**Communications in Mathematics and Applications**
Vol. 16, No. 1, pp. 375–404, 2025
ISSN 0975-8607 (online); 0976-5905 (print)
Published by RGN Publications
DOI: 10.26713/cma.v16i1.3310

http://www.rgnpublications.com

Research Article

# A Multifaceted Novel Approach to Identify Deceptive Reviews Based on Psychology Theories: A New Dataset

Abeer Hassan Asiri* and Fahad Mazaed Alotaibi

*Faculty of Computer and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia*
***Corresponding author:** aassiry@kku.edu.sa

**Received:** September 5, 2024 **Revised:** November 19, 2024 **Accepted:** December 6, 2024

**Abstract.** As deception negatively impacts various areas, deception detection is an important field of study. This paper introduces a framework to detect online review deception. It studied aspects of deceptive reviews, considering the complex nature of deception in textual data and the low chance of direct detection. Furthermore, the paper presents a new corpus for deceptive reviews labeled using deception hints. This dataset was compiled in English and extracted from Google Maps reviews. We focused on the reviews of "restaurants" in New York. The novelty of this dataset is that the truthful and deceptive reviews were not deliberately collected; that is, participants were not requested to write lies, but the texts were collected after the individuals had written them. We used predefined criteria using deception indicators to differentiate deceptive and truthful reviews for this dataset. Each suggested indicator is not a definitive indicator on its own, but we assessed review authenticity using a set of indicators together. This paper aims to discuss lie detection strategies and theories, design a theory-based framework to detect deception in online reviews and implement the framework on a real-world dataset to provide a foundation for future empirical research and practical applications. The experimental results obtained from our labeled benchmark dataset showcase the effectiveness of this approach.

**Keywords.** Deception, Framework, Lying detection, Deception indicators, Online reviews

**Mathematics Subject Classification (2020).** 68T27, 68T42, 91A80

## 1. Introduction

Deceptive reviews refer to the services and product review content that has been forged intentionally to manipulate reader information. These manipulated reviews can be generated through humans or bots. Deception detection in reviews refers to the method of determining if a specific review contains deception or not (Carson [15]). Deception detection in textual data is a challenging task because it misses many non-verbal cues such as facial and voice cues (Vrij *et al*. [62]). Services and product online reviews are one of the most critical and common textual-based communication in e-commerce and one of the easiest deception methods. In general, frauds are thought to cause physiological and behavioral changes in the deceiver, producing clear signs of deception (Carter *et al*. [16]). Identifying deceptive reviews is a complex task for a variety of reasons, such as the huge number of reviews posted on popular platforms, the diverse types of reviews, the complex methods used by those who create fake reviews, and the challenge of distinguishing true reviews that may seem true. Deception theories are psychological theories that try to understand how people behave when deceptive. Those theories proposed that there are verbal and non-verbal signs that can be used to detect deception (Carson [15]). Typically, the various cue sources are referred to as modalities or indicators. For example, indicators detected in the language of a deceiver are considered to originate from the speech channel or to be part of the speech modality. Self-adaptor movements, such as touching one's own body, face, or hair, are examples of non-verbal signs. Both verbal and nonverbal indicators can be used to identify deception using AI techniques, depending on how the deceiver talks and behaves (Carter *et al*. [16]).

Many challenges face researchers in this field such as the continuous changes in deception tactics, the unclarity of features of the deceptive text, and the low quality of the available corpus (Desale *et al*. [19]). The problem of identifying deceptive online reviews has attracted significant interest from researchers, given the commonness of review platforms as a popular medium for these reviews. By nature of the deceptive online reviews, the focus is on verbal lie detection using linguistics data, which is the most precious part besides the behavioral data (Vrij *et al*. [62]).

Current proposed methods for identifying deceptive reviews lack comprehensive frameworks and datasets that capture the diverse and evolving psychological characteristics of deception tactics. While psychological traits can affect the performance of detection cues, prior research on deception detection did not exploit them. In this study, we suggested a set of deception indicators based on deception theories with the support of early studies on deceptive reviews. Then we introduced a deception detection framework that assesses deception probability in reviews based on review-centric features, reviewer-centric features, and metadata. Based on the framework we concluded a set of annotation guidelines used to annotate a new dataset. In this work, we introduced a new labeled reviews dataset collected from Google Maps reviews. The dataset is labeled into two classes, 0 = truthful and 1 = deceptive. Predefined criteria are constructed and used to label the dataset. We developed a restaurant reviews dataset called "RRD" that contains 21476 restaurant reviews.

The rest of the paper is structured as follows: Section 2 introduces the research problem and objectives, Section 3 discusses the research methodology, Section 4 discusses some of the related previous works, Sections 5 discuss some of the deception theories and deception indicators in reviews, Section 6 introduces the details of the proposed framework, Section 7 provides the process of dataset creation as framework implementation, and followed by an explanation of the dataset statistics, validation, and benchmarking. Lastly, we discussed limitations and concluded the paper in sections 9 and 10.

## 2. Research Problem and Objectives

Existing deception detection methods mainly depend on analyzing context-based information. In early studies, truthful review datasets were collected from review websites directly without further analysis and classifications (a well-known problem in this field). In this way, we cannot precisely distinguish truthful reviews from deceptive ones that were prompted by vendors or competitors and posted on these websites (Li *et al*. [36]). We noted that most of the previously annotated datasets were generated by telling participants to intentionally deceive and represent intentional (rather than unintentional) deception. Thus, models trained using crowd sourced generated deceptive reviews are not effective in detecting real-life deceptive reviews on real commercial websites as the detection accuracies are a near chance (Mukherjee *et al*. [44]). So that data misses the real behavior of the deceiver that we try to analyze based on deception theories. This research seeks to fill this gap by analyzing deception theories to generate an effective framework that accurately classifies deceptive and truthful reviews. This work aims to:

- Identify features of the deceptive reviews based on Psychology theories (deception theories) and early studies.
- Provides a new framework to improve deceptive review identification as a basis for further empirical research.
- Introduce a new labeled dataset using the proposed framework.

## 3. Methodology

In this research, we will use a mixed research approach combining a qualitative analysis of literature and available data (Creswell [18]). We will analyze and select deception indicators from deception theories and formulate the framework for detecting deceptive reviews. This will guide us in achieving our first and second research objectives. Then, we go through the experimental approach by implementing the suggested framework to annotate a new dataset. A dataset will be collected from the Google Maps restaurant reviews and labeled based on the framework. Finally, we will implement a traditional machine learning model on the dataset and evaluate it using common metrics. We will select the most suitable features of the deceptive reviews based on deception theories with the support of early research. We will analyze the popular and suitable deception theories that provide deception features in written text. These features will be used in experimental design as follows:

- Develop a comprehensive rule-based framework.

- Implement the framework to extract guidelines to label the dataset.
- Create a new dataset and evaluate it using traditional ML models.

## 4. Literature Review

Research on deceptive reviews detection is a recently developed field of study. Despite that, researchers have designed many methods, the most recent being *Machine Learning* (ML) and *Deep Learning* (DL) techniques (Bathla *et al*. [11], Mohawesh *et al*. [41], and Prome *et al*. [49]). Those intelligent techniques have become the most popular techniques for identifying fraudulent data in recent years. Most of the literature found that supervised methods are the most popular method for detecting deceptive reviews due to the high accuracy provided (Alsubari *et al*. [6], and Prome *et al*. [49]). Three types of features are mainly used for deceptive review detection: text-centric, writer-centric, and other metadata-centric features. First, text-centric focuses on the textual content of the review according to methods such as Bag-of-Words, word frequency, $n$-grams, skip-grams, and word count. Second, writer-centric features, describe user information, their connections, behavior, and timestamps. Finally, the metadata features depend directly on other relevant information such as location data, product or services, type of platform, and appropriate media (Alaskar *et al*. [4], and Li *et al*. [33]). Le and Kim [29] suggested using verbal and non-verbal indicators as features of the reviews dataset to identify deceptive reviews. Verbal cues include the characteristics of the text body of the review, while non-verbal cues include other relevant data such as the posted rating of this review, the number of reviewer friends, and the business rate. The work compared review patterns using classification techniques and applied sentiment analysis to analyze the content of reviews. The result of this work shows the importance of considering both verbal and non-verbal cues. Abdulqader *et al*. [1] suggested using a set of deception theories to detect deception verbal and non-verbal features. This work combines ten theories to generate features and implement classical machine learning models using the Yelp review dataset. This work suggested for future improving the detection process using DL techniques and datasets to test the result of the study. Mohawesh *et al*. [41] introduced an explainable Multiview deep learning model using the product reviews dataset. The proposed model extracts the features of the review text, reviewer data, and product description. It identifies the deceptive reviews and explains why some reviews are classified as deceptive. Preliminary research about fabricated reviews showed how to construct a classification model for detecting whether the text is misleading (Le and Kim [29]).

In the last four years, a significant body of work has proposed methods for deceptive review detection. The work covered a range of methods, including pattern recognition and natural language processing (Li *et al*. [33]). These methods generally depend on labeled datasets. The dataset is the backbone and the key point for training and evaluating the models applied in deception detection. A labeled dataset is essential, and in the case of deceptive reviews, this might be challenging, if not impossible, to get. It is evident from the literature we reviewed that the majority of the datasets that have been utilized in previous studies were produced intentionally, most likely as a result of the scarcity of deceptive review samples and the challenge of accurately labeling them (Abedin *et al*. [2]). The content duplication identification technique

was one of the earliest methods used to identify deceptive reviews. The study of Jindal and Liu [27] was one of the earliest studies in this field. They prepared the dataset from Amazon reviews, and they labeled any repeated reviews as fake. Also, Lin *et al*. [37] used the Jaccard distance method to calculate similarity among review text, and they considered duplicated and near-duplicated reviews deceptive. This method focuses only on similarity and duplication to detect deception. This method overlooked other important deception aspects. So, this annotation approach needs to be improved (Mohawesh *et al*. [42]).

Another type of research suggested hiring crowdsourcing to create artificially deceptive reviews and collect data from opinion websites to represent true reviews such as the dataset prepared by Li *et al*. [33]. They collect their true dataset from TripAdvisor and fake reviews prepared by crowdsourcing. Yoo and Gretzel [66] also collected 40 reviews from TripAdvisor as true reviews and asked marketing students to write deceptive reviews. This method of annotation called "pseudo fake" is not accurate enough to catch and describe real-life review deception cases. Moreover, it requires a lot of manpower and Turkers' behavior is not similar to real fakers, there were some inaccurate labels in this kind of dataset, as it was distinct from a real-life dataset as reported by Mukherjee *et al*. [43].

