IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 November 2022, accepted 4 December 2022, date of publication 8 December 2022,
date of current version 14 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3227631

RESEARCH ARTICLE

Fake Online Reviews: A Unified Detection Model Using Deception Theories

MUJAHED ABDULQADER, ABDALLAH NAMOUN, (Member, IEEE), AND YAZED ALSAAWY
Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah 42351, Saudi Arabia
Corresponding author: Mujahed Abdulqader (mujahid.kamal2013@gmail.com)

ABSTRACT Online reviews influence consumers' purchasing decisions. However, identifying fake online reviews automatically remains a complex problem, and current detection approaches are inefficient in preventing the spread of fake reviews. The literature on fake reviews detection lacks a comprehensive and interpretable theory-based model with high performance, which enables us to understand the phenomenon from a psychological perspective and analyze reviews based on user-generated content as well as consumer behavior. In this research, we synthesized ten well-founded deception theories from psychology, namely leakage theory, four-factor theory, interpersonal deception theory, self-presentational theory, reality monitoring theory, criteria-based content analysis, scientific content analysis, verifiability approach, truth-default theory, and information manipulation theory, and selected nine relevant constructs to develop a unified model for detecting fake online reviews. These constructs include specificity, quantity, non-immediacy, affect, uncertainty, informality, consistency, source credibility, and deviation in behavior. We characterized the selected constructs using verbal and non-verbal features to validate the proposed model empirically. Subsequently, we extracted features from the Yelp datasets and used them to train four machine learning algorithms, specifically Logistic Regression, Naïve Bayes, Decision Tree, and Random Forest. We demonstrated that quantity, non-immediacy, affect, informality, consistency, source credibility, and deviation in behavior are essential constructs for detecting fake reviews. To our surprise, we discovered that non-verbal features are more important than verbal features and that combining features from both types improves the prediction performance. Our theory-based model outperformed most of the state-of-the-art fake review detection models and yielded high interpretability and low complexity.

INDEX TERMS Fake review detection, online reviews, deception detection, feature extraction, machine learning, deception theories.

I. INTRODUCTION

People increasingly use online review applications to convey their thoughts, on various items, such as products and local companies [1]. These reviews tell consumers about the experiences of others using certain items. These items have a quality that can only be judged after usage [2]. Online reviews heavily influence consumers' purchasing decisions. Unfortunately, some companies create fake reviews to influence consumers' impressions of their or their competitors' goods [3].

Fake online reviews have several characteristics. First, they are described as online reviews written by people based on their imaginations without actual experience [4]. Second, fake reviews have a core characteristic which is their ability to mislead consumers [5]. Third, there are two main ways to produce fake reviews, namely: human-generated and computer-generated way [6]. Fourth, fake reviews can be written by different types of consumers, online merchants, or platforms. Fifth, fake reviews are multilingual [7], [8] and come from different cultures. Sixth, reading the text of online reviews is insufficient for humans to differentiate between truthful and fake reviews [9].

Consequently, determining how to automatically identify inaccurate and fraudulent fake reviews is a difficult problem.

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson.

The challenge of identifying fake reviews is referred to as the problem of fake review detection [10]. Current detection approaches are inefficient in preventing the spread of fake reviews because fraudulent users routinely submit fake reviews with new features to avoid detection [5].

When fraudulent users employ basic deceptive techniques, the traditional detection methods fail to distinguish between normal and fraudulent users. For example, they may mimic regular users by posting both truthful and fake reviews [11]. Therefore, an accurate fake reviews detection model is required.

When developing a model to detect fake reviews that use features from deception theories, feature engineering plays a significant role to identify, manipulate, select and extract the most valuable features of fake reviews from raw data. This helps simplify the model and achieve better results in detecting deceptive behavior in fake reviews.

Consumer purchasing decisions, product reputations, sales volumes, and merchant profits are all influenced by reviews [12], [13], [14]. More than 80% of shoppers in the United States read Internet reviews before buying a product [2]. Only a 1% increase in hotel rating scores might result in a 2.6% increase in sales per room [15]. Restaurants sell 19% more frequently when given an extra half-star score [16]. A one-star decrease in a company's Yelp rating results in a 5%–9% decrease in revenue [17]. The percentage of fake reviews can reach up to 33.3% [18]. Approximately 10.3% of online products were subjected to review manipulation [19].

Recently, Amazon observed a significant increase in unverified reviews (reviews lacking the ''verified purchaser'' label) [20]. In March 2019, 99.6% of 1.8 million unverified ratings were five stars. In comparison, from 2017 to 2018, there were an average of 300 thousand unverified reviews every month, with only 75% of them being 5-star [20].

While fake reviews have a significant impact on e-commerce, detecting them is crucial, but complicated. Detection of fake reviews is easy when the user shows apparent suspicious behavior, such as leaving reviews every day using different devices, because normal users do not post reviews daily and do not use various devices to do so [21]. However, this problem has become complicated because of the deception strategies. Fraudulent users change their techniques to avoid detection systems [21]. Some of them attempt to appear normal by including links to well-known entities [11]. Some of them pay people to participate in spam activities through crowdsourcing platforms [22]. Alternatively, they generate fake reviews using deep learning models [23]. For example, on Twitter, they discovered that some fake followers avoided detection systems by writing reviews or following real people [11].

Although almost all studies developing new algorithms and methods to detect fake reviews claim that their algorithms have a high level of accuracy, the frequency of fake reviews continues to rise [5]. Therefore, continuing to develop approaches or algorithms for detecting fake reviews is a priority. To achieve this, we need to investigate the features of fake reviews to distinguish between truthful and fake reviews accurately [5].

New features can be extracted from the data to focus on deception in the behavior of fraudulent users, especially when the detection methods are limited by the available attributes in the data of the user. We cannot listen to users' voices, see facial reactions, or observe body language. Instead, we can only deal with their access to the platform, written reviews, and ratings with time and frequency dimensions.

Consumers have a trust issue when it comes to online reviews; they must read and compare reviews carefully [6]. On the one hand, reviews are incredibly beneficial because they provide important information that helps consumers in the purchase decision to spend their money on high-quality items and services [6]. In addition, online sellers are highly affected by fake reviews, which can damage their reputations and businesses.

Fake reviews give a negative view to consumers, which damages platforms' reputations and reduces the number of consumers [5]. Platforms that allow users to write reviews need to improve their fake reviews detection systems regularly, which is critical for maintaining the platforms' trustworthiness and providing a high-quality user experience for consumers seeking information [6].

The problem of fake reviews requires continuous research to deeply understand it and find effective solutions, while fraudulent users continuously change their techniques to avoid detection systems. This issue is still increasing and affecting platforms, consumers, sellers, and researchers, and it still requires considerable effort to analyze, solve, and reduce the consequences. Consumers need truthful experience information for online products, whereas sellers need to maintain their reputations and businesses. Platforms need to provide trustworthiness for consumers and sellers and guarantee fair competition. Therefore, detecting and cleaning fake reviews from platforms with high accuracy will guarantee more trustworthy and fair platforms for consumers and sellers.

We can summarize the challenges and limitations of fake online reviews detection as follows: First, one of the most important difficulties presented by fake reviews is that even expert customers cannot spot them accurately, in addition to their exponentially growing number. As a result, there are few labeled datasets to be used as the gold standard for training classification methods [3], [7], [24]. Second, reviews are written in many languages [7], [8], with most studies focusing on English. Third, a class imbalance can be observed in several datasets [8], where the proportion of reviews labeled as fake is tiny compared to reviews labeled as truthful [25]. Fourth, the limitation in the number of available data attributes in the public datasets [7]. Some attributes are required, such as the email address, sign-in location, and IP address [7]. Fifth, the problem of concept drift which is the continuous change in the features of online reviews over time [26]. One of the reasons for concept drift is that once fake reviewers learn fake detection criteria, they adjust their behavior to

appear normal, making the criteria useless [6]. Sixth, there is a lack of interpretability in deep-learning-based models for fake reviews detection. Regardless of their high performance, they are still untrusted because of their varying performance from one dataset to another [9]. The study of interpretability can be carried out by focusing on fundamental theories [9].

We still need an interpretable trustful model with high performance for fake reviews detection, which is theory-based and enables us to understand the phenomenon from a psychological perspective using the data of the reviewer's behavior and the review's content. The model needs to be interpretable in a deeper manner so that humans can understand the reason for the model's decision and its psychological interpretation.

Our model covers the most well-known deception theories from two different perspectives in psychology to analyze suspected content and behavior. Unlike previous works that considered a limited number of deception theories, focused only on the old perspective in deception theories, had unclear mapping between theoretical constructs and practical features, focused on the content of reviews and neglected the behavioral side, or were built without considering any fundamental theories.

Our research objectives (ROs) include:

- **RO1**: Synthesize deception aspects from deception theories to formalize a theoretical model of fake reviews detection.
- **RO2**: Specify deception features that can be engineered from the available attributes in open-source datasets.
- **RO3**: Develop a feasible fake online reviews detection model based on the selected deception features.
- **RO4**: Apply feature extraction methods to retrieve the features related to deception aspects from the available attributes in online reviews.
- **RO5**: Test the performance of our unified deception-based fake reviews detection model.

Accordingly, our research questions (RQs) are:

- **RQ1**: What deception aspects from deception theories should be considered to capture the behavior of fraudulent online customers?
- **RQ2**: What are the possible features that can be extracted from the available attributes in open-source customer reviews to reflect the relevant aspects of deceptive behavior?
- **RQ3**: What techniques can be used to extract the features related to deception aspects from the available attributes in user data?
- **RQ4**: Can the deception-based fake reviews detection model improve the performance of fake reviews detection?

Several important research contributions are made through this work:

- First, this is the first time that a fake reviews detection model is built based on synthesizing the most popular deception theories in psychology from both perspectives: the dominant perspective (cue theories) and the new perspective (non-cue theories), integrating the strengths of these two perspectives.
- Second, this study provides unified terms to describe deception constructs and clear mapping between deception theories and the selected verbal and non-verbal features for fake reviews detection while considering fake reviews as a shape of deception.
- Third, this study proposes a pure theory-based model for fake reviews detection that shows high performance regardless of the classification algorithm. The model incorporates the most relevant verbal and non-verbal features and avoids collecting all existing features or random feature selection from the literature.
- Fourth, contrary to the literature on fake reviews detection, which focuses mainly on verbal features, our study proves that the performance of the fake reviews detection model can be improved by balancing this focus with more non-verbal features rather than focusing on one type of features only.
- Fifth, our model outperformed most of the well-known state-of-the-art fake review detection models with a high degree of interpretability, low complexity, and high performance.

II. RELATED WORKS

Studies on the fake reviews problem have varied in focus. Some studies have focused on determining the reasons for writing fake reviews. Some studies have focused on firms and people who have a higher possibility of posting fake reviews. Some studies have focused on techniques for writing fake reviews. Some studies have focused on the impacts on the growth of online reviews or on various stakeholders. Some studies have focused on the impacts on the market or society as a whole. Our interest lies in studies that focus on features and detection methods of fake reviews.

A. ML/DL-BASED FAKE REVIEWS DETECTION METHODS

Machine learning (ML) is an integral part of detecting fake reviews which has been considered as a classification problem. There are three types of machine learning: supervised, unsupervised, and semi-supervised.

The commonly used supervised learning methods include support vector machines (SVM) [3], [24], [27], [28], [29], [37], [47], [48], [49], [50], [51], [52], [53], logistic regression (LR) [27], [30], [31], [32], [53], Naïve Bayes (NB) [27], [28], [29], [33], [34], [35], [36], [38], [51], [53], k-nearest neighbor (kNN) [28], [39], [51], decision trees (DT) [27], [40], [41], random forest (RF) [28], [29], [39], [42], [43], Adaptive Boosting (Adaboost)[44], [45], Sparse Additive Generative Model (SAGE) [46], and multilayer perceptron (MLP) [29].

Because of the limited number of available labeled review datasets [3], some studies have used unsupervised learning methods, such as k-means clustering [54], [55], twice-clustering method [57], unsupervised similarity measurement [58], unsupervised generative Bayesian model [59], topic-sentiment joint probabilistic model [60], matrix iteration

algorithm [61], multi-iterative graph-based [62], statistics-based clustering algorithm [63], unified review deviation models [64], and lexicon-based model [56].

Some studies used semi-supervised learning methods, such as the positive unlabeled (PU) learning approach [65], [66], [67], hybrid positive unlabeled (PU) learning-based approach [68], co-training approach [69], [70], threshold-based detection method [71], multi-task method [72], semi-supervised learning framework (SPR2EP) [73], and Ramp One-Class SVM [74].

Ensemble learning models have also been used in certain studies because they are more effective in detecting fake reviews than single classifiers [25], [27], [89], [90].

Traditional machine learning algorithms are simple to implement, computationally inexpensive, and perform better than deep learning (DL) models on small datasets (see **TABLE 1**). However, with large-scale datasets, they produce lower performance than deep learning models and are not able to capture text sequences [9]. Deep learning models have been used to detect fake reviews, such as Convolutional Neural Networks (CNN) [75], [81], [82], [83], [84], Recurrent Convolutional Neural Networks (RCNN) [85], Long Short-Term Memory (LSTM) [86], [87], [88], Bidirectional Gated Recurrent Unit (Bi-GRU) with attention [76], Generative Adversarial Network (GAN) [77], [78], [79], and Bidirectional Encoder Representations from Transformers (BERT) [80].

Some other deep learning models are still not used in fake reviews detection, but based on the initial experiments performed by R. Mohawesh et al. [9], these models are promising such as convolutional-LSTM (C-LSTM), character-level C-LSTM, Hierarchical Attention Network (HAN), convolutional HAN, distilled version of BERT (DistilBERT), and Robustly Optimized BERT approach (RoBERTa).

Despite the great results that deep learning models have achieved, they lack a conceptual understanding to provide further justifications for the results [9]. All deep learning algorithms for detecting fake reviews are uninterpretable, and it is challenging to trust the model's performance and outcomes, whereas some deep learning models outperform other models on one dataset but underperform others on another [9]. (See **TABLE 1**)

B. THEORY-BASED FAKE REVIEWS DETECTION MODELS

Having a model with good accuracy for fake reviews detection is not sufficient to generalize and trust this model. The model results need sufficient theoretical justification and consistent testing results on different datasets because human deception behavior is complex [91]. Few deception detection models have been built based on fundamental theories, and we do not consider studies that have features or results that are aligned with some fundamental theories. However, we only consider studies that built their models or their feature selection based on fundamental theories from natural or social sciences, such as psychology, sociology, criminology, biology, or linguistics.

S. Banerjee et al. [4], [92] constructed a theoretical model that detects textual cues to differentiate between truthful and fake reviews. They synthesized four deception theories: information manipulation theory (IMT) [93], leakage theory [94], self-presentational theory [95], [96], and reality monitoring theory (RM) [97], [98]. The proposed model identifies four constructs in the content of the review: exaggeration, comprehensibility, specificity, and negligence. These constructs and their cues were tested using logistic regression with negative, positive, and moderate reviews. Accuracy ranged from 78% to 86%.

L. Zhou et al. [99], [100], [101] concentrated on detecting cues employed by deceivers in a textual-based computer-mediated communication context. They selected linguistic-based cues that were grouped into nine linguistic constructs: quantity, complexity, non-immediacy, affect, uncertainty, diversity, specificity, expressivity, and informality. These linguistic constructs were synthesized from media richness theory [102], channel expansion theory [103], interpersonal deception theory (IDT) [91], [104], the model of deceptive communication [105], criteria-based content analysis (CBCA) [106] which is the third stage of statement validity analysis (SVA), derived from the Undeutsch hypothesis [107], reality monitoring theory (RM), scientific content analysis (SCAN) [108], and verbal immediacy theory (VI) [109]. After considering only the essential cues of the linguistic constructs in the model. The classification accuracy ranged from 74% to 80%, the classification precision ranged from 70% to 80%, and the classification specificity ranged from 78% to 81% using discriminant analysis, logistic regression, and neural networks.

T. Qin et al. [110] examined linguistic cues to deceptive behavior across three methods of communication: text, audio, and face-to-face. They synthesized theories and criteria, including criteria-based content analysis (CBCA), reality monitoring theory (RM), scientific content analysis (SCAN), and interpersonal deception theory (IDT). They used linguistic cues which were grouped into seven categories: quantity, complexity, diversity, verb non-immediacy, uncertainty, specificity, and affect.

J. Li et al. [33] attempted to identify general linguistic differences between fake and truthful reviews; for this purpose, they used the research results of applied English linguistics and psycholinguistics with deception research and theories, including reality monitoring theory (RM) and interpersonal deception theory (IDT). They explored LIWC features (sentiment, spatial details, and first-person singular pronouns), part-of-speech (POS) features, and unigram features that distinguish informative (truthful) writing from imaginative (deceptive) writing. They experimented with these features using SAGE and SVM models, starting with intra-domain classification and extending to cross-domain classification. The classification accuracy ranged from 52% to 82%.

B. Kleinberg et al. [111] used named entities to detect verbal deception by modeling and capturing three theoretical concepts: the richness of detail, contextual embedding, and

TABLE 1. Comparison between detection methods for fake reviews.

Detection Method Category	Advantages	Disadvantages	Usage in studies of fake reviews detection
Traditional Supervised Machine Learning Models	• Good results with small datasets. • The most frequently used detection methods for fake reviews. • Model interpretability is not complex.	• High complexity and memory issues with large-scale datasets, such as the Amazon and Yelp datasets. • Efficient feature selection is required. • Labeled data are required when they are not widely available for online reviews.	[3], [24], [27]–[53]
Traditional Unsupervised Machine Learning Models	• Labeled data are not required which can handle the fact that most of the reviews' data are not labelled.	• Lack of successful detection methods.	[54]–[64]
Traditional Semi-supervised Machine Learning Models	• Ability to deal with a large quantity of unlabeled data with few quantity of labeled data in online reviews.	• Inability to correct false predictions, especially with outliers.	[65]–[74]
Deep Learning Models	• Great results and very high performance with large-scale online reviews datasets. • Flexible and efficient.	• Difficult to interpret and lack a conceptual understanding to provide further justification for the results.	[9], [75]–[88]

verifiability of details which were derived from reality monitoring theory (RM), criteria-based content analysis (CBCA), and verifiability approach (VA) [112] respectively.

C. Fuller et al. [113] developed an automated text-based deception detection model by selecting cue set from deception constructs drawn from deception theories, including interpersonal deception theory (IDT), information manipulation theory (IMT), reality monitoring theory (RM), four-factor theory [114], and self-presentational theory. They used confirmatory factor analysis to validate a set of deception constructs, including uncertainty, affect, specificity, and quantity. The overall accuracy ranged from 67% to 74% using logistic regression, decision trees, and neural networks.

D. Derrick et al. [115] built a theoretical model for detecting deceptive chat-based communication as a type of computer-mediated deception, which was mainly based on cognitive load theory [116], [117] and psychological studies that consider the increase of cognitive load as an indication for lying [94], [118], [119]. They hypothesized that deception in chatting affects word count, response time, lexical diversity, and the number of message edits. These four hypotheses were also supported by interpersonal deception theory (IDT), criteria-based content analysis (CBCA), and reality monitoring (RM).

X. Liu et al. [120] focused on the Newman-Pennebaker (NP) model [140] to explore linguistic features from the text for the purpose of detecting deception. They derived four theoretical features from the NP model: negative emotion terms, first-person singular pronouns, action verbs, and exclusive words. These theoretical features were tested using SVM and LR, and the accuracy was approximately 75%, but when these features were combined with other empirically-derived features and then optimized, the accuracy was improved to 86%.

D. Zhang et al. [42] found a set of non-verbal behavioral aspects of reviewers and evaluated their relevance for detecting fake reviews. They applied interpersonal deception theory (IDT) and the concept of non-verbal behavior to fake reviews detection by evaluating reviewers' posting and social behaviors. They combined non-verbal features of reviewers, such as membership, friendship, and posting with verbal features of reviews, such as review length, subjectivity, lexical validity, sentiment, lexical diversity, and self-reference diversity to improve the performance of fake reviews detection. The best accuracy of their model was 84% when using random forest.

T. Ong et al. [121] used expectancy theory [141] and the NP model to develop their hypotheses and show the differences between fake and truthful reviews based on information content, subjectivity, and readability.

K. Yoo et al. [122] examined the linguistic structure of fake and truthful hotel reviews using interpersonal deception theory (IDT) and the NP model. They tested several aspects of reviews, including quantity, lexical complexity, lexical diversity, immediacy, presence of brand names, and sentiment.

