1):7. https://doi.org/10.1186/1751-0473-6-7
Ott M, Choi Y, Cardie C, Hancock JT (2011) Finding Deceptive Opinion Spam by Any Stretch of the Imagination. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technology*
Ott M, Cardie C, Hancock JT (2013) Negative Deceptive Opinion Spam. *In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D (2012) Scikit-learn: machine learning in Python. J Mach Learn Res 12. https://doi.org/10.48550/arXiv.1201.0490
Pennington J, Socher R, Manning C (2014) Glove: global vectors for word representation. Proc 2014 Conf Empir Methods Nat Lang Process (EMNLP) 1532–1543. https://doi.org/10.3115/v1/D14-1162
Pothuganti S (2018) Review on over-fitting and under-fitting problems in machine learning and solutions. Int J Adv Res Eletrical Eletronics Instrum Eng 7(9). https://doi.org/10.15662/IJAREEIE.2018.0709015
Powers D (2008) Evaluation: from precision, recall and F-measure to roc,iinformedness, markedness & correlation. Int J Mach Learn Technol 2:37–64. https://doi.org/10.48550/arXiv.2010.16061
Rehurek R, Sojka P (2011) *Gensim—Statistical Semantics in Python*. https://api.semanticscholar.org/CorpusID:64026679
Richardson L (2019) Beautiful soup documentation. *April*
Robertson S (2004) Understanding inverse document frequency: on theoretical arguments for IDF. J Doc 60(5):503–520. https://doi.org/10.1108/00220410410560582
Regular expression Rossum (2020) re—Regular expression operations. Python Doc. https://docs.python.org/3/library/re.html
Ruan N, Deng R, Su C (2020) GADM: manual fake review detection for O2O commercial platforms. Computers Secur 88:101657. https://doi.org/10.1016/j.cose.2019.101657
Salminen J, Kandpal C, Kamel AM, Jung S, Jansen BJ (2022) Creating and detecting fake reviews of online products. J Retailing Consumer Serv 64:102771. https://doi.org/10.1016/j.jretconser.2021.102771
Satapathy R, Pardeshi SR, Cambria E (2022) Polarity and subjectivity detection with multitask learning and BERT embedding. Future Internet 14(7):191. https://doi.org/10.3390/fi14070191
Budhi GS, Chiong R, Wang Z, Dhakal S (2021) Using a hybrid content-based and behaviour-based featuring approach in a parallel environment to detect fake reviews. Electron Commer Res Appl 47:101048. https://doi.org/10.1016/j.elerap.2021.101048
Shah A, Ranka P, Dedhia U, Prasad S, Muni S, Bhowmick K (2023) Detecting and unmasking AI-Generated texts through explainable artificial intelligence using stylistic features. Int J Adv Comput Sci Appl 14(10). https://doi.org/10.14569/IJACSA.2023.01410110
Sueno HT (2020) Multi-class document classification using support vector machine (SVM) based on improved Naïve Bayes vectorization technique. Int J Adv Trends Comput Sci Eng 9(3):3937–3944. https://doi.org/10.30534/ijatcse/2020/216932020
Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J Royal Stat Soc Ser B: Stat Methodol 58(1):267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
Turing et al (2023), April 12 *Word embeddings in NLP: A Complete Guide*. https://www.turing.com/kb/guide-on-word-embeddings-in-nlp
Vanroose P (2004) Part-of-Speech tagging from an Information—Theoretic point of view. Katholieke Universiteit Leuven
Vidanagama DU, Silva TP, Karunananda AS (2020) Deceptive consumer review detection: A survey. Artif Intell Rev 53(2):1323–1352. https://doi.org/10.1007/s10462-019-09697-5

Wang Z, Chen Q (2020) Monitoring online reviews for reputation fraud campaigns. Knowl Based Syst 195:105685. https://doi.org/10.1016/j.knosys.2020.105685

Wang Y, Pan Y, Yan M, Su Z, Luan TH (2023) A survey on chatgpt: AI–Generated contents, challenges, and solutions. IEEE Open J Comput Soc 4:280–302. https://doi.org/10.1109/OJCS.2023.3300321

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, Von Platen P, Ma C, Jernite Y, Plu J, Xu C, Le Scao T, Gugger S, Rush A (2020) Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

You can't tell whether an online restaurant review is fake—But this AI can. (n.d.). Retrieved 29 December 2023, from https://phys.org/news/2018-09-online-restaurant-fakebut-ai.html

Zhang KZK, Zhao SJ, Cheung CMK, Lee MKO (2014) Examining the influence of online reviews on consumers' decision-making: A heuristic–systematic model. Decis Support Syst 67:78–89. https://doi.org/10.1016/j.dss.2014.08.005