A set of works tried to address this gap using a rule-based dataset to label deceptive reviews.The method based on rules mostly does not depend on manual annotations, which makes the annotations cost relatively low. Annotation using this method is simple but contains some noise. Li *et al*. [32] used some rules to identify deceptive reviews and used human judgment to label the reviews. They invited three undergraduate students to annotate the dataset, and they considered the review deceptive when two out of three annotators supposed that the review was deceptive. Gryka and Janick [21] created a new dataset using Google Maps reviews in Polish and they trained it using the ML model. They detected fake accounts and fake reviews using a set of rules and then trained ML models on the dataset. The model achieved an F1 score of 0.92 when identifying deceptive accounts and about 0.74 for deceptive reviews. The work introduced by Asghar *et al*. [8] They used a spam score to measure deception features in Amazon product reviews. They used a hyperfeatures model to classify the reviews and prioritize the features using a revised feature weighting scheme. Also, Alsubari *et al*. [5] have used the same idea when they defined the Authenticity Score to measure all text features of the Yelp reviews dataset. They considered reviews with a score greater than or equal to 49 % trusted, and the deceptive ones had an authenticity score less than 48.5%. Shahariar *et al*. [51] have created a review dataset of the Belgian language using rule-based techniques. The dataset consisted of 7710 truthful and 1339 fake reviews. Then, they performed a set of experiments on the dataset. They got about 98% for the F1 score.

Table 1 compares studies done in the last ten years and the datasets built using the rule-based method. There is currently no dataset available for deception detection that has been developed using a rule-based psychological framework. All the studies constructed rules based on the researchers' point of view for deception indicators such as mentions of brand names or poor grammar. Hence, our motivation is to propose a framework based on deception theories and construct a dataset for detecting deceptive reviews using the proposed framework.

**Table 1.** Summary of literature that suggested rule-based datasets

| Year | Measurable score | Equation | Explanation | Rule (1 = deceptive, 0 = truthful) | Dataset source | Features | Ref. |
|---|---|---|---|---|---|---|---|
| 2014 | Spamicity | $S = \frac{\sum_1^{i=n} f \times w}{n \times m}$ | They extracted behavioral and linguistic features of reviews, reviewers, and groups.<br>They created a vector of features to label the dataset.<br>They assign weight to each feature and calculate the probability of review spamicity. | If ($S > 0.9$)?<br>label = 1,<br>label = 0 | Amazon products reviews | Review, Reviewers, groups | [53] |
| 2015 | Counter | No | They implemented a framework containing a set of fake hints.<br>A set of 16 questions about reviews and reviewers was used as a rule to classify the dataset.<br>If the counters are greater than the threshold (they set it to 5), the review label will be spam. | For (answer = yes), counter+1<br>If (counter > 5)? 1 : 0 | Amazon products reviews | Review, Reviewers | [54] |
| 2016 | No | No | They used seed words to identify the reviews that might be fake.<br>They removed the reviews with a rating of less than 4.<br>Manually check if the review is deceptive or not. | No | Amazon products reviews | Review Features | [57] |

Table Contd.

| Year | Measurable score | Equation | Explanation | Rule (1 = deceptive, 0 = truthful) | Dataset source | Features | Ref. |
|---|---|---|---|---|---|---|---|
| 2020 | Authenticity score | $A(r) = \sum$ FPS + TPS + TPP + excl (differ) − Negemo − Motion | They calculated a set of review text features based on the LIWC.<br>They found the range of output values is from 1 to 100.<br>They supposed that the truthful reviews were more than or equal to a 49 % score, and the fake ones had a score of less. | If (score <= 50%)? 1 : 0 | Yelp products reviews | Review features | [5] |
| 2020 | Spamicity score | Spamicty_score = $\sum$ Fi × Wi ($F$ = features, $W$ = weights) | They consider a set of review, reviewer, and product features.<br>Assign weight for each feature.<br>Calculated the spam score and normalized it. | If (spamicity_score ≤ 0.5)? 1 : 0 | Amazon products reviews | Review, reviewer, and product features | [8] |
| 2023 | No | No | They set some rules to identify fake accounts and fake reviews then they implemented the rules to label the dataset.<br>Manually they label and test the dataset.<br>No measurable variable was used in this study. | No | Google Maps reviews | Account features, review features. | [21] |
| 2024 | No | No | They set rules to label the dataset.<br>They annotate the dataset based on the rules manually with three human annotators.<br>If the review falls into any of the criteria, they label it FAKE. | No | Food review from Facebook and YouTube | Review features | [51] |

## 5. Deception Theories and Techniques

Deception theories are scientific theories that provide insights into the psychological and linguistic features and aspects of deceptive behavior. These theories try to identify deception features that can be used for automated detection. It can be defined as a set of rules that attempt to discuss how people deceive and how deception can be detected. These theories are an interdisciplinary topic that studies deception, its forms, detection, and psychological, social, and philosophical consequences (Nortje and Tredoux [47]). Implementing deception theories becomes essential in developing effective detection methods. It has been researched in a variety of situations, including law enforcement, psychology, communication, game theory, and artificial intelligence, to name a few (Vrij *et al*. [61]). For a long decade, the studies of deception detection theories shed light on the effects of communication channels or mediums. This type of research discusses which cues can be detected in each communication medium considering the text, audio, audiovisual presentation, and other forms of communication media such as email, chats, social networks, online reviews, etc. While reviews are considered written communication, reviews share many characteristics with spoken language, such as an informal tone and naturally generated phrase fragments. Also, unlike spoken language, reviews have the chance to be edited and a chance to think before writing with the absence of non-verbal deception indicators. Here, we will discuss the theories and techniques of deception that can be implemented for reviews.

### 5.1 Interpersonal Deception Theory

*Interpersonal Deception Theory* (IDT) was developed by David Buller and Judee Burgoon to study the interactions between the deceiver and the receiver of the messages. It examines the various verbal and nonverbal cues used in deception and how they affect the perception and detection of lies. IDT can be applied to analyze linguistic and textual cues that indicate deception in written communication. It studies changes in word choice, sentence structure, and the use of qualifiers or hedges to identify patterns of deception. IDT claims that deceivers are often detected by linguistic cues, such as using more hedging language, avoiding specific details, and making extreme claims (Burgoon and Buller [14]). This theory shows how the deceiver seeks to manipulate details in the language used and behavior. Furthermore, source authentication is one of the critical features considered in this theory (Thomas and Biros [58]). From the language side, this theory suggests that deceivers have fewer words, immediacy, and expressivity but have strong emotions, nervousness, uncertainty, pauses, and response latencies. The deceptive messages are usually brief, contain fewer ideas, and may utilize leveling terms or generalizations such as: ‘always’ or ‘everyone’. Also, they unintentionally use passive and past tense to reduce immediacy and group/others references more than self-references, which reflect no specificity and immediacy (Reddy and Motagi [50], and Wise and Rodriguez [64]). Furthermore, it investigates how deceptive reviewers manipulate information selectively by omitting details, providing biased information, or distorting facts to mislead readers.

### 5.2 Truth-Default Theory (TDT)

*Truth-Default Theory* (TDT) claims that most of the time, humans think that others communicate truthfully. The assumption of honest communication enables efficient contact

and cooperation. But people are naturally suspicious of others when they perceive others to be acting fraudulently, do not act honorably, or when the information in the communication seems contradictory, out of context, or complicated. TDT suggests that people are more likely to believe information that aligns with their existing beliefs or expectations. Deceptive reviews may exhibit language patterns that deviate from typical user reviews, such as excessive use of superlatives, lack of specific details or personal experiences, or an unnatural tone. These linguistic cues can indicate potential deception. Truth-default theory suggests assessing the emotional tone of the review and considering whether it aligns with what would be expected in a truthful review. It also emphasizes the importance of considering the behavior and credibility of the reviewer based on profile content. Look for patterns in the reviewer's behavior, such as a sudden influx of positive or negative reviews, many reviews falling within a narrow time frame, or a pattern of extreme ratings.

### 5.3 Information Manipulation Theory (IMT)

Information Manipulation Theory is one of the most significant new ideas of deception detection perspective theories, which moves the emphasis from non-verbal clues to the whole fraudulent message content and design (Ansari and Gupta [7]). This theory suggests misleading signals occur by subverting the rules that implicitly regulate conversations. These rules are Grice's maxims of conversation: quantity, quality, style, and information relevancy (Howard *et al*. [24]). The expected amount of pertinent information in a communication to make it informative is called its "quantity". The expected validity of information is referred to as quality. The desired avoidance of ambiguity is referred to as manner. Expected relevant information derived from a previous argument is referred to as relevance. IMT states that altering the data in a manner that violates one or more of Grice's four maxims indicates deceptive reviews. Low information in a message indicates that it will be less informative and likely to have a broad tone. This is referred to as a quantity violation (omission). Falsification, also known as quality violations, is the fabrication of the whole message to misrepresent the information. Equivocation, also known as manner violations, is the effort to conceal the truth using conventionally ambiguous language and indirect statements in place of clarity and direct speech. Relevance evasion, also known as relevance violations, deflects attention by offering irrelevant or omitting contextually significant information.

### 5.4 Self-Presentational Theory (SPT)

According to the self-presentational hypothesis, lying happens often in social contexts. They rejected the idea that lying is a difficult process that involves remorse. Rather, they contended that little behavioral leakage is left since most fraudulent presentations are carefully prepared and performed. According to this concept, behavioral control, cognitive load, and emotions all have an impact on how both truth-tellers and deceivers behave (Baumeister and Hutton [12]). According to the theory, people who lie seem more compliant, friendly, and nice than people who speak the truth. Compared to truth-tellers, deceivers are less open, which results in shorter statements with less details, depth, and information. Also, they are more nervous than truthtellers, and deceivers exhibit higher levels of stress and emotions (Marcus [40]).

This approach employed clues to gauge immediacy, such as negations in place of claims and passive voice in misleading communications. They discovered that con artists frequently employed the same words and phrases as they had a low level of fluency. In terms of believability, they discovered that misleading communications were more likely to reflect ambiguity or be internally contradictory.

### 5.5 Reality Monitoring Theory (RMT)

Reality Monitoring is a theory proposed by Charles Johnson and Raye in the early 1980s to explain how people distinguish between real and imagined experiences. It considers the verbal veracity assessment tool based on verbal criteria. It suggests that individuals distinguish between memories of real events and those of imagined or fabricated events by analyzing the sensory and contextual details associated with each type of memory (Vrij and Ganis [60]). Studies based on RM (Li *et al*. [35], and Sporer *et al*. [56]) discuss the differences between memories of imagined events and memories of genuine experiences. Perceptional processes let us retain memories of actual interactions. Because of this, memories of actual experiences include sensory details about taste, smell, vision, and sound, as well as contextual details, including physical characteristics like the event's location of the event and temporal data like the timing and length of the occurrences. When people tell true stories they provide various information, as proven by several empirical research. Also, real tales can be internally constructed with well-balanced sensory and contextual information, usually based on actual experiences and events. Conversely, false narratives could be devoid of these rich details, depend more on information that has been created or obtained from other sources, and frequently reuse terms and phrases as they do not reflect actual occurrences (Li *et al*. [33]).