T. Chang et al. [124] used the rumor model [142] and its conceptual formula to profile the importance and ambiguity in fake reviews by extracting major features of review content: important attribute word, noun-verb ratio, and a specific quantifier. Using SVM, the overall precision of the proposed model was 59.6%.

X. Zhou et al. [123] built a fake news detection model focusing on news content and investigated the relationship between deception and fake news depending on the linguistic cues which were derived from four deception theories: information manipulation theory (IMT), reality monitoring

TABLE 2. Summary of theory-based models.

Study	Theories																							Constructs																									ML/DL	Performance
	Interpersonal Deception	CBCA / SVA / Undeutsch	Reality Monitoring	Scientific Content Analysis	Verbal Immediacy	Verifiability Approach	Information Manipulation	Self-presentational	Leakage	Four Factor	Newman-Pennebaker	Rhetorical Structure	Media Richness	Channel Expansion	Deceptive Communication	Cognitive Load	Dual-Process	Heuristic-Systematic	Elaboration Likelihood	Expectancy	Rumor	Truth-Default	Attitude-Behavior Consistency	Specificity	Quantity	Complexity / Comprehensibility	Non-immediacy	Diversity	Affect / Sentiment	Uncertainty	Objectivity / Subjectivity	Exaggeration	Negligence	Expressivity	Informality	Response Time	Spontaneous Corrections	Distraction	Importance	Consistency	Source Credibility	Information Rating	Coherence	Deviation in Behavior		
[4], [92]			√				√	√	√															√		√					√	√														LR	A=78-86%
[99]–[101]	√	√	√	√	√							√	√	√										√	√	√	√	√	√	√			√	√												LR, NN	A=74-80% P=70-80% S=78-81%
[110]	√	√	√	√																				√	√	√	√	√	√	√																-	-
[33]	√		√																					√			√		√																		SAGE, SVM	A=52-82%
[111]		√	√			√																		√																								-	-
[113]	√		√				√	√		√														√	√				√	√																	LR, DT, NN	A=67-74%
[115]	√	√	√													√									√			√						√	√												-	-
[120]											√															√	√		√								√										LR, SVM	A=75- 86%
[42]	√																								√			√	√		√			√											√	RF	A=84%
[121]											√									√				√	√					√																	-	-
[122]	√										√													√	√	√	√	√	√																			-	-
[123]		√	√				√			√														√	√			√	√		√			√													RF, XGBoost	A=63-76% F=65-76%
[29]									√												√	√																√	√				√					SVM, NB, RF, DT, NN	A=74-93% P=86-94% R=87-93% F=87-93%
[124]																				√				√								√						√										SVM	P=59.6%
[125]																	√							√	√			√		√								√	√	√		√					-	-
[126]																		√						√	√			√											√	√		√					-	-
[127]																		√							√			√																			-	-
[128]																		√						√				√											√		√					-	-
[129]																		√							√			√											√								-	-
[130]																		√																					√	√		√					-	-
[131]																		√						√															√								-	-
[132]																	√	√										√											√			√					-	-
[133]																		√						√				√		√									√	√		√					-	-
[134]																√								√				√											√	√		√					-	-
[135], [136]												√																													√								-	-
[137]												√																													√								-	-
[138], [139]												√																													√								LR	P=80% A=78% R=76%
f_{thr}	7	5	8	2	1	1	3	2	2	2	3	3	1	1	1	1	1	2	8	1	1	1	1	16	7	10	5	6	16	3	5	1	1	2	3	1	1	1	1	2	10	5	3	9		

CBCA: Criteria Based Content Analysis, SVA: Statement Validity Analysis, ML: Machine Learning, DL: Deep Learning,
LR: Logistic Regression, NN: Neural Network, SAGE: Sparse Additive Generative Model, SVM: Support Vector Machine, DT: Decision Tree,
RF: Random Forest, XGBoost: eXtreme Gradient Boosting, NB: Naïve Bayes
A: Accuracy, P: Precision, S: Specificity, F: F1-score, R: Recall
f_{thr}: Usage frequency in theory-based models.

theory (RM), four-factor theory, and the Undeutsch hypothesis. The deception-related attributes extracted from these theories were informality, diversity, subjectivity, sentiment, quantity, and specificity. Random forests (RF) and extreme gradient boosting (XGBoost) were utilized to experiment with these attributes and achieved accuracy from 63% to 76% and an F1-score from 65% to 76%.

Some studies [125], [126], [127], [128], [129], [130], [131], [132], [133], [134] focused on dual-process theory [143] or the widely used dual-process models: the heuristic-systematic model (HSM) [144] and the elaboration likelihood model (ELM) [145]. They developed hypotheses and conceptualized credibility analysis models for online reviews, demonstrating factors that affect the credibility of reviews, such as review sidedness, argument strength, internal consistency, reviewer credibility, information rating, review objectivity, external consistency, review framing, and structural factors.

Some studies [135], [136], [137], [138], [139] employed rhetorical structure theory (RST) [146] with vector space model (VSM) to detect systematic variations in coherence and structure between fake and truthful texts by analyzing the links between the component aspects of discourse. They used RST relations as features.

O. Popoola et al. [138], [139] built a fake reviews detection model from the RST relations using logistic regression. After testing the model, the accuracy, precision, and recall reached 78 %, 80%, and 76 %, respectively.

G. Shan et al. [29] built an online fake reviews detection system by adapting the truth-default theory (TDT) [147], leakage theory [94], and attitude-behavior consistency theory [148]. Three types of review inconsistency were conceptualized and introduced in their study: content inconsistency, rating-sentiment inconsistency, and reviewer language inconsistency. Rating-sentiment inconsistency was derived from coherence. Content inconsistency and reviewer language inconsistency were derived from correspondence. Non-verbal features for reviewer credibility and deviation in reviewing behavior were also incorporated. They tested their hypotheses using support vector machines (SVM), Naïve Bayes (NB), decision tree (DT), random forest (RF), and multilayer perceptron (MLP). After testing the system, accuracy ranged from 74% to 93%, precision ranged from 86% to 94%, recall ranged from 87% to 93%, and F1-score ranged from 87% to 93%.

From the summary in **TABLE 2**, we can see that the most frequently used theories are reality monitoring theory (RM), elaboration likelihood model (ELM), interpersonal deception theory (IDT), Undeutsch hypothesis and the derived criteria-based content analysis (CBCA) respectively. We can also see that the most frequently used constructs across all theory-based models are specificity, affect, complexity, source credibility, deviation in behavior, and quantity, respectively.

The deception detection models for computer-mediated texts, such as fake reviews detection models, mostly focus on verbal behavior through linguistic cues which are derived from cue theories of deception and ignore non-verbal behavior.

The literature on fake reviews detection lacks a comprehensive model that synthesizes relevant fundamental theories to analyze reviews based on their content and writers' behavior.

C. FEATURES USED TO IDENTIFY FAKE REVIEWS

1) REVIEWER FEATURES

These features cover the credibility and non-verbal behavior of the reviewer. The most commonly used reviewer features are:

- **First review ratio:** This feature measures the percentage that the reviewer posts the first review for any service or item. Fake reviews are meant to be posted as early as possible to significantly affect and deceive customers [121], [149].
- **Reviewing burstiness:** This feature computes whether the reviewer posts many reviews within a short period. Posting a large number of reviews in a short period is unusual and might indicate that the reviewer is a spammer and attempts to influence the rating [150], [151], [152].
- **Maximum number of reviews:** This feature measures the largest number of reviews written by a reviewer on a certain day. Truthful reviewers should not exceed a specific threshold in one day [150], and some studies have found that the threshold is five reviews [3], [153].
- **Extreme rating:** This feature computes whether the reviewer always uses extreme ranking, either the highest or the lowest rank on the scale. An extreme rating may indicate an attempt by a fake reviewer to enhance or lower the overall ranking of a product [150].
- **Ratio of positive reviews:** This feature measures the percentage of positive reviews posted by a reviewer, which may indicate spammer behavior if the percentage of positive reviews is high [3], [154].
- **Rating deviation:** This feature measures the divergence between a reviewer's rating and overall rating. An honest reviewer regularly rates items within the range of the overall ratings, whereas suspicious ratings differ significantly from the overall ratings provided by honest reviewers [150], [155].
- **Number of reviews:** This feature measures the reviewer involvement based on the total number of reviews posted. The number of reviews posted by a single reviewer is a crucial feature for distinguishing truthful reviewers from fake ones [43], and the reviewer, who posted more reviews, is more credible than the one who posted fewer reviews [156], [157], [158].
- **Number of friends/followers:** This feature measures a reviewer's sociability based on the total number of friends or followers. Sociability is an important indicator of a reviewer's credibility [156], [158], [159], [160]. A reviewer's reputation can also be measured by the

ratio of followers to the total number of followers and friends [161].
- **Membership length:** This feature measures the age of the reviewer's account from the date of registration. The accounts with longer memberships are more reliable [158], [162].

2) ONLINE REVIEW FEATURES

These features describe the content of the review. The most commonly used review features are:

- **Review length:** This feature measures the number of letters, words, phrases, or paragraphs in a review. Long reviews are considered more trustworthy than short ones for two reasons: lengthy reviews have a better probability of providing consumers with more detailed information [129], [156], and spammers often spend a relatively short period writing fake reviews [163]. Some studies [99], [100], [101], [110], [113] have mapped this feature theoretically using a quantity construct.
- **Reviews content similarity:** This feature measures the similarity between different reviews written by the same reviewer. The presence of similar reviews for different items may indicate that the reviewer is a spammer [3], [164] because they do not want to waste time writing new reviews [59], [149], [165]. The cosine similarity method is primarily used to capture maximum content similarity [150].
- **Bag of Words (BoW) (n-grams):** These features convert a review text into a vector form using individual or small groups of words to describe the frequency of content words, and some studies [33], [48] used these features to detect fake reviews. However, these features cannot capture the meaning of the text.
- **Term frequency-inverse document frequency (TF-IDF):** These features measure the importance and relevance of terms to a review, and are used as word evaluation schemes. R. Barbado et al. [44] used TF-IDF features with bigram features to enhance performance.
- **Language Inquiry and Word Count (LIWC):**LIWC is a text analysis tool that counts words into several categories. M. Ott et al. [24] and D. Plotkina et al. [166] used the LIWC features to detect fake reviews. It consists of a dictionary with a set of categories related to psychology. Some examples of features that can be extracted using LIWC or its dictionary are as follows:
  a) *Personal pronouns*: This feature measures the usage of first-person pronouns and third-person pronouns in a review. When first-person pronouns (e.g.: "I", "my", "we", "our" ...etc.) are used less and third-person pronouns (e.g.: "he", "him", "they", "them" ...etc.) are used more, this may indicate that the review is fake because liars try to disengage themselves from their false information, and because they do not have real experience [91], [167]. Some studies [99], [100], [101], [110] have mapped this feature theoretically using a non-immediacy construct.
  b) *Temporal and spatial information*: This feature measures the usage of locations and times in a review. Legitimate messages are expected to include more information about places and times than deceptive ones [97], [106]. Some studies [99], [100], [101], [110] have theoretically mapped this feature using a specificity construct.
  c) *Positive/negative affect:* These features measure the usage of terms with positive/negative meanings. Deceivers are expected to show more negative affect [91]. Some studies [99], [100], [101], [110] have theoretically mapped these features using an affect construct.
- **Coh-Matrix features:** Coh-Matrix is a tool used for cohesion and coherence measures for texts [168]. It has many features for texts, such as cohesion and narrativity. D. Plotkina et al. [166] have used Coh-Matrix features to detect fake online reviews.
- **Semantic features:** These features capture the meanings of words in a review so that switching between synonyms does not impact the results. Word embedding is one of the distributional semantic methods that represents words in vectors of fixed lengths [169]. Some studies [170], [171] have used word embedding with deep learning to detect fake reviews.
- **Stylometric features:** These features measure the writing style in a review. These features include both syntactic and lexical features. Syntactic features include the presence, frequency, and diversity of specific parts of speech (POS) patterns. Lexical features include lexical diversity, which was mapped theoretically with a diversity construct [99], [100], [101], [110], and lexical validity (ratio of misspellings), which was mapped theoretically with an informality construct [42], [99], [100], [101]. S. Shojaee et al. [47] have used a large number of syntactic and lexical features to identify fake reviews.
- **Discourse/rhetorical relations:** These features capture the coherence of a review text. These relations have different patterns [135], [136]. O. Popoola et al. [138], [139] used RST relations to build a fake reviews detection model.

D. DECEPTION THEORIES

Fake reviews are a form of deception [42], and deception is defined as a message purposely sent by a sender to form a false belief or conclusion by the receiver [167], [172]. Deception theories in psychology that study human cognition and behavior, discovered across disciplines, provide essential clues to deception. These theories have the potential to open new areas for research on large amounts of fake reviews data and can also support the development of fake reviews detection models. In this section, we consider and summarize deception theories that are highly cited, well-founded with

a clear methodology, and tested in the context of deception detection in computer-mediated text.

1) LEAKAGE THEORY

Leakage theory [94] is the first and possibly most prominent deception theory. This theory has dominated deception research, while most of the theories that follow can be considered modifications of the principles of leakage and deception cues [173]. The theory highlights the differentiation between two types of non-verbal behaviors: deception and leakage cues. Deception cues indicate that deception occurs, but they do not reveal what information is hidden. Leakage cues, on the other hand, expose hidden information, which can be considered a leakage of the truth. Both types of cues are mostly clear on the legs and feet as well as micro expressions on the face.

The theory primarily considers high-stakes lies, not insignificant or white lies. According to this idea, deceit must induce emotional reactions in the deceiver, which can only apply to high-stake lies. Negative emotions (guilt, fear, and delight) are deception-related. The consequences of a lie's acceptance or rejection are called stakes. Leakage and deception cues are more likely to occur when stakes are higher.

The theory has evolved over time [174], [175], but it actually kept the same concepts with limited changes, such as considering the verbal content and voice, such as inconsistent content and voice pitch. It considered pauses, indirect speech, long response latency, and speech errors. The only textual content cue that showed a difference was the number of self-references, which was lower in deception. This theory focuses on emotions and facial expressions that can only be clear in face-to-face communication with high-stake lies.

The main criticism of this theory is the lack of evidence that specific physiological states that are thought to be caused by telling a lie cannot be caused by many other emotional states, such as anxiety or fear [176].

2) FOUR-FACTOR THEORY

Four-factor theory [114] suggests that investigating four specific factors could lead to the discovery of deception cues. The four psychological factors are arousal, emotional reactions, cognitive effort, and attempted behavioral control. Arousal and emotional reactions appear to have the same meaning, as mentioned by the authors of the theory, which led other researchers to consider them as three factors [119].

The emotional reactions of deceivers depend on their personalities; they may feel guilty, fearful, or excited, which may cause non-immediacy, anxiety, speech mistakes (stuttering, omission of words, and repetition of words), speech hesitation, or an increased pitch. This theory posits that more cognitive effort is required when telling a lie than when telling the truth. Deceivers try to control their behavior to avoid detection and appear truthful.

The problem with using this theory in the context of online reviews is that it focuses on face-to-face non-verbal behavior.

3) INTERPERSONAL DECEPTION THEORY (IDT)

Interpersonal deception theory (IDT) [91], [104] aims to describe deception from the viewpoint of interpersonal communication in the presence of dynamic interaction between the sender and receiver. According to this theory, the deceiver will participate in strategic behavior changes in response to the receiver's doubts and show nonstrategic leakage signs. On the strategic side, the deceiver tries to manage information in his message, image, and behavior.

Source credibility is one of the attributes considered critical by this theory. Credibility measures the believability of a sender in terms of character, competence, composure, sociability, and dynamism. The IDT posits that as the sender's behavior deviates from normalcy, natural, reciprocity, ideal, and moderate involvement, it should be suspected.

The theory posits that deceivers have less specificity, immediacy, vocal relaxation, vocal pleasantness, and expressivity but show more negative affect, nervousness, arousal, uncertainty, noninvolvement, cognitive load, pauses, response latencies, and non-fluencies.

The theory posits that the behavior of deceivers is reflected in their language in terms of quantity, immediacy, specificity, uncertainty, and vagueness. The deceivers' deceptive messages are brief and reflect less quantity. They may employ leveling terms such as: ''always'' or ''everyone'' which minimize specificity. They may employ indirect forms of expression to modify or objectify their replies. They may use group/others references such as: ''they'' or ''we'' more than self-references such as: ''me'' or ''I'' which reflects non-immediacy. They are more likely to use past-tense verbs than present-tense verbs, reflecting non-immediacy in time.

The problem with using this theory in the context of online reviews is that it considers the interpersonal interactive form of communication, and posits that skilled deceivers are different from unskilled ones, making it very difficult to differentiate between skilled deceivers and truthful senders.

4) SELF-PRESENTATIONAL THEORY

The self-presentational theory [95], [96] first highlights that lying is a regular phenomenon in everyday situations. They disagreed with the concept that lying is usually a complex and guilt-inducing procedure with obvious and powerful cues. Instead, they argued that most false presentations are well-practiced and well-executed and that only little behavioral leakage remains. The theory claims that the behavior of both truth-tellers and deceivers is affected by emotions, cognitive load, and behavioral control. The theory posits that deceivers may appear less convincing, forthcoming, spontaneous, and pleasant, but tenser than truth-tellers.

Deceivers are less convincing than truth-tellers which means they have less plausibility, certainty, involvement, immediacy, and fluency. Deceivers are less forthcoming than truth-tellers which means they have shorter messages with fewer details, less complexity, and less quantity of information. Deceivers are less spontaneous than truth-tellers which

means their deceptive messages are less influenced by narrative mistakes than truthful ones. Deceivers are less pleasant than truth tellers, which means that their deceptive messages are less positive, less cooperative, and more negative. Deceivers are tenser than truth tellers, which means they are more anxious and nervous.

This theory took advantage of former theories such as reality monitoring theory (RM), verbal immediacy theory (VI), and criteria-based content analysis (CBCA). They used the cues mentioned in verbal immediacy theory (VI) [109] to measure immediacy, including the use of passive voice in deceptive messages instead of active voice and negations instead of assertions. Regarding fluency, they found that deceivers repeatedly used the same words and phrases. Regarding plausibility, they found that deceptive messages were more likely to be internally inconsistent or to reflect ambiguity.

The theory was criticized for focusing on similarities instead of the differences between deceivers and truth-tellers, which did not significantly improve deception detection [177].

5) REALITY MONITORING THEORY (RM)

Reality monitoring theory (RM) [97], [98] was originally used to examine the characteristics of memory, and it was not used for deception. This theory was used as a verbal deception detection method because the basis of reality monitoring is the fact that the quality of memories of actually experienced events is different from that of imagined events. The differentiation between memories of experience and imagination is derived from the Undeutsch hypothesis [107]. Researchers have argued that experienced events show truthfulness, whereas imagined events indicate deception.

Perceptual, contextual, and affective information are present in the memories of experienced events. Perceptual (sensory) information refers to sounds, smells, tastes, touches, or visual details that can be memorized from real experiences. Contextual information refers to temporal details (time of occurrence, time order, and duration) and spatial details (places of occurrence and positions of objects or people). Affective information refers to emotions and feelings.

According to reality monitoring, truthful statements exhibit clarity, re-constructability, and realism. Clarity refers to the sharpness and vividness of a statement. Re-constructability refers to the possibility of reconstructing a scenario. Realism refers to the plausibility and feasibility of a scenario.

Cognitive operations are present in the memories of imagined events. Cognitive operations refer to inferences and opinions during the description of a scenario, such as reasoning or thoughts.

In summary, clarity, re-constructability, realism, perceptual information, contextual information, and affective information are the attributes and content of truthful statements. Cognitive operations are present in the deceptive statements.

J. Masip et al. [178] showed that in comparison to deceptive statements, truthful statements contain more evidence of cognitive operations, which contrasts with RM theory.

6) CRITERIA-BASED CONTENT ANALYSIS (CBCA)

Statement validity analysis (SVA) is a verbal deception detection technique used in sexual offense cases to judge the validity of statements of child witnesses [179] and is based on the Undeutsch hypothesis [107]. This technique was also applied to older witnesses in different types of cases and has four stages [119]. The core stage of this technique is the third one which is criteria-based content analysis (CBCA) [106] in which nineteen different criteria are evaluated by qualified evaluators in the written interview. Each of these criteria is believed to appear more repeatedly in truthful statements than in deceptive ones because it is very complicated to fake them [179]. These criteria are judged using a scale of 0 to 2, where "0" indicates the absence of criterion, "1" indicates the presence of criterion, and "2" indicates the strong presence of criterion. It was found that using a scale of 0 to 4 (five points) is preferable to a scale of 0 to 2 (three points) because it is more sensitive to minor variations between deceptive and truthful statements [179]. The nineteen criteria were divided into four categories: general characteristics, specific contents, motivation-related contents, and offense-specific elements.

The general characteristics category contains criteria 1 to 3: logical structure, unstructured production, and quantity of details. Logical structure (criterion 1) refers to the coherence and logical consistency of a statement, but it does not refer to plausibility. Unstructured production (criterion 2) refers to presenting information without considering the order in the time sequence. Quantity of details (criterion 3) refers to the richness of details such as locations, times, people, things, and events.

The specific contents category contains the criteria 4 to 13: contextual embedding, descriptions of interactions, reproduction of conversation, unexpected complications during the incident, unusual details, superfluous details, accurately reported details misunderstood, related external associations, accounts of subjective mental state, and attribution of perpetrator's mental state. Contextual embedding (criterion 4) is present when events are timed and located in a specific place. Descriptions of interactions (criterion 5) include the presence of information that connects the witness with the perpetrator. Reproduction of conversation (criterion 6) refers to the presence of direct dialogue using actual quotations of exact words used by someone. Unexpected complications during the incident (criterion 7) refer to the presence of unexpected elements. Unusual details (criterion 8) refer to the presence of unique, unforeseen, or surprising details regarding individuals, things, or events. Superfluous details (criterion 9) refer to unnecessary details of the event. Accurately reported details misunderstood (criterion 10) refer to giving details that are beyond the understanding of the person. Related external associations (criterion 11) refer to the presence of events that are related to the incident but not part of it. Accounts of

subjective mental state (criterion 12) refer to describing how feelings change and thoughts are mentioned during the incident. Attribution of perpetrator's mental state (criterion 13) refers to describing motives, feelings, or thoughts of the perpetrator during the incident.

The motivation-related contents category contains criteria 14 to 18: spontaneous corrections, admitting lack of memory, raising doubts about one's own testimony, self-deprecation, and pardoning the perpetrator. Spontaneous corrections (criterion 14) refer to adding or correcting information of a previous statement. Admitting lack of memory (criterion 15) refers to forgetting, not remembering, or not knowing. Raising doubts about one's own testimony (criterion 16) refers to indicating the oddness and implausibility of a person's own statement. Self-deprecation (criterion 17) refers to revealing details that are negative or incriminating oneself. Pardoning the perpetrator (criterion 18) refers to excusing or failing to blame the perpetrator.

The offense-specific elements category contains only the details characteristic of the offense. Details characteristic of the offense (criterion 19) refer to describing parts that are considered typical for such a sort of offense by experts.

It is challenging to capture highly subjective criteria automatically; therefore, not all CBCA criteria are applicable for automatic deception detection [110].

7) SCIENTIFIC CONTENT ANALYSIS (SCAN)

Scientific content analysis (SCAN) [108] is a verbal deception detection technique. The SCAN procedure involves asking the person in question to write a detailed report of all the person's activities during a specific timeframe so that a reader with no prior knowledge of the situation can figure out what happened. Subsequently, a SCAN expert examines the handwritten statements using a set of criteria. Some SCAN criteria are assumed to be more probable to take place in truthful statements than in deceptive statements, while others are assumed to be more probable to take place in deceptive statements. SCAN has no fixed criteria list, but only twelve criteria were the focus of the research [107], [180], [181].

The twelve SCAN criteria: denial of allegations, social introduction, spontaneous corrections, lack of conviction and memory, structure of statement, emotions, objective and subjective time, out-of-sequence and extraneous information, missing information, change in language, first person singular past tense, and pronouns.

Denial of allegations (criterion 1) refers to immediately denying the allegations that indicate truthfulness. Social introduction (criterion 2) refers to the clarity of introducing another person when the writer shows ambiguity and failure, such as avoiding mentioning their names or relationships, which indicates deception. Spontaneous corrections (criterion 3) are equivalent to criterion 14 in the CBCA; however, they are considered an indication of deception in the SCAN. Lack of conviction and memory (criterion 4) is equivalent to criterion 15 in CBCA; however, they are considered an indication of deception in SCAN. Structure of statement (criterion 5) refers to the statement's overall balance between describing activities prior to the event, describing the event itself, and describing what happened immediately after the event, while unbalanced statements may indicate deception. Emotions (criterion 6) are equivalent to criterion 12 in CBCA, but in SCAN, the position of emotions in the statement for truth-tellers is considered throughout the story and for deceivers before the story's climax. Objective and subjective time (criterion 7) refer to the coverage of time periods in a statement, where objective time is the actual period of the event, and subjective time is the number of words used to describe it. The correspondence between objective and subjective times indicates truthfulness. Out-of-sequence and extraneous information (criterion 8) is equivalent to criteria 2 and 9 in CBCA; however, they are considered an indication of deception in SCAN. Missing information (criterion 9) refers to omitting some information using words such as: ''finally'', ''shortly thereafter'', ''sometime after'', and ''later on''. First-person singular past tense (criterion 10) refers to using ''I'' and past tense while describing the event, which indicates truthfulness. Pronouns (criterion 11) refer to using pronouns in the statement such as: ''I'', ''they'', ''he'', ''she'', ''my'', or ''his''. The presence of pronouns reflects responsibility, possession, and commitment since the absence of pronouns indicates deception. Change in language (criterion 12) refers to using different terms to describe one thing without sufficient justification, which indicates deception.

The SCAN method had limited attempts from research to validate it and lacks theoretical underpinning and standardization, and there is no theoretical justification for why these criteria can differentiate between truthful and deceptive statements [119], which is the main problem of this method. Another problem is that the common criteria between SCAN and CBCA conflict in the way of interpreting; SCAN experts consider them as indications of deception, but CBCA experts consider them as indications of truthfulness, although CBCA has more support from research [119].

8) VERIFIABILITY APPROACH (VA)

The verifiability approach (VA) [112] is a verbal deception detection technique. This method is based on two assumptions that put liars in a dilemma. In the first assumption, research has repeatedly shown that providing more details indicates truthfulness [119]; therefore, liars want to provide as many details as possible to make a truthful impression [182]. The second assumption is that liars prefer to avoid providing a large number of details because they fear that these details can be checked and that their lies will be discovered [182]. To balance these two opposing targets, liars utilize a strategy that focuses on providing unverifiable details [183].

The verifiability approach posits that information verifiability, or the possibility of verifying information without actually verifying it, can be used to distinguish between truthfulness and deception. Truth-tellers provide more verifiable perceptual, spatial, and temporal details than liars. The quantity of perceptual and contextual details reflects the richness

in details only; however, for this approach, the quality of details is of interest.

An interesting part of their experiments is that telling the participants that they need verifiable details to be checked enlarges the difference between truth-tellers and liars, which improves their ability to detect lies. The difference is enlarged because truth-tellers try to provide more verifiable details. For RM, CBCA, and SCAN, telling participants about the detection method will make it less effective because these methods focus on the number of details and do not distinguish between verifiable and unverifiable details [184]. Looking for verifiable details makes the verifiability approach more effective when dealing with the strategic behavior of deceivers [185]. This approach is not affected by deceivers who have good imaginations or deceivers who describe actual experiences that occurred at other times, similar to CBCA and RM [185].

The number of verifiable details can be assessed in relation to the total number of details. In particular, the quantity of verifiable perceptual and contextual details is divided by the overall quantity of verifiable and unverifiable perceptual and contextual details [186].

9) TRUTH-DEFAULT THEORY (TDT)

As the name of the theory suggests, the basic assumption of truth-default theory (TDT) [147] is that people trust and believe each other by default, which is called ''truth bias'' or ''truth default'' [114], [187]. This theory posits that most people communicate honestly most of the time; therefore, truth bias is beneficial for efficient communication and for improving the accuracy of deception detection, even if this bias causes people to be deceived sometimes. The people's inability to detect lies has been proven to be incorrect. This theory emphasizes the accuracy of deception detection and credibility assessment.

TDT presents a new perspective in deception research that differs from the previous dominant perspective that can be classified as cue theories. It focuses on contextualized communication content (content in the context) rather than on non-verbal cues. Understanding context requires having background information. Context information contains basic data, such as the description of an event, location, or tools. The existence of context information improves the accuracy of detecting deceptive statements [188]. The TDT disagrees with previous studies that posit emotional leakage, cognitive effort, arousal, or self-presentation as indications of deception. The theory states that focusing on non-verbal behavior increases the noise around the deception signal, which decreases the accuracy of deception detection because most lies can be detected by checking correspondence or confession.

This theory does not try to define a new set of deception cues; it criticizes the idea of observing deception cues. Communication content refers to what is said, whereas deception cues refer to how the message is said and how people behave when saying it.

A lack of correspondence and coherence in content triggers suspicion and may indicate deception. Checking correspondence is related to comparing contextualized communication content with known facts and evidence. When evidence is not available, assessing the plausibility of the content should take place, while typical and usual scenarios are known. Logical consistency is also referred to as coherence. Messages from the same person that are truthful and consistent do not conflict with each other. In general, correspondence is more effective in detecting deception than coherence. Coherence was not found to be useful in differentiating between liars and truthful people [189].

This theory posits that strategic questioning and active judgment increase the accuracy, which is not applicable in the case of online reviews. In general, because of the theoretical framework of the dominant perspective, it has been remarkably more useful than this new perspective in detecting computer-mediated deception [190]. The new perspective is still weak in terms of linguistics [191], and the unavailability of evidence or prior knowledge to fact-check the content in some contexts makes the usage of cues more useful [188], which is the case in the context of online reviews. This new perspective is helpful in interrogative contexts [192].

10) INFORMATION MANIPULATION THEORY (IMT)

Information manipulation theory (IMT) [93] is one of the most important theories from the new perspective. IMT shifts the focus from non-verbal cues to deceptive message design. It also suggested a method that can categorize deceptive messages, while previous studies were limited to only three types: distortion, omission, and falsification.

According to IMT, deceptive messages work by violating the principles that govern conversational exchanges in a covert manner. These principles are called Grice's maxims [193]: quantity, quality, manner, and relevance of information. Quantity refers to the expected amount of relevant information provided that makes a message informative. Quality refers to the expected information veracity. Manner refers to the expected avoidance of ambiguity. Relevance refers to expected relevant information based on a prior argument.

Deceptive messages are produced by manipulating the information. Information manipulation refers to violating one or more of Grice's four maxims. Quantity violations (omission) refer to changing the amount of sensitive information revealed in a message; therefore, it will be less informative. Quality violations (falsification) refer to information distortion, either by distorting sensitive information or fabricating the entire message. Manner violations (equivocation) refer to the use of traditionally ambiguous phrases and indirect expressions rather than clarity of expression in attempting to hide the truth. Relevance violations (evasion) refer to providing irrelevant information or failing to provide any contextually relevant information to divert attention.

IMT was extended to information manipulation theory 2 (IMT2) [194]. The IMT2 is a message-production theory for deception. It consists of three proposition groups: intentional

states, cognitive load, and information manipulation. This theory disagrees with the dominant perspective that cognitive load is higher in deception than in truthfulness. IMT2 shows that the most frequent type of deceptive message is quantity violations (omission), followed by quality violations (falsification), manner violations (equivocation), and relevance violations (evasion). IMT2 posits that people with high integrity have almost no motive to lie because they have nothing to hide.

The limitation of our work here is that we are unable to cover all existing deception theories, which may have different directions for interpreting the deception phenomena. We focus only on the most popular and influential deception theories.

III. OUR RESEARCH METHODOLOGY

The proposed method can be summarized as follows: First, ten deception theories were synthesized: leakage theory, four-factor theory, self-presentational theory, IDT, RM, CBCA, SCAN, VA, TDT, and IMT / IMT2. Second, the constructs of deception were collected from the synthesized deception theories. Third, important constructs were selected based on specific criteria. Fourth, a theoretical model was formulated based on the important constructs. Fifth, the theoretical model was empirically validated using online reviews datasets by preprocessing the data, extracting features, applying classification methods, and evaluating the model. (See **FIGURE 1**)

A. SELECTING CONSTRUCTS OF DECEPTION

From the synthesized deception theories and theory-based models, it can be observed that there was a great variety in the use of terms expressing the same constructs. Some studies, such as L. Zhou et al. [99], [100], [101], have determined the constructs that were derived from theories and used them as categories for features. Therefore, before selecting constructs, we put the factors, cues, or criteria derived from deception theories and share the same focus under the same construct. This facilitates understanding of the focus and usage frequency in deception theories (see **TABLE 4**). We determined a group of constructs from the synthesized deception theories and then selected or excluded from them based on a set of criteria, where each construct must satisfy all criteria to be selected (see **TABLE 3**).

The constructs from synthesized deception theories are specificity, quantity, non-immediacy, affect, uncertainty, informality, consistency, source credibility, deviation in behavior, diversity, complexity, spontaneous corrections, response time, body constructs, voice constructs, eye constructs, and face constructs.

The criteria used for construct selection were as follows:

- **Criterion 1 (C1):** ''used in theory-based models''. This refers to whether the construct was used by previous theory-based models of deception detection in computer-mediated texts (see **TABLE 2**).

TABLE 3. Summary of criteria-based constructs selection.

No.	Constructs	Criteria				Selection Decision
		C1	C2	C3	C4	
1	Specificity	Yes	Yes	Yes	Yes	Selected
2	Quantity	Yes	Yes	Yes	Yes	Selected
3	Non-immediacy	Yes	Yes	Yes	Yes	Selected
4	Affect	Yes	Yes	Yes	Yes	Selected
5	Uncertainty	Yes	Yes	Yes	Yes	Selected
6	Informality	Yes	Yes	Yes	Yes	Selected
7	Consistency	Yes	Yes	Yes	Yes	Selected
8	Source Credibility	Yes	Yes	Yes	Yes	Selected
9	Deviation in Behavior	Yes	Yes	Yes	Yes	Selected
10	Lexical Diversity	Yes	No	Yes	Yes	Excluded
11	Complexity	Yes	No	Yes	Yes	Excluded
12	Spontaneous Corrections	Yes	Yes	Yes	No	Excluded
13	Response Time	Yes	Yes	Yes	No	Excluded
14	Body constructs	No	No	No	No	Excluded
15	Voice constructs	No	No	No	No	Excluded
16	Eye constructs	No	No	No	No	Excluded
17	Face constructs	No	No	No	No	Excluded

C1: ''used in theory-based models'', C2: ''had consistent validation results'', C3: ''computationally measurable'', C4: ''available related data attributes''

- **Criterion 2 (C2):** ''had consistent validation results''. This refers to whether the construct had consistent results and received support across previous theory-based models of deception detection in computer-mediated texts.
- **Criterion 3 (C3):** ''computationally measurable''. This refers to the ability to measure a construct computationally in the context of the computer-mediated texts.
- **Criterion 4 (C4):** ''available related data attributes''. This refers to whether the public online reviews datasets have attributes for data points that enable us to measure the construct empirically.

The selected constructs that satisfied all four criteria (from no. 1 to 9 in **TABLE 3**) were specificity, quantity, non-immediacy, affect, uncertainty, informality, consistency, source credibility, and deviation in behavior. The excluded constructs that did not satisfy one or more criteria (from no. 10 to 17 in **TABLE 3**) were diversity, complexity, spontaneous corrections, response time, body constructs, voice constructs, eye constructs, and face constructs. The details of selection and exclusion are provided in the following two subsections.

1) SELECTED CONSTRUCTS AND THEORETICAL MODEL

Verbal constructs:

- Specificity
- Quantity
- Non-immediacy
- Affect
- Uncertainty
- Informality
- Consistency

Non-verbal constructs:

- Source Credibility
- Deviation in behavior

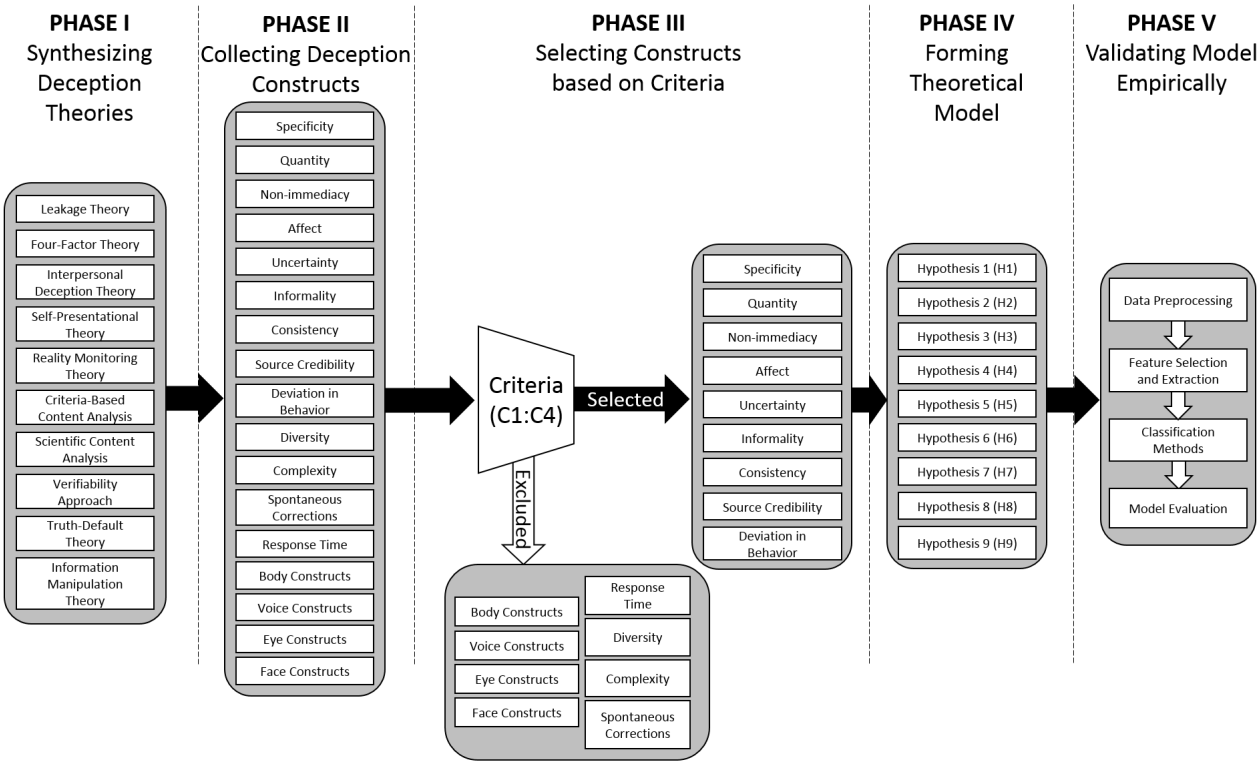

FIGURE 1. Our five-phase research methodology.

The above-selected constructs, the reasons for selection, and the formulation of the theoretical model are explained as follows:

Deception theories use specificity in the language to assess whether the statement is clear, precise, unique, exact, related to the subject, and the number of details included. In the IDT, lack of specificity was used to indicate deception. According to self-presentational theory, deceptive messages are less detailed. In RM, clarity, the existence of perceptual information, and contextual information (temporal and spatial) are used to indicate truthfulness. In CBCA, most of the criteria were only for specificity, criteria 3 to 11: contextual embedding, descriptions of interactions, reproduction of conversation, unexpected complications during the incident, unusual details, superfluous details, accurately reported details misunderstood, and related external associations. In addition to criterion 19: details characteristic of the offense. In SCAN, social introduction (criterion 2) is related to specificity. For VA, since it focuses on verifiable perceptual, spatial, and temporal details and all of them reflect specificity as an indication of truthfulness. In TDT, contextual content requires contextual information. In IMT and IMT2, quality and relevance are two maxims related to specificity to assess the message content. Specificity is the most frequently used construct in theory-based models and deception theories (see **TABLE 2**, **TABLE 4**). This was supported by theory-based models as a construct that discriminates between deception and truthfulness [4], [92], [111], [113], [123]. Therefore, the following hypothesis is proposed:

Hypothesis 1 (H1): Incorporating 'specificity' improves the prediction performance of the fake reviews detection model.

Deception theories use quantity in the language to assess the amount of information in the statement. IDT and self-presentational theory posit that deceptive messages are short and brief, reflecting less quantity while trying to hide and omit information. In SCAN, missing information (criterion 9) is related to quantity. In IMT and IMT2, quantity is one maxim to assess the message content. Quantity is one of the most frequently used constructs in theory-based models (see **TABLE 2**). It was supported by some theory-based models [99], [100], [101], [113], [115], [123] as a construct that discriminates between deception and truthfulness. Therefore, the following hypothesis is proposed:

Hypothesis 2 (H2): Incorporating 'quantity' improves the prediction performance of the fake reviews detection model.