### 5.6 Cognitive Load Theory (CLT)

Cognitive load is defined as the necessary labor needed to do a task, which can significantly affect the person's overall task success. It considers the mental effort required to process information and describes how it is difficult for liars to describe imaging events and details (Cranford *et al*. [17]). So, CL claims that real reviews often reflect personal experiences and provide detailed information. Deceptive reviews, on the other hand, may lack specific details or use generic language, to have a lower cognitive effort (Wielgopolan and Imbir [63]). This theory claims complex writings that reach in detail are likely to be considered more real. On the other hand, simpler materials might be seen as being less profound or serious, which could raise doubts about their veracity. It shows the greater cognitive effort that is often needed in text complexity if the writer lies, which makes the deceiver choose simpler words and content. Based on this technique, to detect deceptive reviews, we need to analyze the linguistic features of reviews, such as language style, coherence, and complexity, which can impose a cognitive load. Also, recognizing patterns associated with deceptive reviews, such as unnatural language, excessive use of superlatives, or overly positive or negative sentiment, requires cognitive effort.

### 5.7 Scientific Content Analysis (Scan)

SCAN is a technique developed by Sapir based on linguistic behavior used by individuals in deceptive communication forms. This method was developed for use in criminal investigations. It relies on analyzing transcripts or written remarks using criteria such as pronoun usage, spontaneous vocabulary, sentiments, and related phrases (Kuzio [28]). This technique is used to detect deception in various forms of communication, including fake news and reviews. It involves the analysis of written statements using specific linguistic and psychological indicators to determine the truthfulness or deception in the content. SCAN can identify patterns or linguistic cues that may appear in fraudulent reviews.

Some common indicators of deceptive reviews using scientific content analysis are the certain language patterns that distinguish deceptive reviews. For example, they may avoid talking about themselves and contain excessive use of superlatives, or generic statements. Analyzing the syntax, grammar, and vocabulary can provide insights into the authenticity of the review. SCAN assumes that changes in the use of pronouns and their omission within a statement is an alert that probable information manipulation may happen[46]. Moreover, it discussed the emotional content and expression where real reviews tend to have a balanced emotional structure, logical flow, and consistency. Deceptive reviews may lack consistency, contain contradictory statements, or display extreme emotion (Smith [55]).

### 5.8 Verifiability Approach (VA)

This approach shows how deceivers and truthtellers differ in their methods of telling the truth. Truthtellers embrace transparency and recount events truthfully, providing all relevant details in their narratives. This makes their language rich in vocabulary, unlike liars, who provide limited information where providing excessive detail would expose their deception to authorities (Nahari [45]).

Based on this principle true reviews are more likely to contain verifiable details and specific information about the reviewer's experience. VA shows how true reviews often include specific information about the service, or experience, product features, dates of interaction, or personal anecdotes. In contrast, manipulated reviews, on the other hand, may lack specific information and instead provide vague or generic statements. Real reviews tend to provide objective information that can be verified or confirmed. They may mention specific locations, prices, names of employees, or other factual details that can be cross-referenced. Fake reviews often tend to be brief, lack such objective information, and instead rely on general opinions or exaggerated claims. Furthermore, this approach, indicates the importance of checking the reviewer's account is verified or linked to a legitimate profile is important (Ioannidis *et al*. [25], Manaskasemsak *et al*. [39], and Verschuere *et al*. [59]).

### 5.9 Indicators of Deceptive Reviews Based on Deception Theories

By matching deception theories and previous empirical studies on online reviews, we derived a set of features indicative of manipulated content. Deception theories claim that deceivers have fewer words, immediacy, and expressivity but reflect more feelings of fear, agitation, confusion, pauses, and non-fluencies. For instance, theories show how deceivers use excessive

positive or negative language to manipulate the content (IDT). They claim the lack of detailed information in reviews is a deliberate tactic to obscure deception, corroborating vague and non-specific statements. The prevalence of overly generic statements in deceptive reviews can be explained by RM Theory, which posits that fabricated memories are less detailed than real ones. The use of emotional extremes further differentiates fake reviews from truthful ones, with the former displaying unnatural levels of emotion. The avoidance of self-references, a tactic highlighted in IDT, is another marker of deception. Contradictory statements, explained by SCAN Theory, are prevalent in fake reviews, reflecting the challenge of maintaining a consistent deceptive narrative. Also, the SCAN strategy indicates that as pronouns could mean accountability, the writer may attempt to relieve themselves of personal responsibility by omitting personal pronouns, especially 'I'. Cognitive Load Theory also accounts for the unnatural sentence structures found in deceptive reviews, where the effort to deceive impacts fluency. This affects the number of words used by the deceiver which may tend to be shorter than in real reviews.

The deception theories show how the deceivers tend to use fewer words and general language, avoid providing detailed information, or intentionally indistinct them to avoid implicating themself. Also, Fake reviewers repeated some words either spontaneously or to appear sincere. The assumption used is that true texts tend to have high levels of complexity. to sound convincing. We discussed the deception indicators that we selected for our framework in detail in Section 6. Table 1 summarizes the deception indicators we selected to identify review truthfulness, the deception theories that discussed them, and some examples of previous studies that prove these indicators in deceptive reviews.

**Table 2.** A summary description of deception indicators

| | Indicators | Description | Related studies | Related theories |
|---|---|---|---|---|
| I1 | Punctuation marks ratio (PMR) sentence structure | While punctuation marks are eliminated in some works as they do not make sense, many studies have explored the correlation between punctuation marks and deceptive writing. This feature is useful in text manipulation studies. Deception theories show how deceptive behavior is reflected in writing styles, such as heightened emotional states, persuasive attempts, and low sentence structure which can be represented in punctuation marks. | SCAN, CLT | [26], [54] |
| I2 | First-Person Singular Pronouns Ratio (FPSPR) | Deceptive reviews often use first-person singular pronouns less frequently. Many studies show that pronoun use can indicate information manipulation and that deceivers tend to use plural or third-person pronouns. In this work, we focus on first-person singular pronouns, and we recommend including other pronoun analyses in the future. | SCAN | [20], [55] |

(Table Contd.)

| | Indicators | Description | Related studies | Related theories |
|---|---|---|---|---|
| I3 | Review Length (RL)-Quantity | The relationship between the review length and its truth is significant and it has been used in deception detection research many times. The longer reviews are often associated with more information and greater detail that may show a real experience. The deception theories and earlier studies show how deceivers tend to use fewer words and avoid providing detailed information. So, extremely short reviews may indicate a manipulation attempt, as they may not be helpful or not provide enough information, and real reviews often provide context, details, and specific examples, which usually result in longer text. | SPT, IMT | [3], [6] |
| I4 | Repeated word Ratio (RWR)-Diversity | A real review reflects real experiences and events and is internally generated with balanced sensory and contextual details. In contrast, deceptive stories may have fewer words, lack rich details, and tend to repeat phrases, words, or actions to seem like reality. Many studies also detected this feature in fake review detection analysis. | SPT, RMT | [35],[60] |
| I5 | Sentiments (S) | Many theories show that deceptive reviews frequently contain excessive positive and negative sentiments. This high level of emotion in deceptive reviews is a tactic usually used by deceivers to affect others, show more emotions, nervousness, and arousal, and seem truthful. Furthermore, this feature is one of the most studied features in this field, and many studies proved this after analyzing the review dataset. | IDT, TDT, SCAN, SPT | [6], [22], [33], [50] |
| I6 | Generalization (G) | One important deceptive indicator is the high use of general language where the deceiver does not have details to add. Truthful content is more likely to include specific temporal, spatial, and perceptual content. General statements with high emotion are usually used by deceivers as they lack experiences to draw from. General statements contain terms like many, lots, most, generally, and commonly. | RMT, VA, IMT | [49], [65] |
| I7 | Passive Voice Ratio (PVR) | Passive voice is a grammatical construction where the object of an action becomes the subject of the sentence. In passive voice, the focus is on the action and the recipient of the action rather than the doer. Passive voice is often characterized using the form of the verb to be (is, was, were, etc.) followed by a past participle of the main verb. passive voice might be used to obscure the agent of an action, which can be relevant in detecting deceptive content. | SCAN, SPT | [22], [52] |

(Table Contd.)

| | Indicators | Description | Related studies | Related theories |
|---|---|---|---|---|
| I8 | Total reviewer reviews (TRR) | Since some reviews may be generated by bots, the user who posted the review may be newly registered, having created their deceptive account solely to leave a deceptive review. Detecting the number of reviews per user can be a useful feature to support other features. | SCAN, SPT, VA | [22], [38] |
| I9 | Account Type (AT) | The reviewer account type is also known as the platform user account type. It identifies whether the account is authenticated. Most platforms give the user "verified reviewers" for high-quality reviewers or users who have purchased items. Reviews of verified users are more trusted. | TDT, SPT, VA | [25], [39], [59] |
| I10 | Likes (L) | Likes on the review may indicate a true review where there is another voice supporting the opinion. | IMT | [31], [34] |
| I11 | Attached Media Count (AMC) | The reviews that contain media may indicate a higher probability of a true review. | IMT | [23], [48] |

## 6. Proposed Framework

Based on the concluded features, we developed a framework to detect deception containing a set of deception indicators. We designed the framework based on the idea that each deception indicator might not be alarming of deception by itself, but they may be helpful and supportive signs with the rest of the factors used to infer the credibility of the reviews.

The framework follows a set of main stages starting with dataset collection and preparation, data analysis, feature engineering, deception indicators computations, deception index calculation, and ending with the review's classification decisions. The decision of the review class is based on the score of the deception index, calculated based on the analysis and calculation of deception features. We divided the deception indicators into three main modules and followed the process of thresholding and Deception Index (DI) calculations as follows:

- *First Module*: This module is related to the deception indicators features of the review text. The main analysis tasks were designed to conclude the textual deception linguistic indicators. In our framework, we will study seven indicators related to deception language (Punctuation Ratios, Pronoun use (self-reference), Review Length (Quantity), Repeated Words Ratio, Sentiments and Emotions, Generalization, and Passive Voice Use(non-immediacy)).
- *Second Module*: This module represents the meta-data features attached to the reviews, such as spatial data, IP address, and Useful votes. In this work, we will consider only two features: Useful Votes (likes) and the Total Attached Media (video and images).
- *Third Module*: It includes features related to the reviewer, such as reviewer account information, reviewer friendship, Membership type, reviewer behavior information, and

total reviewer reviews. This work will focus only on membership type and total reviewer reviews.

- *Deception Index* (DI): According to the literature, calculating the deception score of online reviews is a challenging task since no single formula can accurately capture all deception aspects and it involves analyzing the content of the reviews and detecting potential deception or manipulation (Asghar *et al*. [8]). Furthermore, there is no universally accepted formula to calculate the reliability of online reviews, and determining the reliability score involves subjective judgment, analysis of different features of the review, and the reviewer (Alsubari *et al*. [5]). In this work, we suggested using measurable values that combine all features of deceptive reviews by an aggregated feature approach. After Extracting individual features from each review based on the identified deception hints, we normalize the features to ensure they are on a comparable scale. We transform the feature values to [0 or 1] indicators and combine the indicators into a single score (DI). Finally, we define a threshold value to classify reviews as deceptive or not based on the DI of potentially fraudulent activities.

These three modules complement each other to effectively indicate the credibility of reviews. Each feature will be calculated as feature engineering and normalized to represent them as measurable values. We will assign scores to these indicators to calculate the credibility score of the reviews to calculate the Deception Index (Asghar *et al*. [8]). The deception index will help us determine the deception's probability in review.

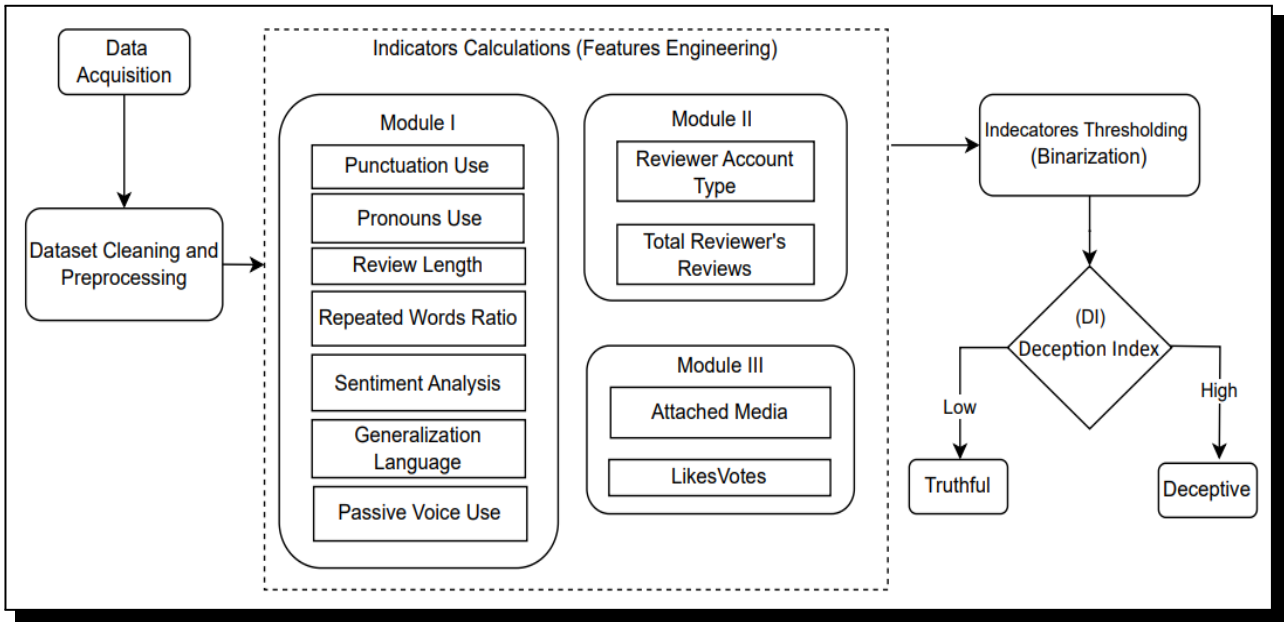


**Figure 1.** Suggested framework

Figure 1 shows the details of the suggested framework. We selected a set of features that are repeatedly used in previous studies and represent deception indicators in deception theories, and we used them to identify the review's authenticity. The framework supports defining the criteria guidelines of the dataset annotation process.

## 7. Framework Implementation

To implement the framework, we start with dataset creation as in Figure 2 which illustrates the dataset creation pipeline. Firstly, we select a source to scrape the dataset. Then, we perform dataset preprocessing to implement the framework modules. Finally, we annotate the dataset as deceptive or truthful.

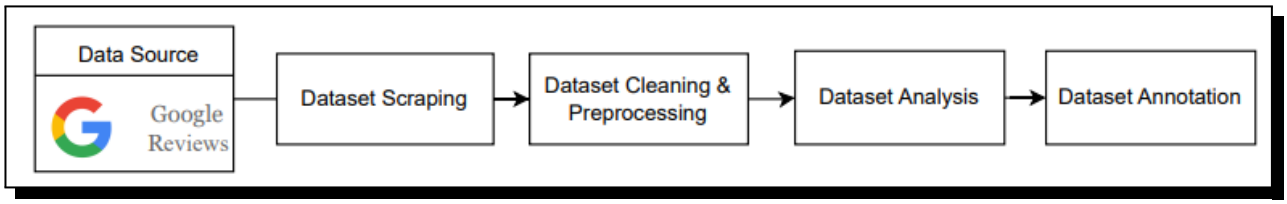


**Figure 2.** Dataset creation pipeline

### 7.1 Dataset Collection

In this work, we collect review datasets from Google Maps reviews, which enable users to rate and review any business. we selected Google Maps Reviews because according to reports [1].It is the most commonly used website for reading and publishing reviews. To scope our work, we targeted English reviews of restaurants (breakfast, lunch, and dinner restaurants) in the USA, New York. We targeted restaurants as they represent one of the biggest economically active sectors in the world and they contain a high number and various review types [44]. We scraped the data using the APIFY scraping tool[2]. We named the dataset RRD (Restaurants Reviews Dataset) and made it publicly available for the research community. We have scraped the reviews of the most popular restaurants there between 1 January 2019 and 1 July 2024. We targeted these years as Google has a great increase in the number of reviews left compared to other review platforms places reviewing through social media was very common during this period [13]. For privacy, we removed Images and URL links for reviewers, review links, and restaurant names. Initially, we got about 51920 records. During the process of data collection, we implement the following specific criteria:

- We collect different restaurant types to cover various linguistics reviews.
- We remove the records that do not contain review text.

### 7.2 Dataset Cleaning and Preprocessing

In this step, we employed several filters on the dataset to enhance consistency and minimize the workload. The following are the procedures used to process the texts:

- We removed the attributes that contain data unsuitable for our study such as postdate, owner response, and URL links.
- We removed emojis and special characters from the review body (after counting punctuation marks).

[1] S. Paget, Local consumer review survey 2023: Customer reviews and behavior, *BrightLocal*, accessed: January 22, 2024, available: URL: https://www.brightlocal.com/research/local-consumer-review-survey/.

[2] Apify, *Apify API*, accessed: January 29, 2024, available: URL: https://apify.com/.

- We count the number of linked media (images and videos) posted with the reviews.
- Duplicate review texts were removed.
- All the records with null values, non-English texts, and URL links are removed.
- We removed the records containing less than two words of review text, as they contained no valuable data.

Applying these filters, we eliminated valueless records and attributes. Our dataset dropped to 21476 records and an initial 5 attributes (review body, total reviewer reviews, review-likes, total images, and account type) that were subsequently passed on to the framework for the annotation process.

## 7.3 Dataset Analysis and Annotation

The following sections show the data analysis of the annotation procedure as pre-defined guidelines from the suggested framework. We calculated the features of reviews in each module, then we implemented averaged classification rules to perform dataset annotation (deceptive or truthful).

### 7.3.1 First Module Implementation

As the first module covers the linguistic characteristics of the review text, we implement it using NLP in Python. By the end of this module implementation, we have new attributes containing Review Length, Ratio of Repeated Words, Ratio of Punctuation Marks, Percent of Review Sentiments, Generalization Score, Informality, Passive Voice, Abstraction Level, And Pronoun Use Ratio (self or group reference).

**– Punctuation Marks Ratio (PMR)**

We start with counting the punctuation marks before the text preprocessing steps. Because in upcoming stages we will remove punctuation as an essential part of the NLP preprocessing. We calculated the ratio of punctuation regarding characters in the review. We used a constant `string.punctuation` from Python's `string` module which includes the following punctuation "#$%&'()*+,-./:;<=>?@[\]^_‘{|}~.

The result shows that the PMR range is 0%-37%. For PMR, we set a red flag for any misuse, including extremely high use. So, we set our red flag at the average which shows a poor structure of the text, a high tone, and superlative language.

**– First-Person Singular Pronouns Ratio (FSPR)**

At this step, we scoped our work to study First-person singular pronouns. We excluded first-person singular pronouns from stopping word removal and lemmatization during text preprocessing. We calculated the FSPR by dividing the number of first-person singular pronouns by the total word count in each review. We set the threshold to the average. If the ratio is lower than the average, we suppose the writer avoids it, and manipulation may occur.

**– Review Length (word count)**

At this step, we preprocessed the text to ensure consistent and accurate NLP results. We performed essential preprocessing steps such as lowercasing, tokenization, and punctuation

removal. Then, to count words in text reviews, we used SpaCy, a Python NLP library. We set the thresholds based on the average review length. If the review is shorter than the average, the deception indicator will increase.

**– Repeated Words Ratio (RWR)**

More repetition in the review text is a red flag of its authenticity as it indicates low memorized details and low valuable facts. We calculate the ratio of repeated words to the other words in each review using the Counter class and the collections module. For tokenization, we additionally used the 'nltk' library. The results show that the reviews contain 50% to 0% repeated words and the average number of repeated words in the dataset is 4%. So, we set the thresholds as:

**– Sentiments (S)**

We used the Valence Aware Dictionary and Sentiment Reasoner (VADER), a rule-based sentiment analysis tool. This tool uses a dictionary called Lexicon optimized for social media text sentiment analysis. This lexicon includes a range of phrases and terms together with the associated sentimental ratings[65]. VADER provides a polarity as a compound, a normalized score between $-1$ (most negative) and $+1$ (most positive). The compound scores were divided into five categories: Strong Positive $(0.6-1.0)$, Positive $(0.2-0.6)$, Neutral $(-0.2-0.2)$, Negative $(-0.6$ and $-0.2)$, Strong Negative $(-1.0--0.6)$. Based on our suggestion in this study the highest emotion reviews represent a high probability of deception, so we assign a 1 as the threshold for reviews with strong positive or negative scores and 0 otherwise.

**– Generalization**

Detecting generic language or generalization speech in a text is a challenge as it requires sophisticated natural language processing (NLP) techniques. In this work, we measured the generalization using Lexical Diversity and Named Entity Recognition (NER) [9]. Lexical diversity measures the range of different words used in a text. Higher lexical diversity indicates more unique words, which are associated with detailed and specific content. Using Named Entity Recognition (NER) to count the number of named entities (people, organizations, locations) can provide insights into the specificity of the text. The text_generalization score function combines lexical diversity and named entity counts to calculate a generalization score. Higher scores indicate more general text. We find the results range from 160% to 11%. The average was 81%, and we set the average as a threshold.

**– Passive Voice Ratio (PVR)**

Detecting the passive voice ratio in a text involves analyzing the grammatical structure of sentences to identify when the subject is the recipient of the action. We calculated the ratio of passive voice in the review to other active statements. We used spaCy library to detect passive voice in each sentence and compute the proportion of passive voice in the text [46]. The result shows that the average PVR in sentences is 3%. Hence, we set the threshold for this indicator to this value: if the PVR is greater than 3%, the probability of deception increases.