Deception theories use non-immediacy in the language to assess warmth, closeness, and involvement. The four-factor theory posits that deceivers are less immediate than truth-tellers. Leakage theory points out that indirect speech is an indication of deception. The IDT and self-presentational theory focus on non-immediacy and non-involvement with clear verbal cues. In SCAN, immediacy is measured verbally from pronoun usage and verb tenses. Non-immediacy is one

of the most frequently used constructs in deception theories (see **TABLE 4**). This was supported by some theory-based models [99], [100], [101], [110], [122] as a construct that discriminates between deception and truthfulness. Therefore, the following hypothesis is proposed:

Hypothesis 3 (H3): Incorporating 'non-immediacy' improves the prediction performance of the fake reviews detection model.

Deception theories use affect in the language to assess emotions and feelings. Leakage theory, four-factor theory, interpersonal deception theory, and self-presentational theory posit that negative emotions indicate deception such as guilt, fear, anxiety, and nervousness. In RM, affective information and cognitive operations are used to assess emotions, thoughts, and opinions. CBCA has two criteria related to affect, criterion 12 and criterion 13: accounts of subjective mental state, and attribution of perpetrator's mental state. In SCAN, emotions (criterion 6) are related to affect. Affect is the second most frequently used construct in both deception theories and theory-based models (see **TABLE 2**, **TABLE 4**). The affect construct was supported by some theory-based models [99], [100], [101], [122], [123] as a construct that discriminates between deception and truthfulness. Therefore, the following hypothesis is proposed:

Hypothesis 4 (H4): Incorporating 'affect' improves the prediction performance of the fake reviews detection model.

Deception theories use uncertainty in the language to assess sureness, doubtfulness, ambiguity, or vagueness. The IDT and self-presentational theory posit that deceptive messages are vague, reflecting less certainty. The CBCA has two criteria for certainty: admitting lack of memory (criterion 15) and raising doubts about one's own testimony (criterion 16). In SCAN, lack of conviction and memory (criterion 4) is related to uncertainty. In IMT and IMT2, manner is one maxim related to uncertainty to assess the message content. Uncertainty is one of the most frequently used constructs in deception theories (see **TABLE 4**). This was supported by some theory-based models [99], [100], [101] as a construct that discriminates between deception and truthfulness. Therefore, the following hypothesis is proposed:

Hypothesis 5 (H5): Incorporating 'uncertainty' improves the prediction performance of the fake reviews detection model.

Deception theories use informality in the language to assess whether a person uses a non-fluent or unofficial language. Leakage theory, four-factor theory, IDT, and self-presentational theory posit that deceptive messages have more speech mistakes and non-fluencies than truthful ones. This was supported by some theory-based models [99], [100], [101], [123] as a construct that discriminates between deception and truthfulness. Therefore, the following hypothesis is proposed:

Hypothesis 6 (H6): Incorporating 'informality' improves the prediction performance of the fake reviews detection model.

Deception theories use consistency in the language to assess logic, plausibility, coherence, or correspondence. Leakage and self-presentational theories posit that deceptive content is less plausible and internally inconsistent. In RM, realism is considered an indication of truthful statements. In CBCA, logical structure (criterion 1) refers to logical consistency and coherence as indications for truthful statements. In TDT, correspondence and coherence are used to assess message credibility. This was supported by strong empirical evidence from G. Shan et al. [29], who investigated three types of inconsistency and their ability to distinguish between fake and truthful reviews. Therefore, the following hypothesis is proposed:

Hypothesis 7 (H7): Incorporating 'consistency' improves the prediction performance of the fake reviews detection model.

The IDT uses source credibility to assess the sender as a source of information if it is reliable, reputable, believable, and trustworthy. In IDT, source credibility is considered by the theory as a critical attribute to measure the believability of the sender in terms of character, competence, composure, sociability, and dynamism. Source credibility is one of the most frequently used constructs in theory-based models (see **TABLE 2**). Therefore, the following hypothesis is proposed:

Hypothesis 8 (H8): Incorporating 'source credibility' improves the prediction performance of the fake reviews detection model.

The IDT uses deviation in behavior to assess the extent to which a person's behavior departs from normal, average, usual, common, and expected behavior. The IDT posits that as the sender's behavior deviates from normalcy, natural, reciprocity, ideal, or moderate involvement, it should be suspected. Deviation in behavior is one of the most frequently used constructs in theory-based models (see **TABLE 2**). Some studies [125], [126], [128], [132], [133], [134] have called this construct external consistency or review consistency. This was supported by strong empirical evidence from D. Zhang et al. [42] and G. Shan et al. [29], who investigated a set of non-verbal behavioral aspects of reviewers and evaluated their relevance for detecting fake reviews. Therefore, the following hypothesis is proposed:

Hypothesis 9 (H9): Incorporating 'Deviation in behavior' improves the prediction performance of the fake reviews detection model.

Based on the above hypotheses, the theoretical model of fake reviews detection was formulated (see **FIGURE 2**). By synthesizing and selecting deception constructs from deception theories and then formulating the theoretical model of fake reviews detection in this section, we achieved our first research objective (RO1).

From **TABLE 4**, we can see that the most frequently used constructs in deception theories are specificity, affect, non-immediacy, uncertainty, and consistency.

2) EXCLUDED CONSTRUCTS

- Body Constructs

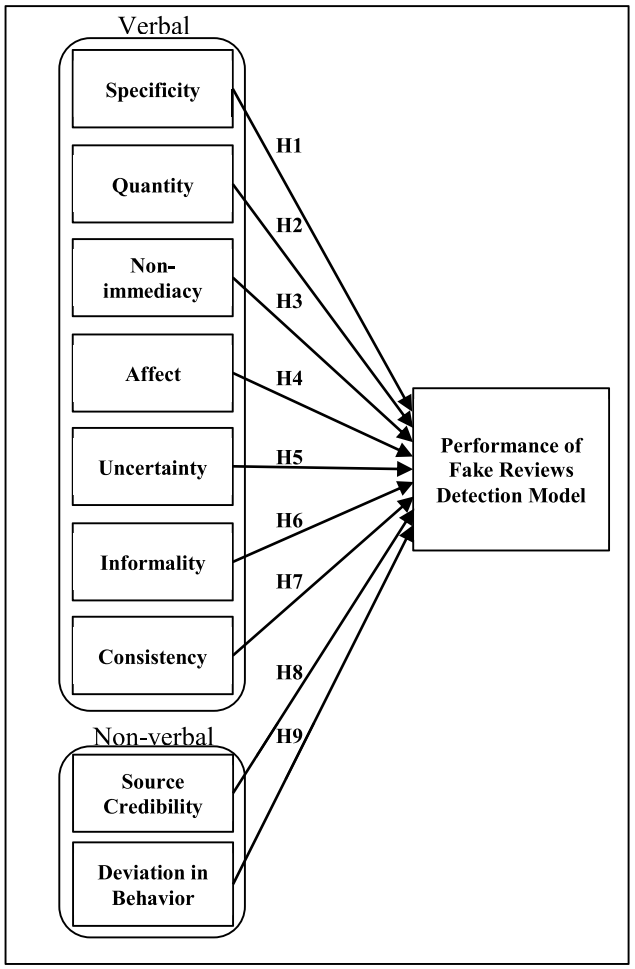

FIGURE 2. Theoretical model of fake reviews detection.

- Voice Constructs
- Eye Constructs
- Face Constructs
- Spontaneous Corrections
- Response Time
- Diversity
- Complexity

The above-excluded constructs and the reasons for exclusion are explained as follows:

The constructs of deception include constructs related to body movements, voice pitch, eye gaze, and facial expressions. These constructs are applicable to face-to-face interactive communication, videotaped interviews, or audiotaped interviews and are not applicable in the context of online reviews or any computer-mediated text. Therefore, based on the four criteria, all these constructs and their related factors, cues, or criteria are neglected.

Some constructs are computationally measurable, but no attributes in the available online reviews datasets can support their measurement. Spontaneous corrections and response time are examples of these constructs. Spontaneous corrections in language are used by deception theories to assess whether the person edits, revises, or rewrites his statements. According to self-presentational theory, deceivers are less spontaneous than truth-tellers, which means that their deceptive messages are less influenced by narrative mistakes than truthful ones. Spontaneous corrections are used as a criterion in CBCA and SCAN; however, they are considered an indication of truthfulness in CBCA and an indication of deception in SCAN. The response time is used by deception theories to assess the latency of responses and unfilled pauses due to cognitive effort. According to the leakage theory and IDT, deceivers may take more time to respond, especially when they are subjected to suspicion. Spontaneous corrections and response time were used in the theory-based model, but in a chat-based communication context using a self-implemented tool [115]. R. Banerjee et al. [195] empirically investigated both spontaneous corrections and response time as indications for deception using keystroke patterns, and the results were promising for both constructs. They did not use an off-the-shelf dataset but instead used a self-implemented key logger and crowd-sourcing approach. Therefore, spontaneous corrections and response time were neglected.

Some studies [99], [100], [101], [110], [121], [122] have considered lexical complexity and lexical diversity in language as constructs derived from deception theories or other psychological studies to differentiate between fake and truthful reviews. These two constructs showed conflicting results from one study to another. Therefore, lexical complexity and diversity were neglected.

TABLE 4. Usage summary of selected constructs in deception theories.

No.	Theories	Selected Constructs							
		Specificity	Quantity	Non-immediacy	Affect	Uncertainty	Informality	Consistency	Source Credibility	Deviation in Behavior
1	Leakage Theory			√	√		√	√		
2	Four-Factor Theory			√	√		√			
3	Interpersonal Deception Theory	√	√	√	√	√	√		√	√
4	Self-Presentational Theory	√	√	√	√	√	√	√		
5	Reality Monitoring Theory	√			√			√		
6	Criteria-Based Content Analysis	√			√	√		√		
7	Scientific Content Analysis	√	√	√	√	√				
8	Verifiability Approach	√								
9	Truth-Default Theory	√						√		
10	Information Manipulation Theory	√	√			√				
Usage frequency in deception theories		**8**	**4**	**5**	**7**	**5**	**4**	**5**	**1**	**1**

B. TEXTUAL DATA PREPROCESSING

1) LOWER CASING

The lower casing is the process of transforming all letters of text into lowercase. The main purpose of this process is to prepare words for the case-sensitive tools. In our case, the LIWC analysis library is case-sensitive, whereas all words in

its main dictionary are lowercase words. If there is only one letter in a word written in the upper case (capital letter), the category of the word will be missed by the LIWC analysis.

2) TOKENIZATION

Tokenization is the process of dividing a review text into smaller units, called tokens, which can be either words or sentences. Word tokenization was used before the LIWC analysis, spell checker, and review length calculation. Whitespace tokenization was used only before LIWC analysis because of its sensitivity in detecting the categories of each word. For spell checker and review length calculation, the NLTK toolkit library was used for this purpose. Sentence tokenization was used for two purposes: before the passive-voice detector (PassivePy) to calculate the percentage of passive-voice sentences and before the location of maximum affect calculation.

3) PUNCTUATION REMOVAL

Punctuation removal is the process of removing all punctuations in a text. This process cleans text before calculating the length of the review to avoid counting the punctuations as words in a review text.

C. FEATURE SELECTION AND EXTRACTION

To validate the theoretical model after selecting the verbal and the non-verbal constructs from deception theories, it is required to select features that can characterize these constructs on the one hand, and that can be measured from the review texts and the reviewer's behavior on the other hand. (For the symbols used in the equations below, see **TABLE 5**).

1) SPECIFICITY FEATURES

- Generalization terms usage (F1)
- Richness of details (F2)
- Richness of verifiable details (F3)

According to IDT, deceivers may employ generalization terms to avoid specificity. These terms are also called leveling terms, all-ness terms, over-generalization terms, words of absoluteness, or absolutist words. Examples of these terms include: ''all'', ''everything'', ''totally'', ''everyone'' ...etc. We used the validated list of nineteen absolutist words which was provided by M. Al-Mosaiwi et al.[196] to calculate the ratio of using these terms in a review text. Consider the following examples using generalization terms:

(1) ''**Everything** in this place is amazing''
(2) ''It's a **totally** bad restaurant. **Nothing** good''

Suppose that G is a set of generalization terms and R_t is the text of a review. Feature F_1 is the ratio of generalization words to the total number of words in a review text. The generalization terms usage in the text can be calculated as follows:

$$F_1 = \frac{|\{w : w \in R_t \wedge w \in G\}|}{|\{w : w \in R_t\}|}$$

where w is a word in the text of review.

TABLE 5. List of symbols for the feature extraction methods.

Symbol	Description
w	A word in a review
S	A sentence in a review
R_t	The whole text of the review
R_*	The star rating of a review [1 to 5]
d	A date in a dataset
R_d	A set of reviews on a date d
P_*	The average star rating of a reviewed product [1 to 5]
U	A set of all reviews for a user in a dataset
U_t	A set of reviews texts for user in a dataset
U_*	A set of reviews ratings for user in a dataset
G	A set of generalization terms
P	A set of words in LIWC category "perceptual processes"
TS	A set of words in LIWC categories "time" and "space"
V	A set of categories of named entities obtained by spaCy
PP	A set of words in LIWC category "personal pronouns"
FS	A set of words in LIWC category "1st personal singular"
FP	A set of words in LIWC category "1st personal plural"
TP	A set of words in LIWC categories "3rd personal singular" and "3rd personal plural"
PV	A set of sentences identified by PassivePy as passive voice
N	A set of words in LIWC category "negations"
AF	A set of words in LIWC category "affective processes"
PE	A set of words in LIWC category "positive emotion"
NE	A set of words in LIWC category "negative emotion"
CR	A set of words in LIWC category "certainty"
UC	A set of words in LIWC category "tentative"
MS	A set of words detected by Pyspellchecker as misspelled
VR	A set of words identified by spaCy as verbs
PR	A set of words identified by spaCy as present verbs
PS	A set of words identified by spaCy as past verbs
$polarity(R)$	Result of polarity for review obtained by TextBlob
$subject(R)$	Result of subjectivity for review obtained by TextBlob
$max(X)$	Maximum value in a set X
$index(x)$	Index of value x in a set
min_p	Minimum sentiment polarity [-1]
min_*	Minimum star rating [1]
max_p	Maximum sentiment polarity [1]
max_*	Maximum star rating [5]

The RM theory posits that perceptual and contextual information are present in memories of real experiences. Perceptual information refers to sounds, smells, tastes, touches, or visual details. Contextual information refers to temporal details (time of occurrence, time order, and duration) and spatial details (places of occurrence and positions of objects or people). CBCA posits that the quantity of details and contextual embedding are two signs of truthfulness. Quantity of details refers to the richness of details such as locations, times, people, things, and events. Contextual embedding is present when events are timed and located in a specific place. We used the LIWC category ''perceptual processes'' to capture the perceptual information from the review text. We also used the LIWC categories ''time'' and ''space'' to capture the temporal and spatial information from the review text. Therefore, these three categories were used to calculate the richness of the details in the review. Consider the following example of perceptual and contextual details:

(3) "I visited this **place** in the **morning**. It is **near** home. The food was **spicy** and **delicious**"

Suppose that TS is a set of words in the "time" and "space" categories and P is a set of words in the "perceptual processes" category inside the LIWC dictionary. Feature F_2 is the ratio of perceptual and contextual words to the total number of words in a review text. The richness of details can be calculated as follows:

$$F_2 = \frac{|\{w : w \in R_t \wedge w \in P \cup \text{TS}\}|}{|\{w : w \in R_t\}|}$$

SCAN has "social introduction" as a criterion which means that the clearness of introducing other persons is an indication of truthfulness, and avoiding mentioning their names or their relationship indicates deception. For VA, truth-tellers provide more verifiable details than liars. The existence of named entities almost meets the real verifiability standards [111], [197]. Therefore, we used NER in the spaCy library to extract eighteen categories of named entities: persons, facilities, money, organizations, geo-political entities, locations, dates, nationalities or religious groups, times, products, events, works of art, law documents, languages, percentages, quantities, ordinals, and cardinals. We then calculated the quantity of these named entities in the review text to represent the quantity of verifiable details. Assessing the quantity of verifiable details can be done in relation to the total quantity of details. In particular, the quantity of verifiable details is divided by the overall quantity of verifiable and unverifiable details [186]. In example (3), there are no verifiable details. Consider the following example for verifiable details:

(4) "When I visited **San Francisco** on **August 12th this year**, I went to this restaurant and at **9 AM**. **One** of them called **Omar** served me the food and he was very cheerful."

Suppose that V is a set of named entities in the review text. Feature F_3 is the ratio of words that provide verifiable details to the total number of words that provide verifiable or unverifiable details in a review text. The richness of verifiable details can be calculated as follows:

$$F_3 = \frac{|\{w : w \in R_t \wedge w \in V\}|}{|\{w : w \in R_t \wedge w \in P \cup \text{TS}\}| + |\{w : w \in R_t \wedge w \in V\}|}$$

2) QUANTITY FEATURES

- Length of the review (F4)

Self-presentational theory and IDT posit that the deceptive messages are short and brief which reflects less quantity of information.

We used the number of words to calculate the quantity of the review. The length of the review can be calculated as follows:

$$F_4 = |\{w : w \in R_t\}|$$

3) NON-IMMEDIACY FEATURES

- First-person singular pronouns usage (F5)
- First-person plural pronouns usage (F6)
- Third-person pronouns usage (F7)
- Present-tense verbs usage (F8)
- Past-tense verbs usage (F9)
- Passive voice usage (F10)
- Negations usage (F11)

IDT, SCAN, and leakage theory posit that deceivers may use group or other references (first-person plural pronouns or third-person pronouns) such as: "they" or "we" more than self-references (first-person singular pronouns) such as: "me" or "I" which reflects non-immediacy. SCAN has a criterion for "pronouns" which considers using personal pronouns in the statement such as: "I", "they", "he", "she", "my", or "his". The presence of pronouns reflects responsibility, possession, and commitment since the absence of pronouns indicates deception. POS tagging provides one tag type for all levels of personal pronouns. Therefore, we used the LIWC category "personal pronouns" and its subcategories "1st personal singular", "1st personal plural", "3rd personal singular", and "3rd personal plural" to capture the usage of pronouns in the review text. Consider the following examples for first-person singular pronouns usage, first-person plural pronouns usage, and third-person pronouns usage, respectively:

(5) "For **me** it was a great meal. **I** sat for a few minutes until **my** order was prepared."

(6) "**We** went to that place in the morning, and everyone served **us** perfectly."

(7) "**They** delayed my order for no reason, but **their** meals are very tasty."

Suppose that FS is a set of words in the "1st personal singular" category and PP is a set of words in the "personal pronouns" category inside the LIWC dictionary. Where FS $\subset$ PP. Feature F_5 is the ratio of first-person singular pronouns to the total number of pronouns in a review text. The first-person singular pronouns usage can be calculated as follows:

$$F_5 = \frac{|\{w : w \in R_t \wedge w \in \text{FS}\}|}{|\{w : w \in R_t \wedge w \in \text{PP}\}|}$$

Suppose that FP is a set of words in the "1st personal plural" category inside the LIWC dictionary. Where FP $\subset$ PP. Feature F_6 is the ratio of first-person plural pronouns to the total number of pronouns in a review text. The first-person plural pronouns usage can be calculated as follows:

$$F_6 = \frac{|\{w : w \in R_t \wedge w \in \text{FP}\}|}{|\{w : w \in R_t \wedge w \in \text{PP}\}|}$$

Suppose that TP is a set of words in the "3rd personal singular" and "3rd personal plural" categories inside LIWC dictionary. Where TP $\subset$ PP. Feature F_7 is the ratio of third-person pronouns to the total number of pronouns in a review text. The third-person pronouns usage can be calculated as follows:

$$F_7 = \frac{|\{w : w \in R_t \wedge w \in \text{TP}\}|}{|\{w : w \in R_t \wedge w \in \text{PP}\}|}$$

For IDT, deceivers are more likely to use past-tense verbs than present-tense verbs, reflecting non-immediacy in time.

However, for SCAN, using the past tense when describing an event indicates truthfulness. We used POS tagging in the spaCy library to identify verbs and their sub-categories, including past and present tenses in a review text. Consider the following examples for using present tense and past tense, respectively:

(8) "We **go** there almost every evening."
(9) "I **visited** this coffee once. The latté **was** tasty."

Suppose that PR is a set of present verbs and VR is a set of all verbs. Where PR $\subset$ VR. Feature F_8 is the ratio of present verbs to the total number of verbs in a review text. The present-tense verbs usage can be calculated as follows:

$$F_8 = \frac{|\{w : w \in R_t \wedge w \in \text{PR}\}|}{|\{w : w \in R_t \wedge w \in \text{VR}\}|}$$

Suppose that PS is a set of past verbs. Where PS $\subset$ VR. Feature F_9 is the ratio of past verbs to the total number of verbs in a review text. The past-tense verbs usage can be calculated as follows:

$$F_9 = \frac{|\{w : w \in R_t \wedge w \in \text{PS}\}|}{|\{w : w \in R_t \wedge w \in \text{VR}\}|}$$

The self-presentational theory posits that the usage of passive voice in the deceptive message is more than the usage of active voice to reduce immediacy. We used the PassivePy tool [198] to identify sentences in passive voice from the review text. Consider the following examples of passive voice:

(10) "The place **is wonderfully arranged** to provide privacy to everyone. All meals **are professionally served**."

Suppose that PV is a set of passive-voice sentences. Feature F_{10} is the ratio of passive voice sentences to the total number of sentences in a review's text. The passive voice usage can be calculated as follows:

$$F_{10} = \frac{|\{S : S \in R_t \wedge S \in \text{PV}\}|}{|\{S : S \in R_t\}|}$$

where S is a sentence in a review's text.

The self-presentational theory posits that the usage of negations in the deceptive message is more than the usage of assertions to reduce immediacy. We used the LIWC category "negations" to capture the negations usage in the review text. Consider the following example of negations usage:

(11) "It **wasn't** as expected, I **couldn't** have imagined this level of irresponsibility."

Suppose that N is a set of words in the "negations" category inside the LIWC dictionary. Feature F_{11} is the ratio of negations to the total number of words in a review text. The negations usage can be calculated as follows:

$$F_{11} = \frac{|\{w : w \in R_t \wedge w \in N\}|}{|\{w : w \in R_t\}|}$$

4) AFFECT FEATURES

- Sentiment polarity (F12)
- Sentiment subjectivity (F13)
- Positive emotion words usage (F14)
- Negative emotion words usage (F15)
- Location of maximum affect (F16)

The leakage and four-factor theories posit that emotional reactions are related to deception. Deceivers may feel guilty, fearful, or excited. The self-presentational theory posits that deceptive messages are less positive and more negative. Deceivers are more anxious and nervous. CBCA has a criterion for the subjective mental state which refers to describing how feelings change and mentioning the thoughts, and another criterion for the perpetrator's mental state which refers to describing the motives, feelings, or thoughts of the perpetrator during the incident. We used the TextBlob sentiment analyzer to extract two dimensions of sentiment from a review: polarity and subjectivity. Polarity indicates whether a sentence is positive or negative. Subjectivity indicates whether the judgment is based on personal opinion or factual information. We also used the LIWC category "affective processes" including its subcategories "negative emotion" and "positive emotion" to capture the usage of positive and negative emotion words in the review text. Consider the following examples for positive emotion usage, negative emotion usage, and subjective sentence, respectively:

(12) "I am **very happy** with this experience. It was **amazing**."
(13) "It was a **bad** experience. Everything is **disgusting**."
(14) "It's a **wonderful** place."

The sentiment polarity can be calculated as follows:

$$F_{12} = \textit{polarity}\,(R_t)\,, \quad [-1, 1]$$

The sentiment subjectivity can be calculated as follows:

$$F_{13} = \textit{subject}\,(R_t)\,, \quad [0, 1]$$

Suppose that PE is a set of words in the "positive emotion" category and AF is a set of words in the "affective processes" category inside the LIWC dictionary. Where PE $\subset$ AF. Feature F_{14} is the ratio of positive emotion words to the total number of affect words in a review text. The positive emotion words usage can be calculated as follows:

$$F_{14} = \frac{|\{w : w \in R_t \wedge w \in \text{PE}\}|}{|\{w : w \in R_t \wedge w \in \text{AF}\}|}$$

Suppose that NE is a set of words in the "negative emotion" category inside the LIWC dictionary. Where NE $\subset$ AF. Feature F_{15} is the ratio of negative emotion words to the total number of affect words in a review text. The negative emotion words usage can be calculated as follows:

$$F_{15} = \frac{|\{w : w \in R_t \wedge w \in \text{NE}\}|}{|\{w : w \in R_t \wedge w \in \text{AF}\}|}$$

SCAN has one criterion for emotions but it does not only consider the existence of emotions, it rather considers the position of emotions in a statement. For truth-tellers, the emotions are expected to be present throughout the story, while deceivers are expected to show emotions before the story's climax. A. Sepehri et al. [199] conducted analyses at the sentence level to investigate the location of maximum

emotion in deception and truth using two datasets, one for news and the other for reviews. Their findings were consistent with the SCAN criterion, and they found that the maximum emotion location for deceptive texts was at the beginning. They suggested a method to measure the location of the maximum affect: (a) tokenize each text at the sentence level, (b) score each sentence using the LIWC category "affective processes" to calculate the emotion ratio, (c) determine the sentence number with the maximum score of affect, and (d) divide the sentence number by the number of sentences in the text. A smaller result indicates that the location of maximum affect is at the beginning. They found no difference in patterns between using sentiment analysis and LIWC-based affect scores. Therefore, we used the LIWC-based method to calculate the location of the maximum affect. Consider the following example of maximum affect located at the beginning of the statement:

(15) "I am **frustrated**, **angry**, and very **upset**. I thought the **worst** thing about this place was the long waiting time. It will be the last time I visit this place."

For each sentence in a review text, it is required to calculate the ratio of affect words to the total number of words, so it will be possible to get the index of the sentence that has the maximum score and determine its location. The location of maximum affect can be calculated as follows:

$$\forall S \in R_t, \quad f_{16}(S) = \frac{|\{w : w \in S \wedge w \in \text{AF}\}|}{|\{w : w \in S\}|}$$
$$F_{16} = \frac{index(\max(f_{16}(S)))}{|\{S : S \in R_t\}|}$$

5) UNCERTAINTY FEATURES

- Certainty words usage (F17)
- Uncertainty words usage (F18)

IMT posits that one of the information manipulation methods used by deceivers is manner violations which refer to the use of traditionally ambiguous phrases and indirect expressions rather than clarity of expression in attempting to hide the truth. We used the LIWC categories "certainty" and "tentative" to capture the usage of certainty and uncertainty words in the review text. Consider the following examples of certainty words usage and uncertainty words usage respectively:

(16) "This is **exactly** what I need. I'm **sure** this is the best Chinese restaurant in my area"

(16) "I was **wondering** why this place **seems** **almost** perfect to me."

Suppose that CR is a set of words in the "certainty" category inside the LIWC dictionary. Feature F_{17} is the ratio of certainty words to the total number of words in a review text. The certainty words usage can be calculated as follows:

$$F_{17} = \frac{|\{w : w \in R_t \wedge w \in \text{CR}\}|}{|\{w : w \in R_t\}|}$$

Suppose that UC is a set of words in the "tentative" category inside the LIWC dictionary. Feature F_{18} is the ratio of uncertainty words to the total number of words in a review text. The uncertainty words usage can be calculated as follows:

$$F_{18} = \frac{|\{w : w \in R_t \wedge w \in \text{UC}\}|}{|\{w : w \in R_t\}|}$$

6) INFORMALITY FEATURES

- Misspelled words ratio (F19)

Leakage theory, four-factor theory, IDT, and self-presentational theory posit that deceptive messages have more speech mistakes and non-fluencies than truthful ones. We used Pyspellchecker, spell checking tool, to detect misspelled words in a review text. Consider the following example of misspelled words:

(18) "**typcially** what I need from any **restrant**."

Suppose that MS is a set of words detected as misspelled in the review's text. Feature F_{19} is the ratio of misspelled words to the total number of words in a review text. The misspelled words ratio can be calculated as follows:

$$F_{19} = \frac{|\{w : w \in R_t \wedge w \in \text{MS}\}|}{|\{w : w \in R_t\}|}$$

7) CONSISTENCY FEATURES

- Rating-sentiment inconsistency (F20)

The self-presentational theory posits that deceptive messages are more likely to be internally inconsistent. We used the TextBlob sentiment analyzer to extract the polarity of a review text. As the polarity lies in the range of $[-1, 1]$ and the star rating lies in the range of $[1, 5]$, we normalized both of them to lie in the range of $[0, 1]$. Consider the following examples of inconsistency between sentiment and rating:

(19) "I **love** to eat breakfast here every morning." ★☆☆☆☆

Suppose that max_p is the maximum sentiment polarity [1], min_p is the minimum sentiment polarity $[-1]$, max_* is the maximum star rating [5], min_* is the minimum star rating [1], and R_* is the review's star rating. Feature F_{20} is the absolute difference between the normalized sentiment polarity of the text and the normalized star rating of a review. The rating-sentiment inconsistency can be calculated as follows:

$$F_{20} = \left| \frac{polarity\,(R_t) - min_p}{max_p - min_p} - \frac{R_* - min_*}{max_* - min_*} \right|$$

The above verbal features from F1 to F20, the textual preprocessing methods for review text, and the programming tools that were used to extract these features were summarized in **FIGURE 3**.

8) SOURCE-CREDIBILITY FEATURES

- Number of reviews (F21)

IDT focuses on source credibility as a measure of the source's believability in terms of different aspects including dynamism and sociability. We used the number of reviews posted by the reviewer to measure dynamism and sociability.

Suppose that U is a set of all reviews posted by a user in a dataset. Feature F_{21} is the number of reviews posted by the

TABLE 6. Summary of selected features for our unified model.

Construct Type	Construct	Feature No.	Feature Description	Feature Extraction Method
Verbal Constructs	Specificity	F1	Generalization terms usage	The ratio of generalization words in a review text.
		F2	Richness of details	The ratio of details in the review text, including perceptual, temporal, and spatial information.
		F3	Richness of verifiable details	The ratio of verifiable details in a review text, including persons, facilities, money, organizations, locations, dates, products, ...etc.
	Quantity	F4	Length of the review	The number of words in a review text.
	Non-immediacy	F5	First-person singular pronouns usage	The ratio of first-person singular pronouns to all personal pronouns in a review text.
		F6	First-person plural pronouns usage	The ratio of first-person plural pronouns to all personal pronouns in a review text.
		F7	Third-person pronouns usage	The ratio of third-person pronouns to all personal pronouns in a review text.
		F8	Present-tense verbs usage	The ratio of present-tense verbs to all verbs in a review text.
		F9	Past-tense verbs usage	The ratio of past-tense verbs to all verbs in a review text.
		F10	Passive voice usage	The percentage of sentences written in passive voice to all sentences in a review text.
		F11	Negations usage	The ratio of negations in a review text.
	Affect	F12	Sentiment polarity	The lexicon-based sentiment polarity calculation
		F13	Sentiment subjectivity	The lexicon-based sentiment subjectivity calculation
		F14	Positive emotion words usage	The ratio of positive emotion words to affect words in a review text.
		F15	Negative emotion words usage	The ratio of negative emotion words to affect words in a review text.
		F16	Location of maximum affect	The location of a sentence which has the maximum affect in a review text.
	Uncertainty	F17	Certainty words usage	The ratio of certainty words in a review text.
		F18	Uncertainty words usage	The ratio of uncertainty words in a review text.
	Informality	F19	Misspelled words ratio	The ratio of misspelled words in a review text.
	Consistency	F20	Rating-sentiment inconsistency	The absolute difference between the normalized polarity of review text and the normalized star rating of the review.
Non-verbal Constructs	Source credibility	F21	Number of reviews	The total number of reviews posted by the reviewer.
	Deviation in behavior	F22	Rating deviation	The normalized absolute difference between the user's rating and the average rating of the product.
		F23	Extreme rating ratio	The ratio of extreme negative or extreme positive reviews rating [1-star or 5-star] rated by the user.
		F24	Maximum reviewing frequency	The highest number of reviews posted by reviewer in one day

user and it can be calculated as follows:

$$F_{21} = |U|$$

9) DEVIATION-IN-BEHAVIOR FEATURES

- Rating deviation (F22)
- Extreme rating ratio (F23)
- Maximum reviewing frequency (F24)

IDT posits that the deceiver's behavior deviates from the normal and ideal behavior. Therefore, the reviewer's rating behavior is important to be captured and compared with the average rating behavior for one product to measure how much it deviates from the normal behavior.

Suppose that P_* is the average star rating of a reviewed product. Feature F_{22} is the normalized absolute difference between the user's rating on a product and the average rating of the reviewed product. The rating deviation can be calculated as follows:

$$F_{22} = \frac{|R_* - P_*|}{max_* - min_*}$$

Fake reviewers tend to rate the products either with the highest rate (5 stars) or with the lowest rate (1 star) to enhance or damage reputation [150], which is deviated from truthful reviewers' rating behavior. Therefore, the extreme rating ratio is important to be measured.

Suppose that U_* is a set of reviews ratings for a user in a dataset. Feature F_{23} is the ratio of reviews with extreme star ratings $\{1, 5\}$ to the total number of reviews posted by the user. The extreme rating ratio can be calculated as follows:

$$F_{23} = \frac{|\{R_* : R_* \in U_* \wedge R_* \in \{1, 5\}\}|}{|\{R_* : R_* \in U_*\}|}$$

Truthful reviewers are not expected to post more than two reviews per day while fake reviewers are expected to write down about seven reviews in one day [150] which is deviated from truthful posting reviewers' behavior. Therefore, the maximum reviewing frequency is important to be measured.

Suppose that R_d is a set of all reviews posted on a specific date d and U is a set of all reviews posted by the user in a dataset. The feature F_{24} is the maximum number of reviews posted by a user in one day. The maximum reviewing frequency can be calculated as follows:

$$\forall d, \quad f_{24}(d) = |R_d \cap U|$$
$$F_{24} = \max(f_{24}(d))$$

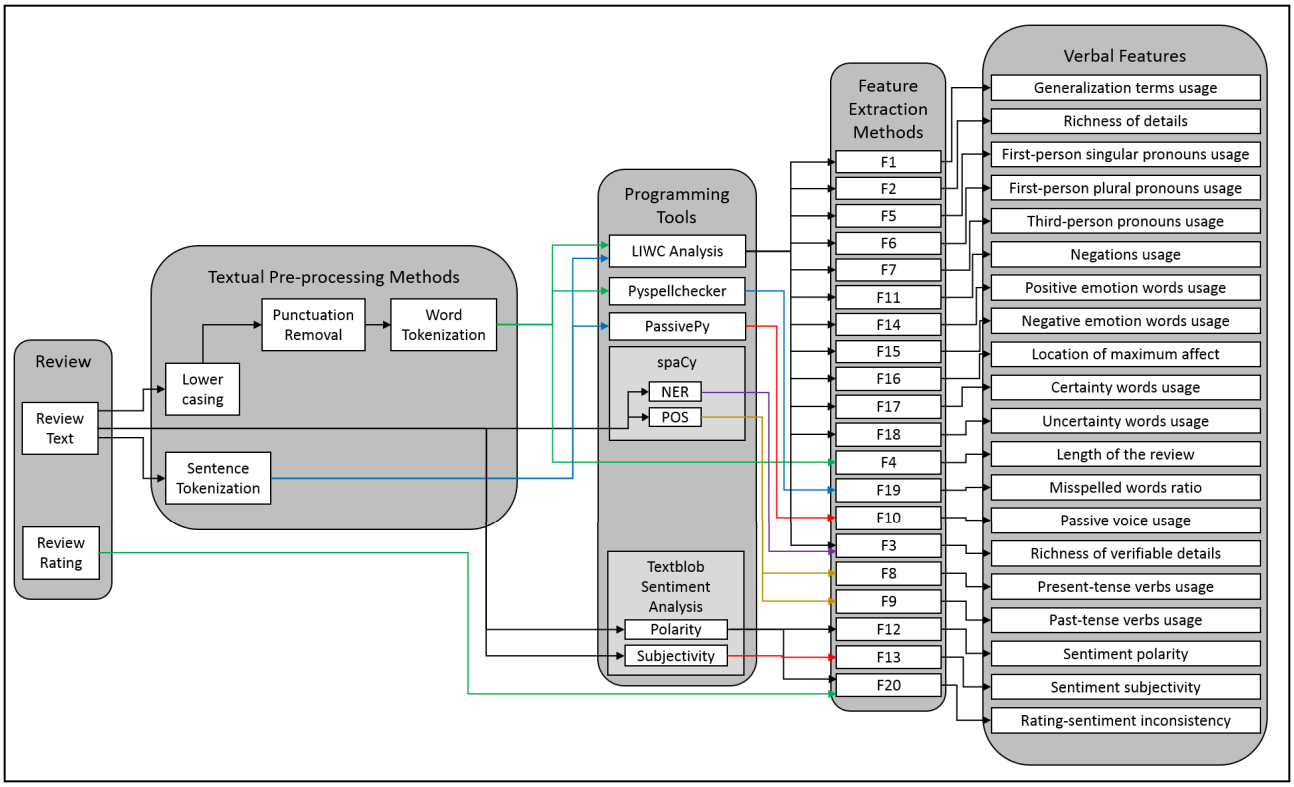

FIGURE 3. Verbal features extraction for our model.

By specifying twenty-four features from the deception constructs and their extraction methods, we achieved the second and fourth research objectives (RO2 and RO4).

IV. EXPERIMENTAL VALIDATION

A. PROPOSED FAKE REVIEWS DETECTION MODEL

To validate the theoretical model empirically, we proposed a fake reviews detection model that can be implemented and validated.

The first stage of the proposed model uses the review text data attribute and pre-processes the text based on the procedures mentioned in the above section (TEXTUAL DATA PREPROCESSING). Non-verbal attributes do not require any pre-processing; these attributes include date, rating, user ID, and product ID.

The next stage uses the pre-processed text data with non-verbal attributes and extracts the verbal and non-verbal features based on the extraction methods mentioned in the above section (FEATURE SELECTION AND EXTRACTION). In general, verbal features are more complicated to extract and computationally costly than non-verbal features (see **FIGURE 3**).

The last stage of the proposed model is the classification stage, in which the extracted features are passed to a trained machine-learning model to classify the review as either fake or truthful.

The model stages are summarized in **FIGURE 4**. By developing a fake reviews detection model based on the selected verbal and non-verbal features, we achieved the third research objective (RO3).

B. ONLINE REVIEWS DATASETS

Finding a ground-truth dataset for the problem of fake reviews is difficult; however, the ground-truth data are not guaranteed. The most popular and near-ground truth datasets are YelpChi, YelpNYC, and YelpZip [3], [163], [200] (see **TABLE 7**), which were crawled from the Yelp website and used widely to benchmark the models for spam detection. YelpChi contains reviews from the Chicago area of a group of hotels and restaurants. YelpNYC contains reviews from New York City of a group of restaurants. YelpZip contains reviews from New York City, New Jersey, Vermont, Connecticut, and Pennsylvania of a group of restaurants. The data attributes include product ID, user ID, date, rating, review text, and label for each review. Yelp filters suspicious reviews but keeps them public. Therefore, the filtered reviews were considered fake, whereas the recommended reviews were considered truthful in the three datasets. These datasets do not offer adequate behavioral details because of the high proportion of reviewers with a single review and products with a single review [201], which makes it important to extract verbal and non-verbal features when using these datasets.

C. CLASSIFICATION METHODS

Four classification algorithms, logistic regression (LR), Naïve Bayes (NB), decision tree (DT), and random

TABLE 7. Descriptive statistics of Yelp datasets.