#### 7.3.2 Second Module Implementation

This module contains the features related to the reviewer. In this work, we considered two features, reviewer account authentication and total reviewer reviews.

**– Total Reviews of Reviewer (TRR)**

For the TRR we set the threshold to the average of the data. When the total reviewer reviews less than the average, we assign 1 to the indicator I9.

**– Account Type (AT)**

We used the local guide account from Google Maps for the type of reviewer account. Google assigns a local guide rank to the helpful, active reviewers who regularly contribute to the Google Maps community with high-quality contributions by a valid Google account. Users who provide valueless reviews, duplicate photos across locations, stolen photos, or repeat some reviews are not allowed to be local guides[3]. In our dataset, we assign Boolean values of 0 for local Guide reviewers and 1 for others, as other accounts increase the probability of deceptive reviews compared to the local guide.

#### 7.3.3 Third Module Implementation

In the third module, we measure the indicators of other metadata we consider the likes and the count of media with the reviews.

**– Likes (L)**

The number of likes or the useful votes of reviews can be considered a good helpful sign of review credibility analysis. In our collected dataset we noticed the votes ranged from 0 to 194. We transformed this feature to be a binary feature. For the indicator, we assigned 1 for reviews with less than 1.

**– Attached Media (AM)**

In our dataset, the attached media ranged from 0-10. This feature also is transformed into a binary feature (0 if true and 1 if false). Then, we set 1 as the minimum level of media linked with reviews. If the linked media is less than 1, we assign 1 for the indicator I11.

### 7.4 Annotation Guidelines

Based on the pre-defined criteria of the Framework modules, we set up clear guidelines to ensure annotations' consistency and accuracy. Measures were implemented to assess and maintain the quality of the annotations. We implemented the annotation rules using Excel to ensure and test the accuracy of dataset annotations. Table 3 shows the details of each rule, including the features and criteria used. By implementing these guidelines, we performed feature binarization of the reviews to normalize the scales of different features and classify them. Binarization is the process of converting continuous features to binary based on whether the value is greater or smaller than the threshold (any value above the threshold is set to 1, and anything below is set to 0 or vice versa) (Lebanon and El-Geish [30]).

[3]*About Local Guides - Google Maps Help*, accessed: August 01, 2024, [Online]. Available: URL: https://support.google.com/maps/answer/6225846?hl=en&ref_topic=14986415&sjid=225219747278477716-EU.

**Table 3.** Guideline rules of review annotation

| Feature | Does the review fall in these criteria? | Indicator | Annotation rule | Average |
|---|---|---|---|---|
| PM Ratio | 1: Extreme use of punctuation | I1 | if (PMR > AVG),<br>I1=1,<br>else I1=0 | 0.0300 |
| FPSP Ratio | 2: Avoiding personal voice | I2 | If (FSPR < AVG),<br>I2=1,<br>else I3=0 | 0.0499 |
| Review Length | 3: Concise content | I3 | if (RL < AVG),<br>I3=1,<br>else I3=0 | 28.01 |
| RW Ratio | 4: Extreme positive or negative emotion | I4 | if (RWR > AVG)<br>I4=1,<br>else I4=0 | 0.0422 |
| Sentiment | 5: Extreme duplication of words | I5 | if ((S < 1.0 AND S > 0.6)<br>OR<br>(S < −1.0 AND S > −0.6))<br>I5=1,<br>else I5=0 | 0.4728 |
| Generalization | 6: Extreme generic pattern | I6 | If (G > AVG),<br>I6=1,<br>else I6=0 | 0.8076 |
| Passive Voice | 7: Extreme passive voice | I7 | If (PVR > AVG)?<br>I7=1,<br>else I7=0 | 0.0314 |
| Total reviews | 8: Unreliable reviewer account | I8 | If (TRR < AVG)?<br>I8=1,<br>else I8=0 | 78.33 |
| Account type | 9: Low numbers of reviewer reviews | I9 | If (RAT = 0)?<br>I9=0,<br>else I9=1 | 0.4302 |
| Likes (L) | 10: Has likes vote? | I10 | If (LC < 1)?<br>I0=1,<br>else I10=0 | 0.3800 |
| Attached Media | 11: Has media attached? | I11 | If (AM < 1)?<br>I11=1,<br>else I11=0 | 0.326 |

## 7.5 The Deception Index (DI)

We calculated the Deception Index for each review using 11 predefined indicators. We assign each rule a value of 0 or 1, where 1 indicates a higher probability of deception. To calculate the credibility of reviews, we sum all the values of the indicators, find the average among all reviews, and then assign a label to the reviews based on the resulting value. Reviews with

**Table 4.** Example of review annotation on criteria analysis

| Review Body | Analysis | Explanation | Label |
|---|---|---|---|
| "a bad experience I will never go to that place again the lousy service disgusting tables with flies for your food and they still charge you 15 percent of their lousy service, the manager very bad, the place very bad not recommended to spend your money" | *Punctuation Ratio*: 0.008<br>*FSPR*: 0.045<br>*Word Count*: 46<br>*Repeated Word Ratio*: 0.20<br>*Sentiment*: -96% (negative sentiment closer to −1)<br>*Generalization Score*: 0.72<br>*Passive Voice Proportion*: 0%<br>*Total reviewer's reviews:* 135<br>*Account type:* not local guide<br>*Likes*: False<br>*Media*: 0 | The review content is short but contains enough detail to show the experience. Specific details about the experience may indicate a truthful review because they refer to aspects of the experience, making the review seem more credible. It also has a 0.20 repetitive ratio. The ratios of punctuation and first singular pronouns are normal. The review contains extremely negative language. All statements are active and direct. Regarding the reviewer account it contains 135 reviews, but it is not a local guide account. The review does not contain media and has no likes on it.<br>In this way, the total indicators $= 0+1+0+1+1+0+0+0+1+1+1 = 5 < 6$ (the average of DI). Based on this analysis, this review seems likely to be not deceptive. | Truthful (0) |
| "Great food and service. There is street parking if you get lucky. But it is worth the wait. Everything we ordered was d-e-l-ous." | *Punctuation Ratio*: 0.05<br>*FSPR*: 0<br>*Word Count*: 23<br>*Repeated Word Ratio*: 0<br>*Sentiment*: 0.89<br>*Generalization Score*: 1.03<br>*Passive Voice Proportion*: 0<br>*Total reviewer's reviews*: 7<br>*Account type*: Local guide<br>*Likes*: True<br>*Media*: 0 | The review content is short and does not contain enough details about the experience. It also does not contain repetitive words. The ratio of punctuation is high. The first singular pronouns were missed in this review. The review contains extremely positive language. No passive voice, where all statements are active and direct. The reviewer account contains 77 reviews, and it is a local guide account. The review does not contain media, and it has other users like it.<br>In this way, the total indicators $= 1+1+1+0+1+1+0+1+0+0+1 = 7 > 6$ (the average DI). Based on this analysis, this review seems likely to be deceptive. | Deceptive (1) |

credibility scores greater than the average are labeled by 1 as “deceptive”, while those with scores equal to or less than the average are labeled by 0 as “truthful” (equations (7.1) and (7.2)).

$$DI = \sum_{n=1}^{I11} I_n, \tag{7.1}$$

$$\text{if } (DI > AVG) \rightarrow \text{Review}_{\text{label}} = 1, \text{ else: Review}_{\text{label}} = 0. \tag{7.2}$$

Table 4 shows examples of reviews from the dataset and explains the reason for label choosing based on the guideline.

## 8. Dataset Validation and Statistic

We conducted a validation procedure to ensure the correctness of the dataset based on the defined guidelines. We assessed the distribution of dataset features to identify inconsistencies or anomalies, verify that the annotations align with the expected patterns, and validate the dataset. First, we check the completeness and consistency of each record. Then, we double-check reviews to make sure that all the reviews are truly labeled and follow the defined rules. We tested the conditional checks (if-else rules). We selected a random subset of the dataset. Then we manually test each rule and label the review. Finally, we compared the results of the manually labeled instances to those of the earlier labels and addressed any mistakes or null values. The final dataset contained 21476 processed and annotated reviews. The truthful reviews represent 12583 records while the deceptive are 8893. The dataset contains 8 continuous value attributes and 3 Boolean (account type, likes, Attached media). Figure 3 shows the dataset distribution in classes 0 and 1.

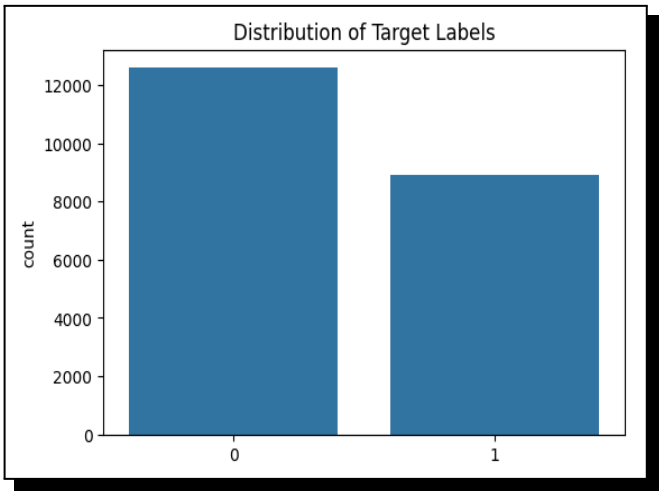


**Figure 3.** Dataset distribution

The longest review in the dataset contains 772 words while the shortest one contains 3. Figure 4 shows the histogram of the dataset’s continuous features with the distribution across the two classes.

Figure 5 shows the top ten words in the dataset classified by classes.

Figure 6 illustrates the RRD dataset’s features correlation matrix. Most features exhibit either a negligible or negative correlation (values that are close to or below 0), indicating that the features do not exhibit multicollinearity.