Dataset	Business Domain	Total Number of Reviews	Number of Truthful Reviews	Number of Fake Reviews (%)	Number of Reviewers (%spammer)	Mean length of a fake review	Mean length of a truthful review	Mean number of reviews posted by a fake reviewer	Mean number of reviews posted by a truthful reviewer	Percentage of extreme ratings (1 or 5) in fake reviews	Percentage of extreme ratings (1 or 5) in truthful reviews
YelpChi	Hotels & Restaurants	67,395	58,476	8,919 (13.2%)	38,063 (20.33%)	105	148	1.153	1.920	56.67%	40.77%
YelpNYC	Restaurants	359,052	322,167	36,885 (10.3%)	160,225 (17.79%)	82	123	1.294	2.434	55.77%	41.84%
YelpZip	Restaurants	608,598	528,132	80,466 (13.2%)	260,277 (23.91%)	88	122	1.293	2.638	58.90%	41.20%

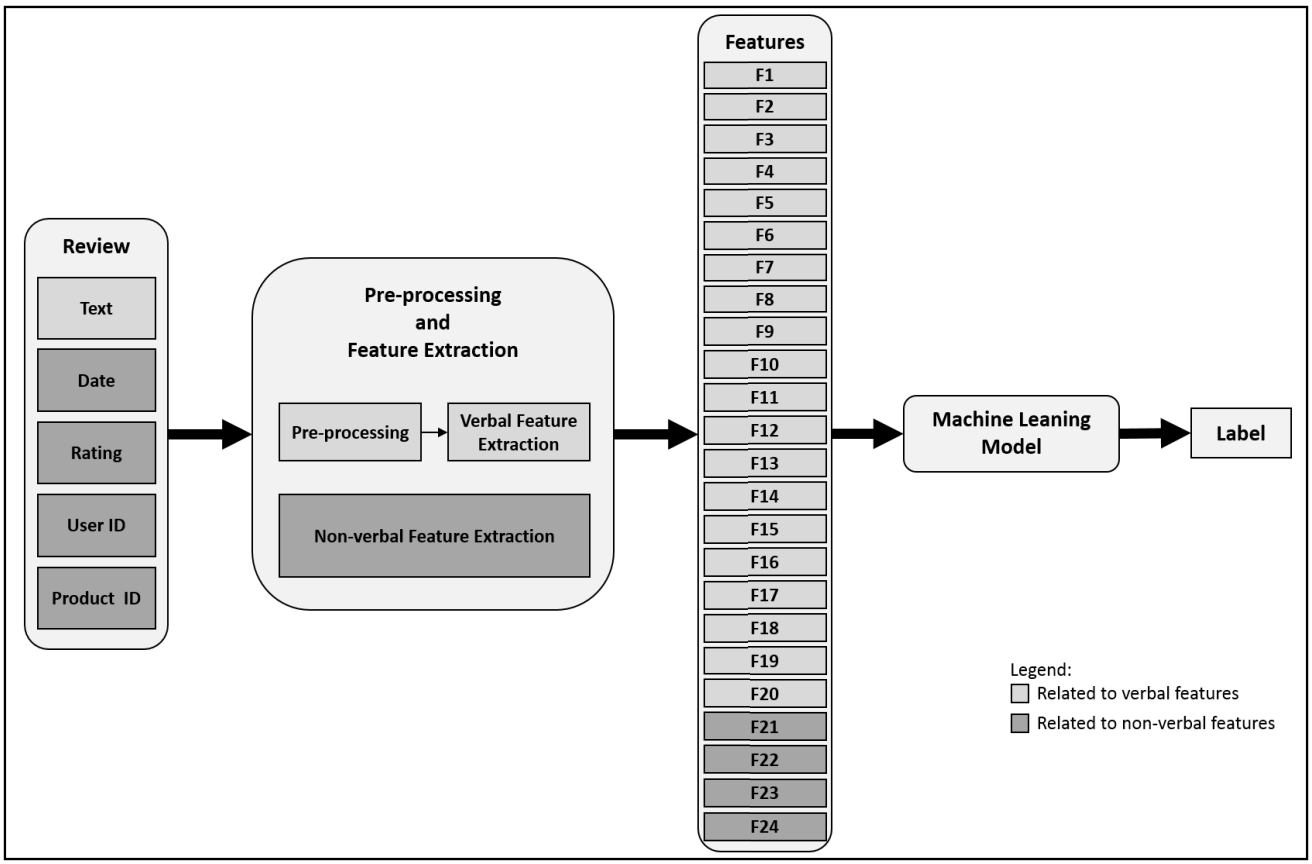

FIGURE 4. Fake reviews detection steps.

forest (RF), were chosen for this research. They were our first choice because they are frequently used in the literature and their performance results are promising. We used the Scikit-learn library [202] to implement these methods.

1) LOGISTIC REGRESSION (LR)

Logistic regression (LR) is a linear model used for classification. It utilizes an additional logistic function (sigmoid) and converts linear probabilities into logistic ones, in contrast to linear regression, which assumes linear correlations between the output and features. The estimated probabilities that fall between zero and one are constrained by the logit distribution.

2) NAIVE BAYES (NB)

Naïve Bayes (NB) classifiers [203] use Bayes' theorem and are probabilistic classifiers. NB is one of the earliest classification methods and is simple to build without the use of iterative parameter estimation schemes. As a result, they are extremely scalable and easily trainable, especially when the input dimensions are high, the NB classification method is well suited.

3) DECISION TREE (DT)

A decision tree (DT) models the process of converting input features into one of the defined class labels using a treelike graph. Each leaf node in the decision tree indicates a class

label for a given item. It is simple to interpret because it can be seen as a set of if-then rules. In this study, we used the classification and regression tree (CART) algorithm [204]. By continuously separating a node into two child nodes, starting with the root node that includes the entire learning sample, CART creates a binary decision tree.

4) RANDOM FOREST (RF)

Random forest (RF) [205] is an ensemble learning method for classification that creates several decision trees during training and outputs the class, representing the average of the classes produced by the individual trees.

D. MODEL EVALUATION

To validate the fake reviews detection model in our experiment, we performed 10-fold cross-validation with five evaluation metrics and the average values of the scores were reported for each of the four classification methods on the three datasets. We used the Scikit-learn library [202] to measure these metrics.

The area under the receiver operating characteristic curve (ROC AUC), F1-score (F1), accuracy (A), recall (R), and precision (P) were used as evaluation metrics with the YelpChi, YelpNYC, and YelpZip datasets using all twenty-four selected features (see **TABLE 6**).

The receiver operating characteristic (ROC) curve is the true positive rate vs. false positive rate curve, where the model is evaluated using different thresholds and the area under this curve (AUC) is used as an evaluation metric. The ROC AUC shows the ability of the model to discriminate between classes [206] regardless of the threshold or class distribution. A valuable feature of ROC curves is their insensitivity to class distribution because they are based on the true positive rate and false positive rate [207]. The three Yelp datasets are highly imbalanced, and fake reviews account for only 10-13% (see **TABLE 7**). Therefore, we used the ROC AUC for model evaluation and comparison with other models.

The F1-score (F1), accuracy (A), recall (R), and precision (P) are commonly used to evaluate machine-learning classifiers. Precision (P) is the ratio of fake reviews that are correctly classified as fake reviews. This metric evaluates the ability of the classifier not to classify a truthful review as fake. Recall (R) is the ratio of the total fake reviews in the dataset that are correctly classified, which evaluates the ability of the classifier to find all fake reviews in the dataset. F1-score (F1) is the harmonic mean of the recall and precision. Accuracy (A) is the ratio of correctly classified reviews to whether they are truthful or fake from all tested reviews. These four metrics are calculated as follows.

$$P = \frac{t_p}{t_p + f_p}$$

$$R = \frac{t_p}{t_p + f_n}$$

$$F1 = \frac{2P * R}{P + R}$$

TABLE 8. Experimental Results of Performance on Yelp Datasets (without hyperparameter optimization).

Dataset	ML model	AUC	F1	A	P	R
YelpChi	LR	0.7904	0.8078	0.8675	0.8151	0.8675
	NB	0.7560	0.7913	0.7672	0.8275	0.7672
	DT	0.5812	0.8022	0.7984	0.8063	0.7984
	RF	0.7322	0.8176	0.8608	0.8065	0.8608
YelpNYC	LR	0.7543	0.8487	0.8973	0.8342	0.8973
	NB	0.7104	0.8169	0.7945	0.8460	0.7945
	DT	0.5775	0.8383	0.8342	0.8428	0.8342
	RF	0.7231	0.8551	0.8940	0.8462	0.8940
YelpZip	LR	0.7909	0.8073	0.8678	0.8195	0.8678
	NB	0.7330	0.7830	0.7574	0.8209	0.7574
	DT	0.5961	0.8094	0.8060	0.8131	0.8060
	RF	0.7574	0.8252	0.8637	0.8189	0.8637

$$A = \frac{t_p + t_n}{t_p + t_n + f_p + f_n}$$

t_p is the true positive which indicates the number of fake reviews that were correctly classified as fake reviews. f_p is the false positive which indicates the number of truthful reviews that were falsely classified as fake ones. f_n is the false negative which indicates the number of fake reviews that were falsely classified as truthful reviews. t_n is the true negative which indicates the number of truthful reviews that were correctly classified as truthful reviews.

According to the high imbalance of labels in the datasets, we calculated the average weighted by support (the number of true instances for each label) for F1-score, precision, and recall.

E. ANALYSES OF RESULTS

To evaluate the impact of incorporating deception-based constructs on the prediction performance of the fake reviews detection model, we evaluated the selected features (see **TABLE 6**) that characterize these constructs in the reviews data.

First, we extracted the features from the three Yelp datasets. Second, we performed 10-fold cross-validation for each of the four classification methods on the three datasets without hyperparameter optimization. Third, we evaluated the impact and contribution of the selected features using the permutation feature importance technique. Fourth, we optimized the hyperparameters for the AUC scores of all classifiers using the exhaustive grid search with cross-validation. Fifth, we performed 10-fold cross-validation for each of the four classification methods on the three datasets using four feature sets.

The experimental results of the performance of the classification models using all selected features with 10-fold cross-validation are listed in **TABLE 8**. Logistic regression (LR) showed an AUC in the range of 0.75 to 0.79, F1-score in the range of 0.80 to 0.85, accuracy in the range of 0.86 to 0.89,

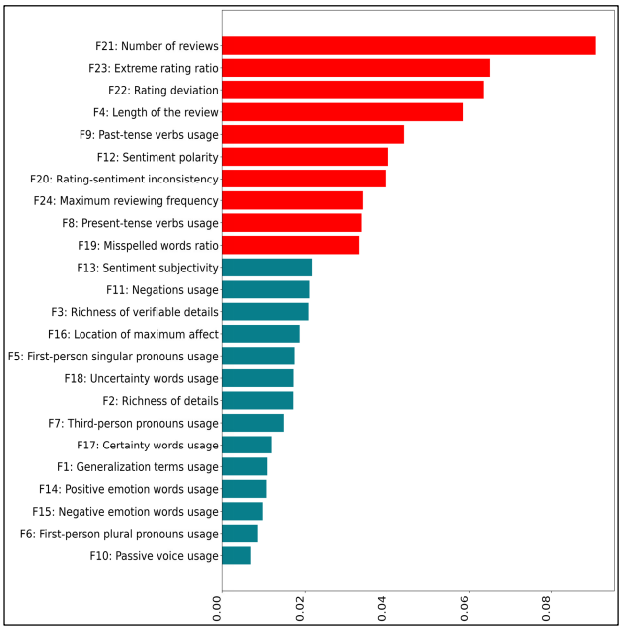

FIGURE 5. Fake reviews feature importance ranking.

precision in the range of 0.81 to 0.83, and recall in the range of 0.86 to 0.89.

Naïve Bayes (NB) showed an AUC in the range of 0.71 to 0.75, F1-score in the range of 0.78 to 0.81, accuracy in the range of 0.75 to 0.79, precision in the range of 0.82 to 0.84, and recall in the range of 0.75 to 0.79. The decision tree (DT) showed an AUC in the range of 0.57 to 0.59, F1-score in the range of 0.80 to 0.83, accuracy in the range of 0.79 to 0.83, precision in the range of 0.80 to 0.84, and recall in the range of 0.79 to 0.83. Random forest (RF) showed an AUC in the range of 0.72 to 0.75, F1-score in the range of 0.81 to 0.85, accuracy in the range of 0.86 to 0.89, precision in the range of 0.80 to 0.84, and recall in the range of 0.86 to 0.89.

We selected the permutation feature importance technique [208] rather than the biased impurity-based importance to evaluate the contribution of each feature and its related construct in the model prediction performance. The permutation feature importance is calculated by shuffling the features randomly one by one and then checking the decrement in the evaluation score. The experimental importance scores of the permutation feature importance for all selected features are shown in **TABLE 9** and **FIGURE 5**.

The top ten features that had sufficient importance scores (greater than 0.03): number of reviews, extreme rating ratio, rating deviation, length of the review, past-tense verbs usage, sentiment polarity, rating-sentiment inconsistency, maximum reviewing frequency, present-tense verbs usage, and misspelled words ratio, respectively. These features are related to the constructs of source credibility, deviation in behavior, quantity, non-immediacy, affect, consistency, and informality.

We optimized the hyperparameters of all classifiers using the exhaustive grid search with 10-fold cross-validation to improve the AUC score. We then performed 10-fold cross-validation using four feature sets: all selected features, important features (see the top ten in **FIGURE 5**), verbal features (see F1 to F20 in **TABLE 6**), and non-verbal features (see F21 to F24 in **TABLE 6**).

The experimental results of the performance of the ML classification methods using each feature set with 10-fold

TABLE 9. Experimental results of importance scores.

No.	Feature No.	Feature Description	Construct	Construct Type	Importance Score
1	F1	Generalization terms usage	Specificity	Verbal	0.011
2	F2	Richness of details	Specificity	Verbal	0.017
3	F3	Richness of verifiable details	Specificity	Verbal	0.021
4	F4	Length of the review	Quantity	Verbal	0.058
5	F5	First-person singular pronouns usage	Non-immediacy	Verbal	0.017
6	F6	First-person plural pronouns usage	Non-immediacy	Verbal	0.009
7	F7	Third-person pronouns usage	Non-immediacy	Verbal	0.015
8	F8	Present-tense verbs usage	Non-immediacy	Verbal	0.034
9	F9	Past-tense verbs usage	Non-immediacy	Verbal	0.044
10	F10	Passive voice usage	Non-immediacy	Verbal	0.007
11	F11	Negations usage	Non-immediacy	Verbal	0.021
12	F12	Sentiment polarity	Affect	Verbal	0.040
13	F13	Sentiment subjectivity	Affect	Verbal	0.022
14	F14	Positive emotion words usage	Affect	Verbal	0.011
15	F15	Negative emotion words usage	Affect	Verbal	0.010
16	F16	Location of maximum affect	Affect	Verbal	0.019
17	F17	Certainty words usage	Uncertainty	Verbal	0.012
18	F18	Uncertainty words usage	Uncertainty	Verbal	0.017
19	F19	Misspelled words ratio	Informality	Verbal	0.033
20	F20	Rating-sentiment inconsistency	Consistency	Verbal	0.040
21	F21	Number of reviews	Source credibility	Non-Verbal	0.090
22	F22	Rating deviation	Deviation in behavior	Non-Verbal	0.063
23	F23	Extreme rating ratio	Deviation in behavior	Non-Verbal	0.065
24	F24	Maximum reviewing frequency	Deviation in behavior	Non-Verbal	0.034

TABLE 10. Experimental results of AUC scores using different feature sets (with hyperparameter optimization).

Dataset	ML model	AUC			
		All Features	Important Features	Verbal Features	Non-Verbal Features
YelpChi	LR	0.7905	0.7887	0.6534	0.7725
	NB	0.7560	0.7605	0.6565	0.7435
	DT	0.7907	0.7915	0.6699	0.7798
	RF	0.8010	**0.8038**	0.6889	0.7922
YelpNYC	LR	0.7543	0.7529	0.6533	0.7393
	NB	0.7104	0.7192	0.6483	0.7003
	DT	0.7921	0.7922	0.6727	0.7786
	RF	0.8048	**0.8095**	0.6882	0.8005
YelpZip	LR	0.7909	0.7898	0.6476	0.7846
	NB	0.7330	0.7436	0.6442	0.7364
	DT	0.8121	0.8121	0.6769	0.8051
	RF	0.8210	**0.8262**	0.6932	0.8217

cross-validation are presented in **TABLE 10** and **FIGURE 6**. The performance results are similar among different classification methods on different datasets, proving that the model is consistent and its strength depends on the selected features, regardless of the dataset or the method used for the purpose of classification. By testing the performance of the proposed model, the fifth research objective was achieved (RO5).

V. DISCUSSION

This study examined how deception constructs from deception theories can be used to build a fake reviews detection model and improve its performance. Incorporating verbal and non-verbal deception-related features was theory-driven. The main purpose was to improve the performance of the fake reviews detection model and to avoid the incorporation of irrelevant features. We evaluated how deception-related features differentiate fake and truthful reviews.

If we return to deception theories, we can see that reality monitoring theory (RM), criteria-based content analysis (CBCA), scientific content analysis (SCAN), verifiability approach (VA), truth-default theory (TDT), and information manipulation theory (IMT) focus only on the verbal behavior and the content of the deceptive message. Leakage theory, four-factor theory, interpersonal deception theory (IDT), and self-presentational theory consider non-verbal behavior. The only theory that had computationally measurable non-verbal constructs in the context of the computer-mediated text is the interpersonal deception theory (IDT). Therefore, the bias found in deception theories towards verbal deception was clearly reflected in the selection of constructs and, thus, the selection of features.

To test the nine hypotheses, we empirically evaluated the permutation importance for each feature using the three Yelp datasets. If a construct has at least one related feature with a sufficient importance score (greater than 0.03), then

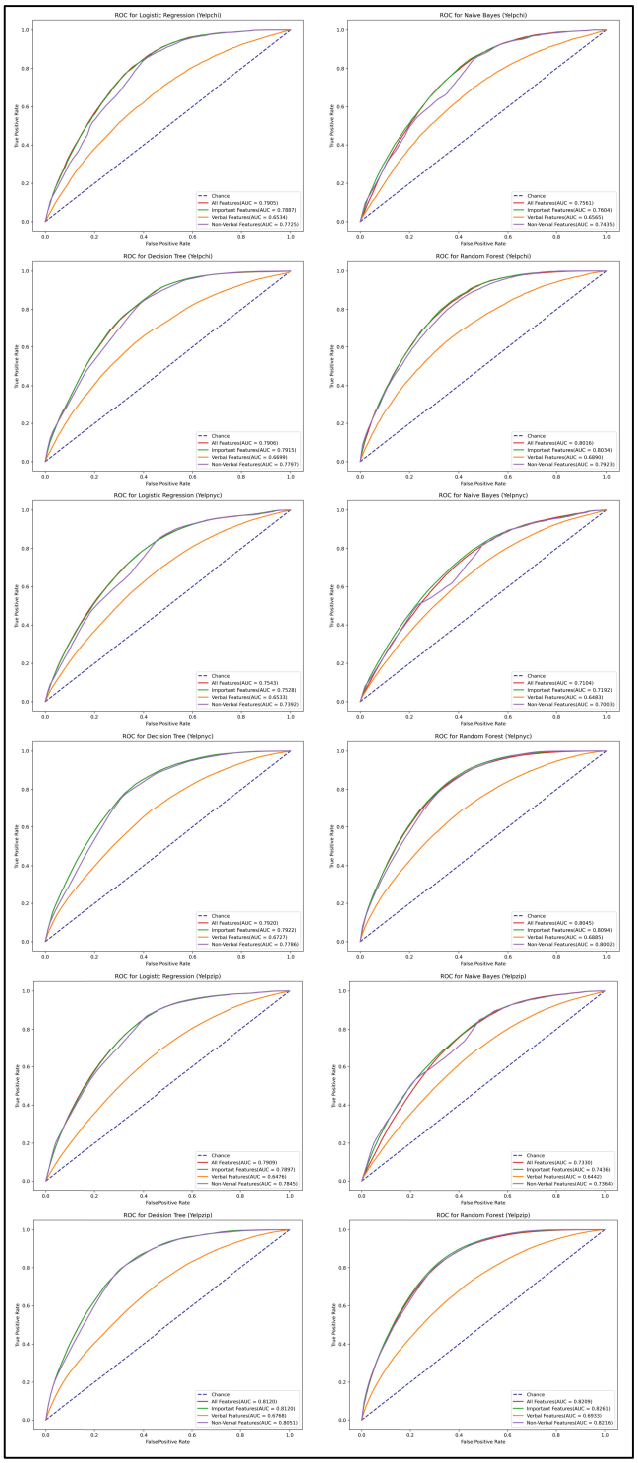

FIGURE 6. ROC curves using all classification methods with different feature sets.

incorporating this construct improves the prediction performance of the fake reviews detection model, thereby supporting the hypothesis that considers this construct.

Although the number of selected non-verbal features is five times less than the number of selected verbal features, and the computational cost of extraction for non-verbal features

is less than that for verbal features, non-verbal features are still more important than verbal features in terms of their ability to differentiate between fake and truthful reviews (see **TABLE 9** and **FIGURE 5**). The top three features – number of reviews, extreme rating ratio, and rating deviation– are related to non-verbal constructs, source credibility, and deviation in behavior, respectively. The last non-verbal feature is the maximum reviewing frequency, which also has a sufficient importance score (see **TABLE 9** and **FIGURE 5**). Therefore, these results support Hypotheses H8 and H9.

The results (see **TABLE 10** and **FIGURE 6**) generally show a performance improvement using combined verbal and non-verbal important features over using all features. The highest performance for all datasets was achieved using the RF model with important features. A performance decrement is shown between using non-verbal features and the use of all or important features. These results support the use of a combination of verbal and non-verbal features to improve the performance of the fake reviews detection model. On the other hand, using non-verbal features alone still shows good performance, which is another strong piece of evidence that supports Hypotheses H8 and H9.

The length of the review, past-tense verbs usage, sentiment polarity, rating-sentiment inconsistency, present-tense verbs usage, and misspelled words ratio are the six verbal features with sufficient importance scores that represent the importance and impact of the verbal constructs: quantity, non-immediacy, affect, consistency, and informality. Therefore, these results support Hypotheses H2, H3, H4, H6, and H7.

The remaining features are related to specificity and uncertainty. Although these constructs have received more attention in deception theories than non-verbal constructs, they were not able to show enough importance as features in the fake reviews detection model. Therefore, these results do not support Hypotheses H1 and H5. Based on these results, fake reviews are sufficiently certain and can be rich in details whether they are verifiable or unverifiable, and cannot be distinguished from truthful reviews by specificity or certainty.

Using verbal features (**TABLE 10** and **FIGURE 6**) shows the worst results, with a decrement between 6% in the best case and 14% in the worst case, which proves the importance of non-verbal features over verbal features. In general, the results show the difficulty for fake reviewers to imitate the non-verbal behavior of truthful reviewers, while verbal behavior is easier to imitate because the review content can be manipulated and prepared to look truthful.

According to the results, the supported deception constructs that can improve the fake reviews detection are quantity, non-immediacy, affect, consistency, informality, source credibility, and deviation in behavior. These results can answer the first research question (RQ1) ''what deception aspects from deception theories should be considered to capture the behavior of fraudulent online customers?''

According to the results, the crucial features that can improve fake reviews detection are the number of reviews, extreme rating ratio, rating deviation, length of the review, past-tense verbs usage, sentiment polarity, rating-sentiment inconsistency, maximum reviewing frequency, present-tense verbs usage, and misspelled words ratio. These results can answer the second research question (RQ2) ''What are the possible features that can be extracted from the available attributes in open-source customer reviews to reflect the relevant aspects of deceptive behavior?''

To extract the selected features from the online reviews data, we used programming tools and methods, summarized in **TABLE 6** and **FIGURE 3**. This answers the third research question (RQ3) ''What techniques can be used to extract the features related to deception aspects from the available attributes in user data?''

T. Vantan et al. [209] combined verbal features with a bag of words. They proposed four classification methods with their combined verbal features. Although they chose CNN+LSTM as one of their proposed methods, they could not achieve an accuracy of more than 0.78 using only a balanced subset of the YelpNYC dataset.

M. Ferreira Uchoa [210] used n-grams (verbal features) with SVM and NN. They used the YelpChi, YelpNYC, and YelpZIP datasets for the testing. Although their proposed neural networks reached 1500 hidden layers, they could only reach an accuracy in a range of 0.54 to 0.65.

A. Rastogi et al. [201], [211] used behavioral and textual features with LR, SVM, and MLP. They filtered YelpNYC and YelpZip to consider products and reviewers that had at least three reviews, which means that they filtered out approximately 75% from YelpNYC and 42% from YelpZip. The filtered-out data contain 80% of fake reviews in each dataset. Though they filtered out most of the fake reviews from both datasets before validation, they could only reach an AUC in a range of 0.73 to 0.88 for YelpNYC and 0.70 to 0.84 for YelpZIP using behavioral features.

C. Yuan et al. [212] tested a group of well-known fake reviews detection models as a baseline to compare the results with those of their proposed model (HFAN). They used the YelpChi, YelpNYC, and YelpZIP datasets for benchmarking. The models tested were RSD [213], SpEagle [163], TDSD [214], CHMM [215], Spam2Vec [216], CNN-GRNN [76], SWNN [217], ABNN [218], and AEDA [78] (see their results in **TABLE 11**). They considered the text at the user and product levels. Although their proposed model (HFAN) has a highly complex architecture, they could only reach AUC scores of 0.83, 0.85, and 0.87 for YelpChi, YelpNYC, and YelpZIP, respectively.

Our unified model outperformed most of the well-known fake review detection models (see **TABLE 11**). Although A. Rastogi et al.[201], [211] obtained high AUC scores, their model was validated after filtering out 80% of the fake reviews from YelpNYC and YelpZip datasets, making the results unreliable for comparison. For HFAN [212], their model obtained high AUC scores (only 3-5% higher than our model), but it suffers from high complexity and low interpretability of the results. Our model is interpretable, with favorable theory-based features and mapping to deception

TABLE 11. Comparing AUC scores of prominent fake reviews detection models using the Yelp datasets.

Model	Best AUC Score			Data used from the dataset (%)	Model Complexity
	YelpChi	YelpNYC	YelpZip		
T. Vantan *et al.* [209]	N/A	0.7840	N/A	22%	Medium
A. Rastogi *et al.* [201], [211]	N/A	0.8800	0.8400	25-58%	Low
RSD [213]	0.5062	0.5415	0.5982	100%	Low
SpEagle [163]	0.7887	0.7829	0.8040	100%	Medium
TDSD [214]	0.7882	0.7886	0.8163	100%	High
CHMM [215]	0.7868	0.7871	0.8264	100%	High
Spam2Vec [216]	0.7861	0.7835	0.8121	100%	Low
CNN-GRNN [76]	0.7868	0.7904	0.8187	100%	High
SWNN [217]	0.7857	0.7857	0.8125	100%	High
ABNN [218]	0.7853	0.7883	0.8082	100%	High
AEDA [78]	0.7914	0.7892	0.8132	100%	High
HFAN [212]	**0.8324**	**0.8478**	**0.8728**	**100%**	**High**
Our model	**0.8038**	**0.8095**	**0.8262**	**100%**	**Low**

theories. In addition, our model has low complexity and high performance.

After implementing and validating the proposed fake reviews detection model and then comparing it with state-of-the-art models, the improvement in performance, complexity, and interpretability was proved by results which, in turn, answered the fourth research question (RQ4) "Can the deception-based fake reviews detection model improve the performance of fake reviews detection?".

VI. CONCLUSION AND FUTURE DIRECTIONS

In this study, we developed a pure theory-based model for fake reviews detection. To achieve the objectives of the research and answer the research questions, we began by synthesizing ten popular deception theories and analyzing them thoroughly. Next, we derived important constructs of deception from the deception theories to build a unified theoretical model. We selected features that could characterize the derived constructs and could be measured from the review texts and the reviewer's behavior. Finally, our fake reviews detection model was empirically validated using three famous Yelp reviews datasets after extracting the selected features.

Some limitations of this work present ample opportunities for future research as suggested below.

- First, the synthesized deception theories in this study are limited by those that are the most influential and popular in computer-mediated text contexts. More deception theories can be synthesized and merged with other well-founded fundamental theories from sociology, criminology, biology, or linguistics.
- Second, our selected constructs from deception theories were limited to computer-mediated text. Therefore, we encourage other researchers to select constructs from the same deception theories that are applicable to deception detection for other types of media such as voice and video.
- Third, more features can be added to our feature set to characterize the derived constructs more widely in the context of online reviews.
- Fourth, more complex semantic features using deep learning methods (e.g., [219], [220], [221]) and other empirically-derived features can be combined with theory-based features to enhance the prediction performance.
- Fifth, the benchmarking Yelp datasets used for model validation include reviews of hotels and restaurants only, which raises questions regarding the generalizability of our model and the study's conclusions. Therefore, it would be worthwhile to investigate whether the results of this study can be reproduced for online reviews of other business categories.
- Sixth, we used only four classical machine-learning algorithms for model validation. Therefore, we encourage other researchers to validate the model using different algorithms, particularly neural networks, which are expected to achieve better predictions.

REFERENCES

[1] R. Zhang, C. Sha, M. Zhou, and A. Zhou, "Exploiting shopping and reviewing behavior to re-score online evaluations," in *Proc. 21st Int. Conf. Companion World Wide Web (WWW Companion)*, 2012, pp. 649–650, doi: 10.1145/2187980.2188171.

[2] M. Anderson and A. Smith. (2016). *Online Shopping and E-Commerce*. Pew Research Center. [Online]. Available: https://www.pewresearch.org/internet/2016/12/19/online-shopping-and-e-commerce/

[3] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "What yelp fake review filter might be doing?" in *Proc. 7th Int. Conf. Weblogs Social Media (ICWSM)*, 2013, pp. 409–418.

[4] S. Banerjee and A. Y. K. Chua, "Theorizing the textual differences between authentic and fictitious reviews: Validation across positive, negative and moderate polarities," *Internet Res.*, vol. 27, no. 2, pp. 321–337, Apr. 2017, doi: 10.1108/IntR-11-2015-0309.

[5] Y. Wu, E. W. T. Ngai, P. Wu, and C. Wu, "Fake online reviews: Literature review, synthesis, and directions for future research," *Decis. Support Syst.*, vol. 132, May 2020, Art. no. 113280, doi: 10.1016/j.dss.2020.113280.

[6] J. Salminen, C. Kandpal, A. M. Kamel, S.-G. Jung, and B. J. Jansen, "Creating and detecting fake reviews of online products," *J. Retailing Consum. Services*, vol. 64, Jan. 2022, Art. no. 102771, doi: 10.1016/j.jretconser.2021.102771.

[7] N. Hussain, H. T. Mirza, G. Rasool, I. Hussain, and M. Kaleem, "Spam review detection techniques: A systematic literature review," *Appl. Sci.*, vol. 9, no. 5, p. 987, Mar. 2019, doi: 10.3390/app9050987.

[8] S. Kaddoura, G. Chandrasekaran, D. Elena Popescu, and J. H. Duraisamy, "A systematic literature review on spam content detection and classification," *PeerJ Comput. Sci.*, vol. 8, p. e830, Jan. 2022, doi: 10.7717/PEERJ-CS.830.

[9] R. Mohawesh, S. Xu, S. N. Tran, R. Ollington, M. Springer, Y. Jararweh, and S. Maqsood, "Fake reviews detection: A survey," *IEEE Access*, vol. 9, pp. 65771–65802, 2021, doi: 10.1109/ACCESS.2021.3075573.

[10] X. Wu, Y. Dong, J. Tao, C. Huang, and N. V. Chawla, "Reliable fake review detection via modeling temporal and behavioral patterns," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 494–499, doi: 10.1109/BigData.2017.8257963.

[11] B. Hooi, H. A. Song, A. Beutel, N. Shah, K. Shin, and C. Faloutsos, "FRAUDAR: Bounding graph fraud in the face of camouflage," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 895–904, doi: 10.1145/2939672.2939747.

[12] A. Heydari, M. A. Tavakoli, N. Salim, and Z. Heydari, "Detection of review spam: A survey," *Expert Syst. Appl.*, vol. 42, no. 7, pp. 3634–3642, 2015, doi: 10.1016/j.eswa.2014.12.029.

[13] R. Filieri, S. Alguezaui, and F. McLeay, "Why do travelers trust TripAdvisor? Antecedents of trust towards consumer-generated media and its influence on recommendation adoption and word of mouth," *Tour. Manag.*, vol. 51, pp. 174–185, Dec. 2015, doi: 10.1016/j.tourman.2015.05.007.

[14] C. Dellarocas, "Strategic manipulation of internet opinion forums: Implications for consumers and firms," *Manage. Sci.*, vol. 52, no. 10, pp. 1577–1593, Oct. 2006, doi: 10.1287/mnsc.1060.0567.

[15] S. Gössling, C. M. Hall, and A.-C. Andersson, "The manager's dilemma: A conceptualization of online review manipulation strategies," *Current Issues Tourism*, vol. 21, no. 5, pp. 484–503, Mar. 2018, doi: 10.1080/13683500.2015.1127337.

[16] M. Anderson and J. Magruder, "Learning from the crowd: Regression discontinuity estimates of the effects of an online review database," *Econ. J.*, vol. 122, no. 563, pp. 957–989, 2012, doi: 10.1111/j.1468-0297.2012.02512.x.

[17] M. Luca, "Reviews, reputation, and revenue: The case of yelp.Com," *SSRN Electron. J.*, pp. 1–40, Sep. 2011, doi: 10.2139/ssrn.1928601.

[18] S. Salehi-Esfahani and A. B. Ozturk, "Negative reviews: Formation, spread, and halt of opportunistic behavior," *Int. J. Hospitality Manage.*, vol. 74, pp. 138–146, Aug. 2018, doi: 10.1016/j.ijhm.2018.06.022.

[19] N. Hu, I. Bose, N. S. Koh, and L. Liu, "Manipulation of online reviews: An analysis of ratings, readability, and sentiments," *Decis. Support Syst.*, vol. 52, no. 3, pp. 674–684, Feb. 2012, doi: 10.1016/j.dss.2011.11.002.

[20] *Amazon Flooded With Millions of Fake Reviews in 2019*. Accessed: Aug. 23, 2022. [Online]. Available: https://reviewmeta.com/blog/amazon-flooded-with-millions-of-fake-reviews-in-2019/

[21] J. Wang and C. Wu, "Camouflage is NOT easy: Uncovering adversarial fraudsters in large online app review platform," *Meas. Control*, vol. 53, nos. 9–10, pp. 2137–2145, Nov. 2020, doi: 10.1177/00202940209702l3.

[22] P. Kaghazgaran, J. Caverlee, and A. Squicciarini, "Combating crowdsourced review manipulators: A neighborhood-based approach," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, Feb. 2018, pp. 306–314, doi: 10.1145/3159652.3159726.

[23] Y. Yao, B. Viswanath, J. Cryan, H. Zheng, and B. Y. Zhao, "Automated crowdturfing attacks and defenses in online review systems," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2017, pp. 1143–1158, doi: 10.1145/3133956.3133990.

[24] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2011, pp. 309–319.

[25] J. Yao, Y. Zheng, and H. Jiang, "An ensemble model for fake online review detection based on data resampling, feature pruning, and parameter optimization," *IEEE Access*, vol. 9, pp. 16914–16927, 2021, doi: 10.1109/ACCESS.2021.3051174.

[26] R. Mohawesh, S. Tran, R. Ollington, and S. Xu, "Analysis of concept drift in fake reviews detection," *Expert Syst. Appl.*, vol. 169, May 2021, Art. no. 114318, doi: 10.1016/j.eswa.2020.114318.

[27] F. Khurshid, Y. Zhu, Z. Xu, M. Ahmad, and M. Ahmad, "Enactment of ensemble learning for review spam detection on selected features," *Int. J. Comput. Intell. Syst.*, vol. 12, no. 1, p. 387, 2019, doi: 10.2991/ijcis.2019.125905655.

[28] E. F. Cardoso, R. M. Silva, and T. A. Almeida, "Towards automatic filtering of fake reviews," *Neurocomputing*, vol. 309, pp. 106–116, Oct. 2018, doi: 10.1016/j.neucom.2018.04.074.

[29] G. Shan, L. Zhou, and D. Zhang, "From conflicts and confusion to doubts: Examining review inconsistency for fake review detection," *Decis. Support Syst.*, vol. 144, May 2021, Art. no. 113513, doi: 10.1016/j.dss.2021.113513.

[30] S. Banerjee and A. Y. K. Chua, "A theoretical framework to identify authentic online reviews," *Online Inf. Rev.*, vol. 38, no. 5, pp. 634–649, Jul. 2014, doi: 10.1108/OIR-02-2014-0047.

[31] N. Jindal and B. Liu, "Analyzing and detecting review spam," in *Proc. 7th IEEE Int. Conf. Data Mining (ICDM)*, Oct. 2007, pp. 547–552, doi: 10.1109/ICDM.2007.68.

[32] N. Jindal and B. Liu, "Review spam detection," in *Proc. 16th Int. Conf. World Wide Web (WWW)*, 2007, pp. 1189–1190, doi: 10.1145/1242572.1242759.

[33] J. Li, M. Ott, C. Cardie, and E. Hovy, "Towards a general rule for identifying deceptive opinion spam," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2014, pp. 1566–1576, doi: 10.3115/v1/p14-1147.

[34] A. Jeyapriya and C. S. K. Selvi, "Extracting aspects and mining opinions in product reviews using supervised learning algorithm," in *Proc. 2nd Int. Conf. Electron. Commun. Syst. (ICECS)*, Feb. 2015, pp. 548–552, doi: 10.1109/ECS.2015.7124967.

[35] A. A. Hammad and A. El-Halees, "An approach for detecting spam in Arabic opinion reviews," *Int. Arab J. Inf. Technol.*, vol. 12, no. 1, pp. 10–16, 2015.

[36] L. Zhang, Z. Wu, and J. Cao, "Detecting spammer groups from product reviews: A partially supervised learning model," *IEEE Access*, vol. 6, pp. 2559–2568, 2018, doi: 10.1109/ACCESS.2017.2784370.

[37] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting burstiness in reviews for review spammer detection," in *Proc. 7th Int. Conf. Weblogs Social Media (ICWSM)*, 2013, pp. 175–184.

[38] B. Heredia, T. M. Khoshgoftaar, J. D. Prusa, and M. Crawford, "Improving detection of untrustworthy online reviews using ensemble learners combined with feature selection," *Social Netw. Anal. Mining*, vol. 7, no. 1, pp. 1–18, Dec. 2017, doi: 10.1007/s13278-017-0456-z.

[39] J. K. Rout, A. K. Dash, and N. K. Ray, "A framework for fake review detection: Issues and challenges," in *Proc. Int. Conf. Inf. Technol. (ICIT)*, Dec. 2018, pp. 7–10, doi: 10.1109/ICIT.2018.00014.

[40] L. Ball and J. Elworthy, "Fake or real? The computational detection of online deceptive text," *J. Marketing Analytics*, vol. 2, no. 3, pp. 187–201, Sep. 2014, doi: 10.1057/jma.2014.15.

[41] Z. Sedighi, H. Ebrahimpour-Komleh, and A. Bagheri, "RLOSD: Representation learning based opinion spam detection," in *Proc. 3rd Iranian Conf. Intell. Syst. Signal Process. (ICSPIS)*, Dec. 2017, pp. 74–80, doi: 10.1109/ICSPIS.2017.8311593.

[42] D. Zhang, L. Zhou, J. L. Kehoe, and I. Y. Kilic, "What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews," *J. Manage. Inf. Syst.*, vol. 33, no. 2, pp. 456–481, Apr. 2016, doi: 10.1080/07421222.2016.1205907.

[43] N. Kumar, D. Venugopal, L. Qiu, and S. Kumar, "Detecting review manipulation on online platforms with hierarchical supervised learning," *J. Manage. Inf. Syst.*, vol. 35, no. 1, pp. 350–380, Jan. 2018, doi: 10.1080/07421222.2018.1440758.

[44] R. Barbado, O. Araque, and C. A. Iglesias, "A framework for fake review detection in online consumer electronics retailers," *Inf. Process. Manage.*, vol. 56, no. 4, pp. 1234–1244, Jul. 2019, doi: 10.1016/j.ipm.2019.03.002.

[45] F. Khurshid, Y. Zhu, C. W. Yohannese, and M. Iqbal, "Recital of supervised learning on review spam detection: An empirical analysis," in *Proc. 12th Int. Conf. Intell. Syst. Knowl. Eng. (ISKE)*, Nov. 2017, pp. 1–6, doi: 10.1109/ISKE.2017.8258755.

[46] Y. Lin, T. Zhu, X. Wang, J. Zhang, and A. Zhou, "Towards online review spam detection," in *Proc. 23rd Int. Conf. World Wide Web*, Apr. 2014, pp. 341–342, doi: 10.1145/2567948.2577293.

[47] S. Shojaee, M. A. A. Murad, A. B. Azman, N. M. Sharef, and S. Nadali, "Detecting deceptive reviews using lexical and syntactic features," in *Proc. 13th Int. Conf. Intellient Syst. Design Appl.*, Dec. 2013, pp. 53–58, doi: 10.1109/ISDA.2013.6920707.

[48] M. Ott, C. Cardie, and J. T. Hancock, "Negative deceptive opinion spam," in *Proc. 2nd Workshop Comput. Linguistics Literature, CLfL Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, 2013, pp. 497–501.

[49] H. Li, Z. Chen, A. Mukherjee, B. Liu, and J. Shao, "Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns," in *Proc. 9th Int. Conf. Web Social Media (ICWSM)*, 2015, pp. 634–637.

[50] L. C. Cagnina and P. Rosso, "Detecting deceptive opinions: Intra and cross-domain classification using an efficient representation," *Int. J. Uncertainty, Fuzziness Knowl. Based Syst.*, vol. 25, no. 2, pp. 151–174, Oct. 2017, doi: 10.1142/S0218488517400165.

[51] S. P. Rajamohana and K. Umamaheswari, "Hybrid approach of improved binary particle swarm optimization and shuffled frog leaping for feature selection," *Comput. Electr. Eng.*, vol. 67, pp. 497–508, Apr. 2018, doi: 10.1016/j.compeleceng.2018.02.015.