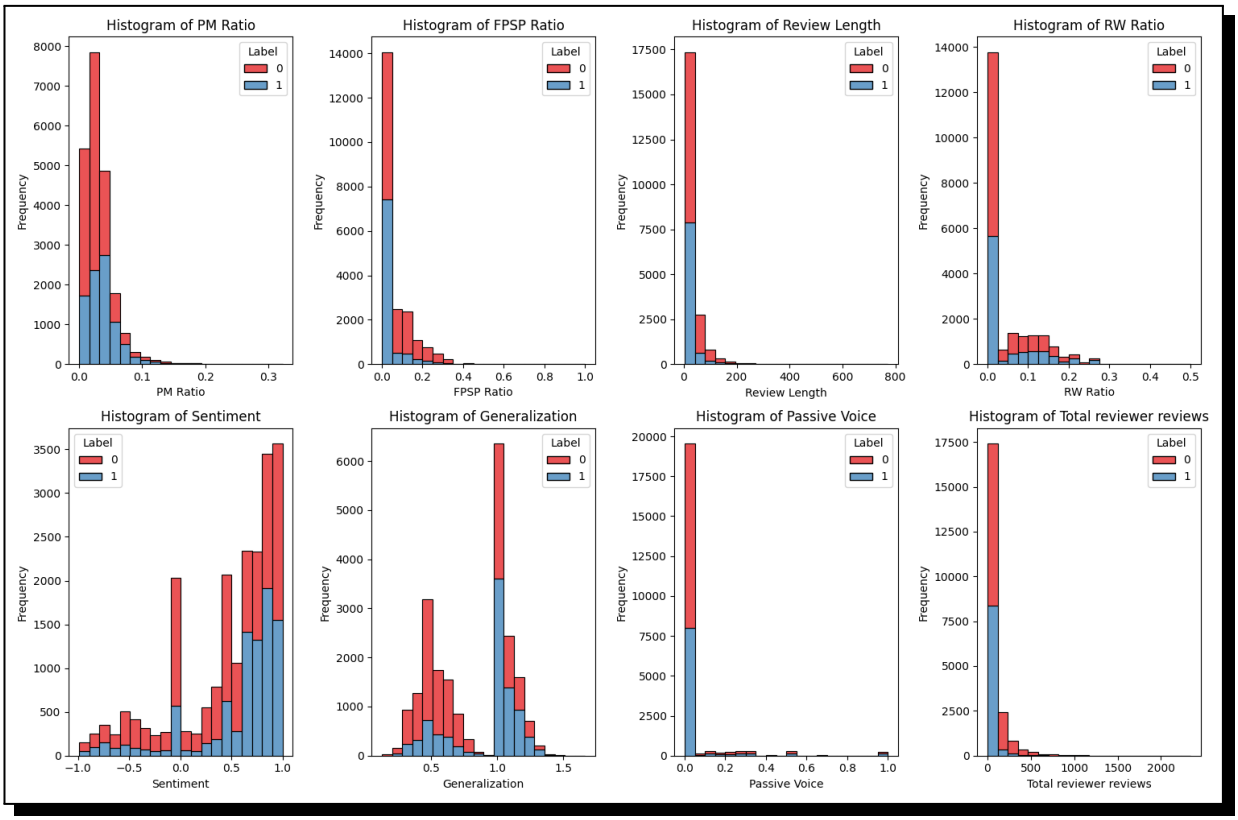


**Figure 4.** Histograms of the continuous features of the dataset

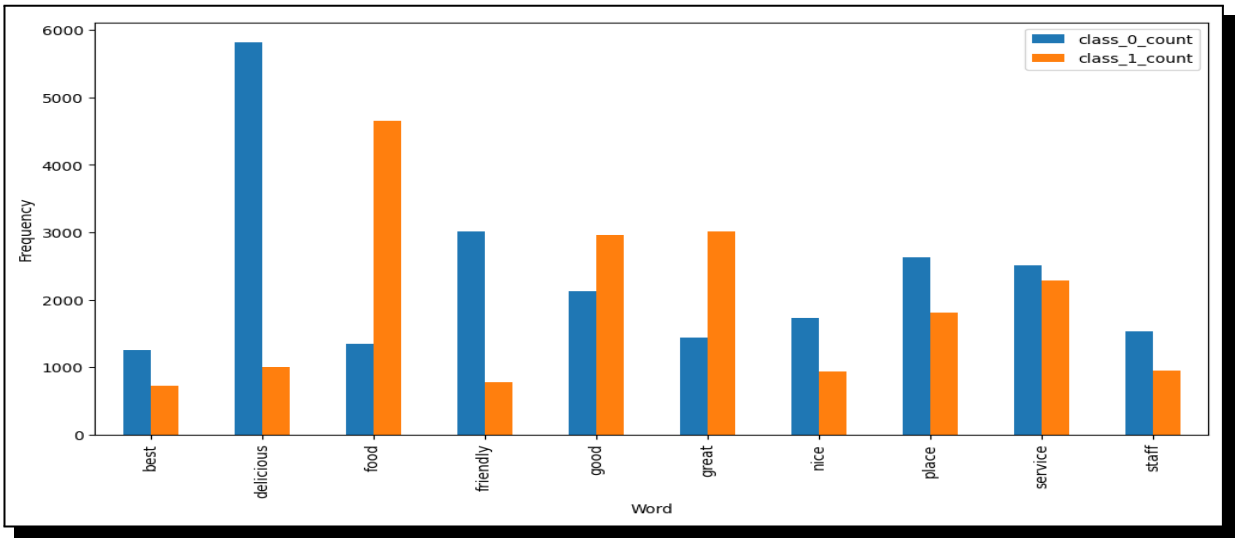


**Figure 5.** Top 10-word frequency in the dataset

Finally, to evaluate the datasets, we trained different traditional machine learning models to establish baseline performance metrics and to make dataset benchmarking using traditional machine learning models. This will help identify the most suitable model for future application and further optimization. We evaluated it regarding accuracy, precision, recall, and f1 score. However, since the dataset is imbalanced, we implemented the Synthetic Minority Oversampling Technique (SMOTE). Figure 5 shows the dataset distribution before and after SMOTE implementation.

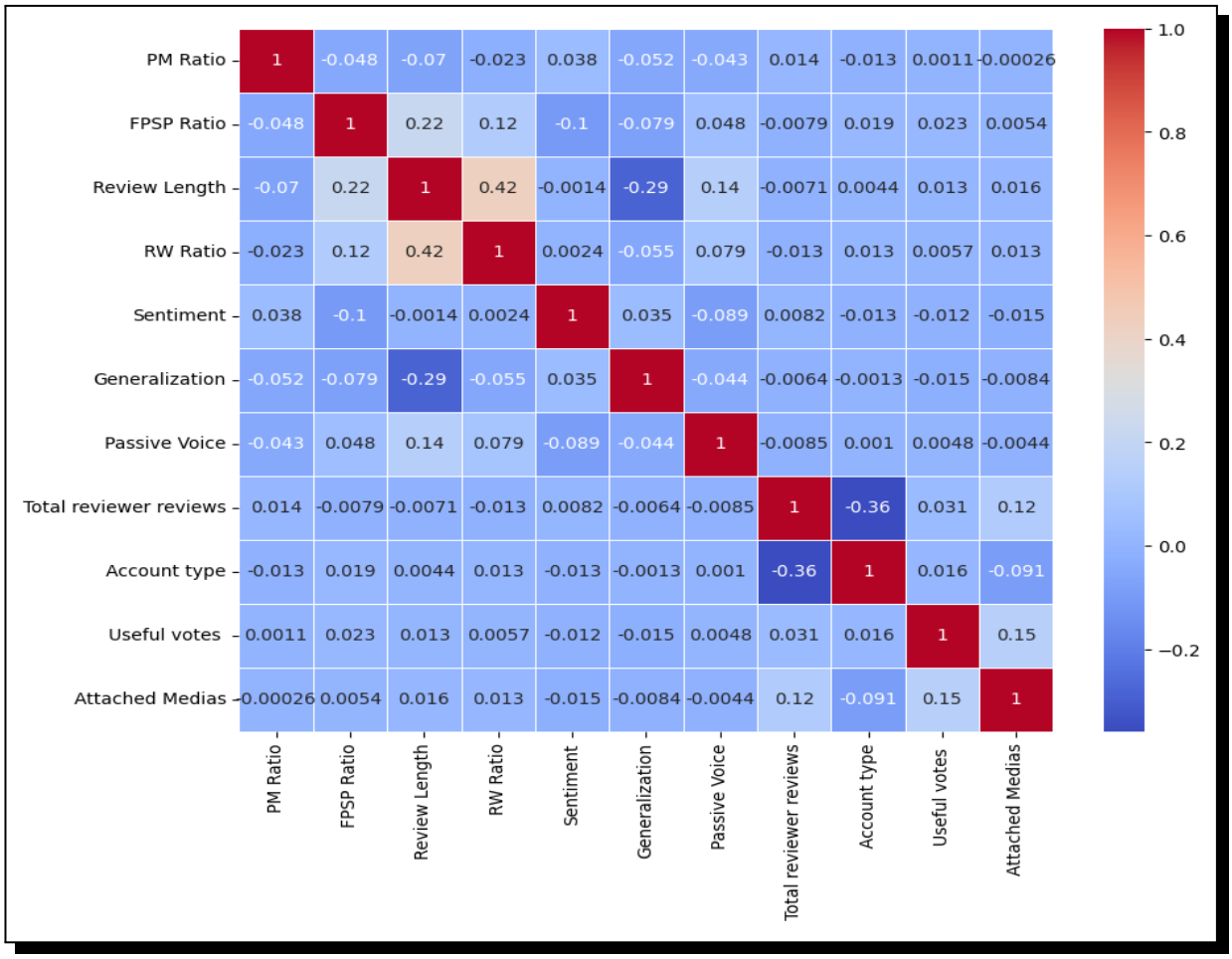


**Figure 6.** Figure6 RRD's correlation matrix

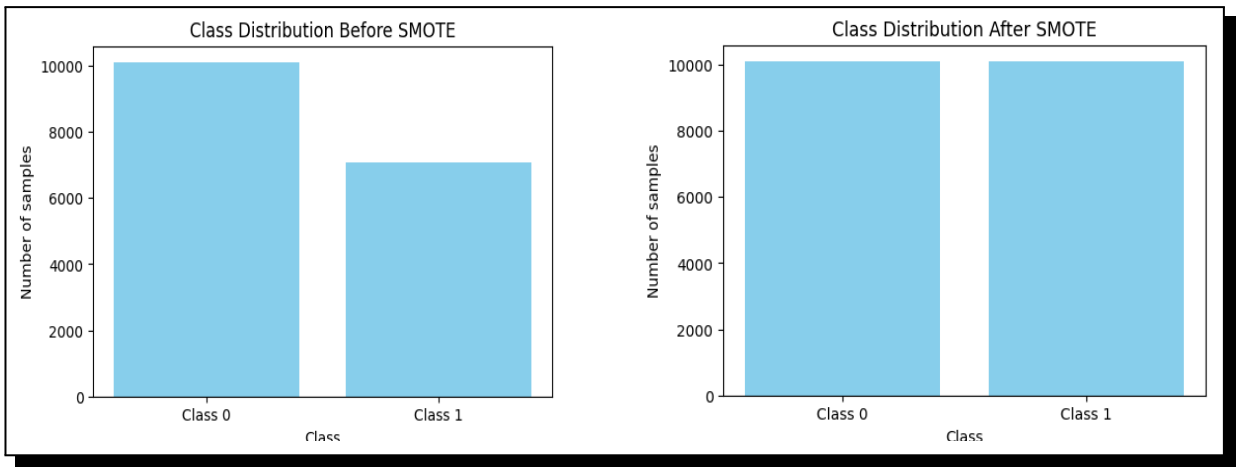


**Figure 7.** Dataset distribution before and after SMOTE

We used the K-Nearest Neighbors, Naive Bayes, Logistic Regression Model, Support Vector Machine, and Random Forest Model. Random Forest reached an accuracy of 98%. It performed significantly better than other models, while KNN had the lowest result. The baseline performance of the standard models is summarized in Table 1. Figure 8 shows the comparison of the accuracy of the machine learning classifier.

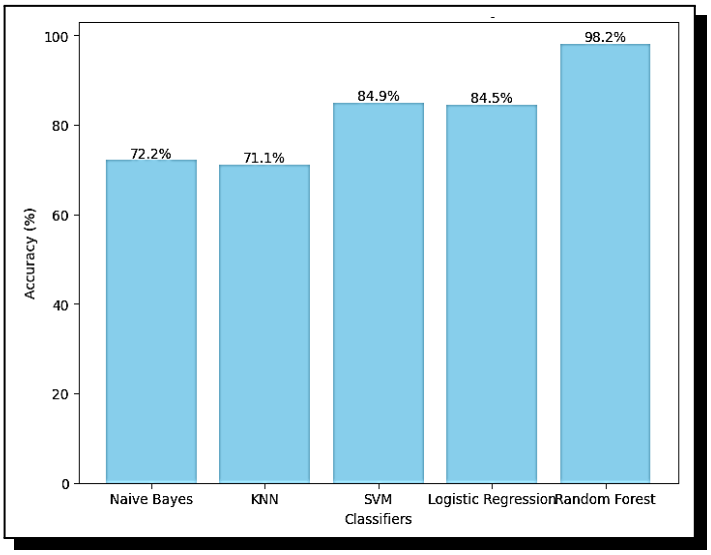


**Figure 8.** Comparison of the classifier accuracy

**Table 5.** Standard models performance

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| KNN | 0.71089 | 0.721952 | 0.71089 | 0.71273 |
| Naive Bayes | 0.72183 | 0.753071 | 0.72183 | 0.72251 |
| Logistic Regression | 0.84497 | 0.848479 | 0.84497 | 0.84567 |
| SVM | 0.84939 | 0.852908 | 0.84939 | 0.85007 |
| Random Forest | 0.98245 | 0.980473 | 0.98045 | 0.98043 |

## 9. Limitations and Future Work

While our dataset is extensive, this work is restricted by being confined to the English language, one platform (Google Maps), one country (USA), and one sector (restaurants). So, we cannot generalize the findings of this paper to all other types of reviews. Those writing reviews in Google Maps do not fully represent the user base. Also, we suggested designing a fully automated feature calculation system using machine learning models to yield better and more accurate results. For the future, we recommend using the proposed framework using new datasets in different areas and considering other languages. Adding spatial and timing features will be valuable to enhance deception detection. We will further improve the proposed framework by investigating more indicators and evaluating the system using more extensive datasets. We did not consider other features of the reviewers' social networks. The inclusion of such characteristics in further studies is expected to raise detection accuracy. We recommend assigning weight to each deception indicator to describe its importance in detection. Finally, we did not consider the emerging problem of spammer groups; this should be investigated in future studies.

## 10. Conclusion

Deception in online reviews is one of the most complex issues in the business world. Detection of manipulated information in reviews is a critical and challenging issue with the absence of many

deceptions' noticeable nonverbal cues. This study aimed to analyze and integrate deception theories and techniques to effectively identify deception cues in online reviews.

We suggested a multifaceted novel approach to detect deception in online reviews. Then, we implemented the framework to annotate a new dataset of reviews collected from Google Maps reviews. Unlike other datasets, our dataset represents unintentional deception that was analyzed and classified based on deception theories. The dataset is analyzed and annotated based on review and reviewer features with psychological theories. The dataset is anticipated to be a valuable tool for researchers working on deception detection algorithms. It may be applied to giving companies valuable tools to protect their platforms against fraudulent reviews. Our proposed framework can be implemented and improved to serve more domains. Also, we implemented the rule-based guidelines of the framework to annotate a new dataset that will support the research community. Finally, we trained traditional ML models on the dataset and compared the results.

## Competing Interests

The authors declare that they have no competing interests.

## Authors' Contributions

All the authors contributed significantly in writing this article. The authors read and approved the final manuscript.

## References

[1] M. Abdulqader, A. Namoun and Y. Alsaawy, Fake online reviews: A unified detection model using deception theories, *IEEE Access* **10** (2022), 128622 – 128655, DOI: 10.1109/ACCESS.2022.3227631.

[2] E. Abedin, A. Mendoza and S. Karunasekera, Credible vs fake: A literature review on differentiating online reviews based on credibility, in: *Forty-First International Conference on Information Systems, India 2020*, pp. 1 – 17 (2020).

[3] F. Abri, L. F. Gutiérrez, A. S. Namin, K. S. Jones and D. R. W. Sears, Linguistic features for detecting fake reviews, in: *2020 19th IEEE International Conference on Machine Learning and Applications* (ICMLA, Miami, FL, USA, 2020), pp. 352 – 359 (2020), DOI: 10.1109/ICMLA51294.2020.00063.

[4] H. Alaskar, Z. Sbaï, W. Khan, A. Hussain and A. Alrawais, Intelligent techniques for deception detection: A survey and critical study, *Soft Computing* **27**(7) (2023), 3581 – 3600, DOI: 10.1007/s00500-022-07603-w.

[5] S. N. Alsubari, M. B. Shelke and S. N. Deshmukh, Fake reviews identification based on deep computational linguistic features, *International Journal of Advanced Science and Technology* **29**(8s) (2020), 3846 – 3856.

[6] S. N. Alsubari, S. N. Deshmukh, A. A. Alqarni, N. Alsharif, T. H. H. Aldhyani, F. W. Alsaade and O. I. Khalaf, Data analytics for the identification of fake reviews using supervised learning, *Computers, Materials & Continua* **70**(2) (2022), 3189 – 3204, DOI: 10.32604/cmc.2022.019625.

[7] S. Ansari and S. Gupta, Review manipulation: Literature review, and future research agenda, *Pacific Asia Journal of the Association for Information Systems* **13**(1) (2021), Article 4, DOI: 10.17705/1pais.13104.

**[8]** M. Z. Asghar, A. Ullah, S. Ahmad and A. Khan, Opinion spam detection framework using hybrid classification scheme, *Soft Computing* **24**(5) (2020), 3475 – 3498, DOI: 10.1007/s00500-019-04107-y.

**[9]** I. Augenstein, L. Derczynski and K. Bontcheva, Generalisation in named entity recognition: A quantitative analysis, *Computer Speech & Language* **44** (2017), 61 – 83, DOI: 10.1016/j.csl.2017.01.012.

**[10]** K. Barik and S. Misra, Analysis of customer reviews with an improved VADER lexicon classifier, *Journal of Big Data* **11** (2024), Article number: 10, DOI: 10.1186/s40537-023-00861-x.

**[11]** G. Bathla, P. Singh, R. K. Singh, E. Cambria and R. Tiwari, Intelligent fake reviews detection based on aspect extraction and analysis using deep learning, *Neural Computing and Applications* **34** (2022), 20213 – 20229, DOI: 10.1007/s00521-022-07531-8.

**[12]** R. F. Baumeister and D. G. Hutton, Self-presentation theory: Self-construction and audience pleasing, in: *Theories of Group Behavior*, B. Mullen and G. R. Goethals (editors), Springer Series in Social Psychology, Springer, New York, pp. 71 – 87 (1987), DOI: 10.1007/978-1-4612-4634-3_4.

**[13]** S. Berry, Fake Google restaurant reviews and the implications for consumers and restaurants, *arXiv:2401.11345* (2024), 1 – 158, DOI: 10.48550/arXiv.2401.11345.

**[14]** J. K. Burgoon and D. B. Buller, Interpersonal deception theory, in: *The International Encyclopedia of Interpersonal Communication*, C. R. Berger, M. E. Roloff, S. R. Wilson, J. P. Dillard, J. Caughlin and D. Solomon (editors), John Wiley & Sons, Ltd., (2015), DOI: 10.1002/9781118540190.wbeic170.

**[15]** T. L. Carson, *Lying and Deception: Theory and Practice*, Oxford University Press, Oxford, (2010), DOI: 10.1093/acprof:oso/9780199577415.001.0001.

**[16]** M. Carter, M. Tsikerdekis and S. Zeadally, Approaches for fake content detection: Strengths and weaknesses to adversarial attacks, *IEEE Internet Computing* **25**(2) (2021), 73 – 83, DOI: 10.1109/MIC.2020.3032323.

**[17]** E. A. Cranford, C. Gonzalez, P. Aggarwal, M. Tambe, S. Cooney and C. Lebiere, Towards a cognitive theory of cyber deception, *Cognitive Science* **45**(7) (2021), e13013, DOI: 10.1111/cogs.13013.

**[18]** J. W. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 3rd edition, Sage Publications, Inc., Thousand Oaks, CA, xxix + 260 pages (2009).

**[19]** K. S. Desale, S. Shinde, N. Magar, S. Kullolli and A. Kurhade, Fake review detection with concept drift in the data: A survey, in: *Proceedings of Seventh International Congress on Information and Communication Technology*, X.-S. Yang, S. Sherratt, N. Dey and A. Joshi (editors), Springer Nature, Singapore, pp. 719 – 726 (2023), DOI: 10.1007/978-981-19-1610-6_63.

**[20]** I. Goormans, L. Mergaerts and C. Vandeviver, SCANning for truth. Scholars' and practitioners' perceptions on the use(fulness) of Scientific Content Analysis in detecting deception during police interviews, *Psychology, Crime & Law* **30**(9) (2022), 971 – 993, DOI: 10.1080/1068316X.2022.2139828.

**[21]** P. Gryka and A. Janicki, Detecting fake reviews in Google Maps — A case study, *Applied Sciences* **13**(10) (2023), 6331, DOI: 10.3390/app13106331.

**[22]** P. Hajek and J.-M. Sahut, Mining behavioural and sentiment-dependent linguistic patterns from restaurant reviews for fake review detection, *Technological Forecasting and Social Change* **177** (2022), 121532, DOI: 10.1016/j.techfore.2022.121532.

**[23]** S. Hlee, H. Lee, C. Koo and N. Chung, Fake reviews or not: Exploring the relationship between time trend and online restaurant reviews, *Telematics and Informatics* **59** (2021), 101560, DOI: 10.1016/j.tele.2020.101560.

[24] H. A. Howard, D. J. Griffin and Z. W. Arth, Information manipulation and cognitive trust: An organizational replication and extension of IMT, *Atlantic Journal of Communication* **30**(3) (2002), 219 – 230, DOI: 10.1080/15456870.2021.1884078.

[25] K. Ioannidis, T. Offerman and R. Sloof, Lie detection: A strategic analysis of the verifiability approach, *American Law and Economics Review* **24**(2) (2022), 659 – 705, DOI: 10.1093/aler/ahac005.

[26] A. A. Iswara and K. A. Bisena, Manipulation and persuasion through language features in fake news, *Retorika: Jurnal Ilmu Bahasa* **6**(1) (2020), 26 – 32, DOI: 10.22225/jr.6.1.1338.26-32.

[27] N. Jindal and B. Liu, Opinion spam and analysis, in: *Proceedings of the 2008 International Conference on Web Search and Data Mining* (WSDM'08), Association for Computing Machinery, New York, NY, USA, pp. 219 – 230 (2008), DOI: 10.1145/1341531.1341560.

[28] A. Kuzio, *Cross-cultural Deception in Polish and American English in Computer-Mediated Communication*, Cambridge Scholars Publishing, Newcastle upon Tyne, UK, (2018).

[29] H. Le and B. Kim, Detection of fake reviews on social media using machine learning algorithms, *Issues in Information Systems* **21**(1) (2020), 185 – 194, DOI: 10.48009/1_iis_2020_185-194.

[30] G. Lebanon and M. El-Geish, *Computing with Data: An Introduction to the Data Industry*, Springer, xvii + 576 pages (2018), DOI: 10.1007/978-3-319-98149-9.

[31] L. Li, B. Qin, W. Ren and T. Liu, Document representation and feature combination for deceptive spam review detection, *Neurocomputing* **254** (2017), 33 – 41, DOI: 10.1016/j.neucom.2016.10.080.

[32] F. Li, M. Huang, Y. Yang and X. Zhu, Learning to identify review spam, in: *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence* (IJCAI'11, Barcelona, Catalonia, Spain), Vol. 3, AAAI Press, pp. 2488 – 2493 (2011), DOI: 10.5591/978-1-57735-516-8/IJCAI11-414.

[33] J. Li, M. Ott, C. Cardie and E. Hovy, Towards a general rule for identifying deceptive opinion spam, in: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), Association for Computational Linguistics, Baltimore, Maryland, pp. 1566 – 1576 (2014), DOI: 10.3115/v1/P14-1147.

[34] J. Li, X. Xu and E. W. T. Ngai, How review content, sentiment and helpfulness votes jointly affect trust of reviews and attitude, *Internet Research* **34**(6) (2024), 2232 – 2256, DOI: 10.1108/INTR-01-2023-0025.

[35] Y. Li, Z. Liu and X. Liu, Who did I lie to that day? Deception impairs memory in daily life, *Psychological Research* **86**(6) (2022), 1763 – 1773, DOI: 10.1007/s00426-021-01619-x.

[36] Y. Li, Z. Zhang, S. Pedersen, X. Liu and Z. Zhang, The influence of relative popularity on negative fake reviews: A case study on restaurant reviews, *Journal of Business Research* **162** (2023), 113895, DOI: 10.1016/j.jbusres.2023.113895.

[37] Y. Lin, T. Zhu, H. Wu, J. Zhang, X. Wang and A. Zhou, Towards online anti-opinion spam: Spotting fake reviews from the review sequence, in: *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (ASONAM2014, Beijing, China, 2014), IEEE, pp. 261 – 264 (2014), DOI: 10.1109/ASONAM.2014.6921594.

[38] P. Liu, Z. Xu, J. Ai and F. Wang, Identifying indicators of fake reviews based on spammer's behavior features, in: *2017 IEEE International Conference on Software Quality, Reliability and Security Companion* (QRS-C, Prague, Czech Republic, 2017), pp. 396 – 403 (2017), DOI: 10.1109/QRS-C.2017.72.

[39] B. Manaskasemsak, J. Tantisuwankul and A. Rungsawang, Fake review and reviewer detection through behavioral graph partitioning integrating deep neural network, *Neural Computing and Applications* **35** (2023), 1169 – 1182, DOI: 10.1007/s00521-021-05948-1.

**[40]** B. Marcus, 'Faking' from the applicant's perspective: A theory of self-presentation in personnel selection settings, *International Journal of Selection and Assessment* **17**(4) (2009), 417 – 430, DOI: 10.1111/j.1468-2389.2009.00483.x.

**[41]** R. Mohawesh, S. Xu, M. Springer, Y. Jararweh, M. Al-Hawawreh and S. Maqsood, An explainable ensemble of multi-view deep learning model for fake review detection, *Journal of King Saud University - Computer and Information Sciences* **35**(8) (2023), 101644, DOI: 10.1016/j.jksuci.2023.101644.

**[42]** R. Mohawesh, S. Xu, S. N. Tran, R. Ollington, M. Springer and Y. Jararweh, Fake reviews detection: A survey, *IEEE Access* **9** (2021), 65771 – 65802, DOI: 10.1109/ACCESS.2021.3075573.

**[43]** A. Mukherjee, V. Venkataraman, B. Liu and N. Glance, Fake review detection: Classification and analysis of real and pseudo reviews, Technical Report, Department of Computer Science (UIC-CS-2013-03), University of Illinois at Chicago, USA, (2013), URL: https://www2.cs.uh.edu/~arjun/tr/UIC-CS-TR-yelp-spam.pdf.

**[44]** A. Mukherjee, V. Venkataraman, B. Liu and N. Glance, What yelp fake review filter might be doing?, *Proceedings of the International AAAI Conference on Web and Social Media* **7**(1) (2013), 409 – 418, DOI: 10.1609/icwsm.v7i1.14389.

**[45]** G. Nahari, Chapter 14 – The applicability of the verifiability approach to the real world, *Detecting Concealed Information and Deception: Recent Development*, J. P. Rosenfeld (editor), Academic Press, pp. 329 – 349 (2018), DOI: 10.1016/B978-0-12-812729-2.00014-8.

**[46]** M. Neumann, D. King, I. Beltagy and W. Ammar, ScispaCy: fast and robust models for biomedical natural language processing, in: *Proceedings of the 18th BioNLP Workshop and Shared Task*, Florence, Italy, Association for Computational Linguistics, pp. 319 – 327 (2019), DOI: 10.18653/v1/W19-5034.

**[47]** A. Nortje and C. Tredoux, How good are we at detecting deception? A review of current techniques and theories, *South African Journal of Psychology* **49**(4) (2019), 491 – 504, DOI: 10.1177/0081246318822953.

**[48]** Y. Pan and L. Xu, Detecting fake online reviews: An unsupervised detection method with a novel performance evaluation, *International Journal of Electronic Commerce* **28**(1) (2024), 84 – 107, DOI: 10.1080/10864415.2023.2295067.

**[49]** S. A. Prome, N. A. Ragavan, M. R. Islam, D. Asirvatham and A. J. Jegathesan, Deception detection using machine learning (ML) and deep learning (DL) techniques: A systematic review, *Natural Language Processing Journal* **6** (2024), 100057, DOI: 10.1016/j.nlp.2024.100057.

**[50]** R. Reddy and A. U. Motagi, Fake review detection and emotion recognition based on semantic feature selection with bi-directional long short term memory, *International Journal of Intelligent Engineering and Systems* **15**(5) (2022), 473 – 482, DOI: 10.22266/ijies2022.1031.41.

**[51]** G. M. Shahariar, Md. T. R. Shawon, F. M. Shah, M. S. Alam and Md. S. Mahbub, Bengali fake reviews: A benchmark dataset and detection system, *Neurocomputing* **592**, 127732, DOI: 10.1016/j.neucom.2024.127732.

**[52]** L. Shi, S. Xie, L. Wei, Y. Tao, A. W. Junaid and Y. Gao, Joint sentiment topic model with word embeddings for fake review detection, *SSRN*, DOI: 10.2139/ssrn.4096565.

**[53]** N. Shirin, F. Erfan, N. Salim and G. S. Hamid, A novel approach for opinion spam detection in e-commerce, in: *Proceedings of the 8th International Conference on E-Commerce with a Focus on E-Trust*, Mashhad, Iran (2014).

**[54]** S. Shojaee, A. Azman, M. Murad, N. Sharef and N. Sulaiman, A framework for fake review annotation, in: *Proceedings of the 2015 17th UKSIM-AMSS International Conference on Modelling and Simulation* (UKSIM'15), IEEE Computer Society, pp. 153 – 158 (2015).

**[55]** N. Smith, Reading between the lines: An evaluation of the Scientific Content Analysis technique (SCAN), C. F. Willis (editor), Great Britain Home Office, England, United Kingdom, 53 pages (2001), URL: https://www.ojp.gov/ncjrs/virtual-library/abstracts/reading-between-lines-evaluation-scientific-content-analysis.

**[56]** S. L. Sporer, A. L. Manzanero and J. Masip, Optimizing CBCA and RM research: recommendations for analyzing and reporting data on content cues to deception, *Psychology, Crime & Law* **27**(1) (2021), 1 – 39, DOI: 10.1080/1068316X.2020.1757097.

**[57]** C. Sun, Q. Du and G. Tian, Exploiting product related review features for fake review detection, *Mathematical Problems in Engineering* **2016** (2016), e4935792, DOI: 10.1155/2016/4935792.

**[58]** J. Y. Thomas and D. P. Biros, An empirical evaluation of interpersonal deception theory in a real-world, high-stakes environment, *Journal of Criminal Psychology* **10**(3) (2020), 185 – 199, DOI: 10.1108/JCP-07-2019-0025.

**[59]** B. Verschuere, G. Bogaard and E. Meijer, Discriminating deceptive from truthful statements using the verifiability approach: A meta-analysis, *Applied Cognitive Psychology* **35**(2) (2020), 374 – 384, DOI: 10.1002/acp.3775.

**[60]** A. Vrij and G. Ganis, Chapter 7 – Theories in deception and lie detection, *Credibility Assessment*, D. C. Raskin, C. R. Honts and J. C. Kircher (editors), Academic Press, San Diego, pp. 301 – 374 (2014), DOI: 10.1016/B978-0-12-394433-7.00007-5.

**[61]** A. Vrij, M. Hartwig and P. A. Granhag, Reading lies: Nonverbal communication and deception, *Annual Review of Psychology* **70**(1) (2019), 295 – 317, DOI: 10.1146/annurev-psych-010418-103135.

**[62]** A. Vrij, P. A. Granhag, T. Ashkenazi, G. Ganis, S. Leal and R. P. Fisher, Verbal lie detection: Its past, present and future, *Brain Sciences* **12**(12) (2022), 1644, DOI: 10.3390/brainsci12121644.

**[63]** A. Wielgopolan and K. K. Imbir, Cognitive load and deception detection performance, *Cognitive Science* **47**(7) (2023), e13321, DOI: 10.1111/cogs.13321.

**[64]** M. Wise and D. Rodriguez, Detecting deceptive communication through computer-mediated technology: Applying interpersonal deception theory to texting behavior, *Communication Research Reports* **30**(4) (2013), 342 – 346, DOI: 10.1080/08824096.2013.823861.

**[65]** Y. Wu, E. W. T. Ngai, P. Wu and C. Wu, Fake online reviews: Literature review, synthesis, and directions for future research, *Decision Support Systems* **132** (2020), 113280, DOI: 10.1016/j.dss.2020.113280.

**[66]** K.-H. Yoo and U. Gretzel, Comparison of deceptive and truthful travel reviews, in :*Information and Communication Technologies in Tourism*, W. Höpken, U. Gretzel and R. Law (editors), Springer, Vienna, pp. 37 – 47 (2009), DOI: 10.1007/978-3-211-93971-0_4.