[52] H. Li, B. Liu, A. Mukherjee, and J. Shao, "Spotting fake reviews using positive-unlabeled learning," *Computación y Sistemas*, vol. 18, no. 3, pp. 467–475, Sep. 2014, doi: 10.13053/CyS-18-3-2035.

[53] Y. Liu, B. Pang, and X. Wang, "Opinion spam detection by incorporating multimodal embedded representation into a probabilistic review graph," *Neurocomputing*, vol. 366, pp. 276–283, Nov. 2019, doi: 10.1016/j.neucom.2019.08.013.

[54] Q. T. Ha, T. T. Vu, H. T. Pham, and C. T. Luu, "An upgrading feature-based opinion mining model on Vietnamese product reviews," in *Active Media Technology* (Lecture Notes in Computer Science), vol. 6890. Berlin, Germany: Springer, 2011, pp. 173–185, doi: 10.1007/978-3-642-23620-4_21.

[55] L. Liu, Z. Lv, and H. Wang, "Opinion mining based on feature-level," in *Proc. 5th Int. Congr. Image Signal Process. (CISP)*, Oct. 2012, pp. 1596–1600, doi: 10.1109/CISP.2012.6469929.

[56] M. D. Kamalesh and H. K. Diwedi, "Extracting product features from consumer reviews and its applications," *Int. J. Appl. Eng. Res.*, vol. 10, no. 2, pp. 2345–2350, 2015.

[57] W.-J. Jia, S. Zhang, Y.-J. Xia, J. Zhang, and H. Yu, "A novel product features categorize method based on twice-clustering," in *Proc. Int. Conf. Web Inf. Syst. Mining*, Oct. 2010, pp. 281–284, doi: 10.1109/WISM.2010.71.

[58] H. Ahmed, "Detecting opinion spam and fake news using n-gram analysis and semantic similarity," Ph.D. thesis, 2017. [Online]. Available: http://dspace.library.uvic.ca/bitstream/handle/1828/8796/Ahmed_Hadeer_Masc_2017.pdf

[59] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," in *Proc. 21st Int. Conf. World Wide Web*, Apr. 2012, pp. 191–200, doi: 10.1145/2187836.2187863.

[60] L.-Y. Dong, S.-J. Ji, C.-J. Zhang, Q. Zhang, D. W. Chiu, L.-Q. Qiu, and D. Li, "An unsupervised topic-sentiment joint probabilistic model for detecting deceptive reviews," *Expert Syst. Appl.*, vol. 114, pp. 210–223, Dec. 2018, doi: 10.1016/j.eswa.2018.07.005.

[61] C. Yu, Y. Zuo, B. Feng, L. An, and B. Chen, "An individual-group-merchant relation model for identifying fake online reviews: An empirical study on a Chinese e-commerce platform," *Inf. Technol. Manage.*, vol. 20, no. 3, pp. 123–138, Sep. 2019, doi: 10.1007/s10799-018-0288-1.

[62] S. Noekhah, N. B. Salim, and N. H. Zakaria, "Opinion spam detection: Using multi-iterative graph-based model," *Inf. Process. Manage.*, vol. 57, no. 1, Jan. 2020, Art. no. 102140, doi: 10.1016/j.ipm.2019.102140.

[63] S. Deng, "Deceptive reviews detection of industrial product," *Int. J. Services Oper. Informat.*, vol. 8, no. 2, pp. 122–135, 2016, doi: 10.1504/IJSOI.2016.080090.

[64] Y. Liu and B. Pang, "A unified framework for detecting author spamicity by modeling review deviation," *Expert Syst. Appl.*, vol. 112, pp. 148–155, Dec. 2018, doi: 10.1016/j.eswa.2018.06.028.

[65] D. H. Fusilier, M. Montes-y-Gómez, P. Rosso, and R. Guzmán Cabrera, "Detecting positive and negative deceptive opinions using PU-learning," *Inf. Process. Manage.*, vol. 51, no. 4, pp. 433–443, Jul. 2015, doi: 10.1016/j.ipm.2014.11.001.

[66] J. K. Rout, A. Dalmia, K.-K. R. Choo, S. Bakshi, and S. K. Jena, "Revisiting semi-supervised learning for online deceptive review detection," *IEEE Access*, vol. 5, pp. 1319–1327, 2017, doi: 10.1109/ACCESS.2017.2655032.

[67] H. Deng, L. Zhao, N. Luo, Y. Liu, G. Guo, X. Wang, Z. Tan, S. Wang, and F. Zhou, "Semi-supervised learning based fake review detection," in *Proc. IEEE Int. Symp. Parallel Distrib. Process. with Appl. IEEE Int. Conf. Ubiquitous Comput. Commun. (ISPA/IUCC)*, Dec. 2017, pp. 1278–1280, doi: 10.1109/ISPA/IUCC.2017.00195.

[68] Z. Wu, J. Cao, Y. Wang, Y. Wang, L. Zhang, and J. Wu, "HPSD: A hybrid PU-learning-based spammer detection model for product reviews," *IEEE Trans. Cybern.*, vol. 50, no. 4, pp. 1595–1606, Apr. 2020, doi: 10.1109/TCYB.2018.2877161.

[69] W. Zhang, C. Bu, T. Yoshida, and S. Zhang, "CoSpa: A co-training approach for spam review identification with support vector machine," *Information*, vol. 7, no. 1, p. 12, Mar. 2016, doi: 10.3390/info7010012.

[70] W. Zhang, C. Bu, T. Yoshida, and S. Zhang, "CoFea: A novel approach to spam review identification based on entropy and co-training," *Entropy*, vol. 18, no. 12, p. 429, Nov. 2016, doi: 10.3390/e18120429.

[71] Y. Li, Y. Lin, J. Zhang, J. Li, and L. Zhao, "Highlighting the fake reviews in review sequence with the suspicious contents and behaviours," *J. Inf. Comput. Sci.*, vol. 12, no. 4, pp. 1615–1627, Mar. 2015, doi: 10.12733/jics20105452.

[72] Z. Hai, P. Zhao, P. Cheng, P. Yang, X.-L. Li, and G. Li, "Deceptive review spam detection via exploiting task relatedness and unlabeled data," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 1817–1826, doi: 10.18653/v1/d16-1187.

[73] C. M. Yilmaz and A. O. Durahim, "SPR2EP: A semi-supervised spam review detection framework," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 306–313, doi: 10.1109/ASONAM.2018.8508314.

[74] Y. Tian, M. Mirzabagheri, P. Tirandazi, and S. M. H. Bamakan, "A non-convex semi-supervised approach to opinion spam detection by ramp-one class SVM," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102381, doi: 10.1016/j.ipm.2020.102381.

[75] S. Yuan, X. Wu, and Y. Xiang, "Task-specific word identification from short texts using a convolutional neural network," *Intell. Data Anal.*, vol. 22, no. 3, pp. 533–550, May 2018, doi: 10.3233/IDA-173413.

[76] Y. Ren and Y. Zhang, "Deceptive opinion spam detection using neural network," in *Proc. 26th Int. Conf. Comput. Linguistics (COLING)*, 2016, pp. 140–150.

[77] H. Aghakhani, A. Machiry, S. Nilizadeh, C. Kruegel, and G. Vigna, "Detecting deceptive reviews using generative adversarial networks," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, May 2018, pp. 89–95, doi: 10.1109/SPW.2018.00022.

[78] Z. You, T. Qian, and B. Liu, "An attribute enhanced domain adaptive model for cold-start spam review detection," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 1888–1895.

[79] X. Tang, T. Qian, and Z. You, "Generating behavior features for cold-start spam review detection with adversarial learning," *Inf. Sci.*, vol. 526, pp. 274–288, Jul. 2020, doi: 10.1016/j.ins.2020.03.063.

[80] S. Kennedy, N. Walsh, K. Sloka, A. McCarren, and J. Foster, "Fact or factitious? Contextualized opinion spam detection," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, Student Res. Workshop*, 2019, pp. 344–350, doi: 10.18653/v1/p19-2048.

[81] T. Kumaravel and B. Bizu, "Convolutional neural network for customer's opinion on Amazon products," *Int. J. Recent Technol. Eng.*, vol. 8, no. 3, pp. 6634–6643, 2019, doi: 10.35940/ijrte.C5670.098319.

[82] L. Li, W. Ren, B. Qin, and T. Liu, "Learning document representation for deceptive opinion spam detection," in *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data* (Lecture Notes in Computer Science), vol. 9427. Nanjing, China: Springer, 2015, pp. 393–404, doi: 10.1007/978-3-319-25816-4_32.

[83] S. Zhao, Z. Xu, L. Liu, M. Guo, and J. Yun, "Towards accurate deceptive opinions detection based on word order-preserving CNN," *Math. Problems Eng.*, vol. 2018, pp. 1–9, Sep. 2018, doi: 10.1155/2018/2410206.

[84] Q. Li, Q. Wu, C. Zhu, J. Zhang, and W. Zhao, "An inferable representation learning for fraud review detection with cold-start problem," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8, doi: 10.1109/IJCNN.2019.8852437.

[85] W. Zhang, Y. Du, T. Yoshida, and Q. Wang, "DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network," *Inf. Process. Manage.*, vol. 54, no. 4, pp. 576–592, 2018, doi: 10.1016/j.ipm.2018.03.007.

[86] C.-C. Wang, M.-Y. Day, C.-C. Chen, and J.-W. Liou, "Detecting spamming reviews using long short-term memory recurrent neural network framework," in *Proc. 2nd Int. Conf. E-commerce, E-Business E-Government (ICEEG)*, 2018, pp. 16–20, doi: 10.1145/3234781.3234794.

[87] W. Liu, W. Jing, and Y. Li, "Incorporating feature representation into BiLSTM for deceptive review detection," *Computing*, vol. 102, pp. 701–715, Nov. 2020, doi: 10.1007/s00607-019-00763-y.

[88] Z.-Y. Zeng, J.-J. Lin, M.-S. Chen, M.-H. Chen, Y.-Q. Lan, and J.-L. Liu, "A review structure based ensemble model for deceptive review spam," *Information*, vol. 10, no. 7, p. 243, Jul. 2019, doi: 10.3390/info10070243.

[89] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proc. Int. Conf. Web Search Web Data Mining (WSDM)*, 2008, pp. 219–229, doi: 10.1145/1341531.1341560.

[90] S. Mani, S. Kumari, A. Jain, and P. Kumar, "Spam review detection using ensemble machine learning," in *Machine Learning and Data Mining in Pattern Recognition* (Lecture Notes in Computer Science), vol. 10935. New York, NY, USA: Springer, 2018, pp. 198–209, doi: 10.1007/978-3-319-96133-0_15.

[91] D. B. Buller and J. K. Burgoon, "Interpersonal deception theory," *Commun. Theory*, vol. 6, no. 3, pp. 203–242, Aug. 1996, doi: 10.1111/j.1468-2885.1996.tb00127.x.

[92] S. Banerjee and A. Y. K. Chua, "Authentic versus fictitious online reviews: A textual analysis across luxury, budget, and mid-range hotels," *J. Inf. Sci.*, vol. 43, no. 1, pp. 122–134, Feb. 2017, doi: 10.1177/0165551515625027.

[93] S. A. McCornack, "Information manipulation theory," *Commun. Monographs*, vol. 59, no. 1, pp. 1–16, Mar. 1992, doi: 10.1080/03637759209376245.

[94] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, Feb. 1969, doi: 10.1080/00332747.1969.11023575.

[95] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception," *Psychol. Bull.*, vol. 129, no. 1, pp. 74–118, 2003, doi: 10.1037/0033-2909.129.1.74.
[96] B. M. DePaulo, "Nonverbal behavior and self-presentation," *Psychol. Bull.*, vol. 111, no. 2, pp. 203–243, Mar. 1992, doi: 10.1037/0033-2909.111.2.203.
[97] M. K. Johnson and C. L. Raye, "Reality monitoring," *Psychol. Rev.*, vol. 88, no. 1, pp. 67–85, Jan. 1981, doi: 10.1037/0033-295X.88.1.67.
[98] M. K. Johnson and C. L. Raye, "False memories and confabulation," *Trends Cognit. Sci.*, vol. 2, no. 4, pp. 137–145, 1998, doi: 10.1016/S1364-6613(98)01152-8.
[99] L. Zhou, D. P. Twitchell, T. Qin, J. K. Burgoon, and J. F. Nunamaker, "An exploratory study into deception detection in text-based computer-mediated communication," in *Proc. 36th Annu. Hawaii Int. Conf. Syst. Sci.*, Jan. 2003, p. 10, doi: 10.1109/HICSS.2003.1173793.
[100] L. Zhou, J. K. Burgoon, J. F. Nunamaker, and D. Twitchell, "Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications," *Group Decis. Negotiation*, vol. 13, no. 1, pp. 81–106, Jan. 2004, doi: 10.1023/B:GRUP.0000011944.62889.6f.
[101] L. Zhou, J. K. Burgoon, D. P. Twitchell, T. Qin, and J. F. Nunamaker, "A comparison of classification methods for predicting deception in computer-mediated communication," *J. Manage. Inf. Syst.*, vol. 20, no. 4, pp. 139–166, Mar. 2004, doi: 10.1080/07421222.2004.11045779.
[102] R. L. Daft and R. H. Lengel, "Organizational information requirements, media richness and structural design," *Manage. Sci.*, vol. 32, no. 5, pp. 554–571, May 1986, doi: 10.1287/mnsc.32.5.554.
[103] J. R. Carlson and R. W. Zmud, "Channel expansion theory and the experiential nature of media richness perceptions," *Acad. Manage. J.*, vol. 42, no. 2, pp. 153–170, Apr. 1999, doi: 10.2307/257090.
[104] D. B. Buller, J. K. Burgoon, A. Buslig, and J. Roiger, "Testing interpersonal deception theory: The language of interpersonal deception," *Commun. Theory*, vol. 6, no. 3, pp. 268–289, Aug. 1996, doi: 10.1111/j.1468-2885.1996.tb00129.x.
[105] G. R. Miller and J. B. Stiff, "Applied issues in studying deceptive communication," in *The Applications of Nonverbal Behavioral Theories and Research*. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1992, pp. 217–237.
[106] M. Steller and G. Köhnken, "Criteria-based content analysis," in *Psychological Methods in Criminal Investigation and Evidence*, D. Raskin, Ed. New York, NY, USA: Springer-Verlag, 1989, pp. 217–245.
[107] U. Undeutsch, "Beurteilung der Glaubhaftigkeit von Aussagen," in *Handbuch der Psychologie Bd. 11: Forensische Psychologie*. Göttingen, Germany: Hogrefe, 1967, pp. 26–181.
[108] A. Sapir, "The LSI course on scientific content analysis (SCAN)," Lab. Sci. Interrogation, Phoenix, ZA, USA, Tech. Rep., 1987.
[109] A. Mehrabian and M. Wiener, "Non-immediacy between communicator and object of communication in a verbal message: Application to the inference of attitudes," *J. Consulting Psychol.*, vol. 30, no. 5, pp. 420–425, 1966, doi: 10.1037/h0023813.
[110] T. Qin, J. K. Burgoon, J. P. Blair, and J. F. Nunamaker, "Modality effects in deception detection and applications in automatic-deception-detection," in *Proc. 38th Annu. Hawaii Int. Conf. Syst. Sci.*, 2005, p. 23, doi: 10.1109/hicss.2005.436.
[111] B. Kleinberg, M. Mozes, A. Arntz, and B. Verschuere, "Using named entities for computer-automated verbal deception detection," *J. Forensic Sci.*, vol. 63, no. 3, pp. 714–723, May 2018, doi: 10.1111/1556-4029.13645.
[112] G. Nahari, A. Vrij, and R. P. Fisher, "Exploiting liars' verbal strategies by examining the verifiability of details," *Legal Criminol. Psychol.*, vol. 19, no. 2, pp. 227–239, Sep. 2014, doi: 10.1111/j.2044-8333.2012.02069.x.
[113] C. M. Fuller, D. P. Biros, and R. L. Wilson, "Decision support for determining veracity via linguistic-based cues," *Decis. Support Syst.*, vol. 46, no. 3, pp. 695–703, Feb. 2009, doi: 10.1016/j.dss.2008.11.001.
[114] M. Zuckerman, B. M. Depaulo, and R. Rosenthal, "Verbal and nonverbal communication of deception," *Adv. Exp. Soc. Psychol.*, vol. 14, pp. 1–59, Jan. 1981, doi: 10.1016/S0065-2601(08)60369-X.
[115] D. C. Derrick, T. O. Meservy, J. L. Jenkins, J. K. Burgoon, and J. F. Nunamaker, "Detecting deceptive chat-based communication using typing behavior and message cues," *ACM Trans. Manage. Inf. Syst.*, vol. 4, no. 2, pp. 1–21, Aug. 2013, doi: 10.1145/2499962.2499967.
[116] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cogn. Sci.*, vol. 12, no. 2, pp. 257–285, 1988, doi: 10.1016/0364-0213(88)90023-7.
[117] A. Vrij, R. Fisher, S. Mann, and S. Leal, "A cognitive load approach to lie detection," *J. Investigative Psychol. Offender Profiling*, vol. 5, nos. 1–2, pp. 39–43, 2008, doi: 10.1002/jip.82.
[118] A. Vrij, S. Mann, S. Kristen, and R. P. Fisher, "Cues to deception and ability to detect lies as a function of police interview styles," *Law Hum. Behav.*, vol. 31, no. 5, pp. 499–518, 2007, doi: 10.1007/s10979-006-9066-4.
[119] A. Vrij, *Detecting Lies and Deceit: Pitfalls and Opportunities* (Series in the Psychology of Crime, Policing and Law). Hoboken, NJ, USA: Wiley, 2008. [Online]. Available: http://books.google.com/books?hl=en&lr=&id=20pg76wmAucC&oi=fnd&pg=PR7&dq=A.+Vrij.+2008.+Detecting+lies+and+deceit:+Pitfalls+and+opportunities&ots=wdmrfkCidc&sig=Rmj4auTiHr2s9Lomo5ytytrj7QY#v=onepage&q&f=false
[120] X. Liu, J. Hancock, G. Zhang, R. Xu, D. Markowitz, and N. Bazarova, "Exploring linguistic features for deception detection in unstructured text," in *Proc. Rapid Screening Technol., Deception Detection Credibility Assessment Symp.*, 2012, pp. 1–10, doi: 10.1109/.HICSS.2003.1173793.
[121] T. Ong, M. Mannino, and D. Gregg, "Linguistic characteristics of shill reviews," *Electron. Commerce Res. Appl.*, vol. 13, no. 2, pp. 69–78, Mar. 2014, doi: 10.1016/j.elerap.2013.10.002.
[122] K.-H. Yoo and U. Gretzel, "Comparison of deceptive and truthful travel reviews," in *Information and Communication Technologies in Tourism*. Vienna, Austria: Springer, 2009, pp. 37–47, doi: 10.1007/978-3-211-93971-0_4.
[123] X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, "Fake news early detection: A theory-driven model," *Digit. Threats, Res. Pract.*, vol. 1, no. 2, pp. 1–25, Jun. 2020, doi: 10.1145/3377478.
[124] T. Chang, P. Y. Hsu, M. S. Cheng, C. Y. Chung, and Y. L. Chung, "Detecting fake review with rumor model—Case study in hotel review," in *Intelligence Science and Big Data Engineering. Big Data and Machine Learning Techniques* (Lecture Notes in Computer Science), vol. 9243. Cham, Switzerland: Springer, 2015, pp. 181–192, doi: 10.1007/978-3-319-23862-3_18.
[125] E. Abedin, A. Mendoza, and S. Karunasekera, "Towards a credibility analysis model for online reviews," in *Proc. 23rd Pacific Asia Conf. Inf. Syst., Secure ICT Platform 4th Ind. Revolution (PACIS)*, 2019, pp. 1–9.
[126] C. Luo, J. Wu, Y. Shi, and Y. Xu, "The effects of individualism–collectivism cultural orientation on eWOM information," *Int. J. Inf. Manage.*, vol. 34, no. 4, pp. 446–456, Aug. 2014, doi: 10.1016/j.ijinfomgt.2014.04.001.
[127] Y. Huang, C. Li, J. Wu, and Z. Lin, "Online customer reviews and consumer evaluation: The role of review font," *Inf. Manage.*, vol. 55, no. 4, pp. 430–440, Jun. 2018, doi: 10.1016/j.im.2017.10.003.
[128] C. M.-Y. Cheung, C. L. Sia, and K. K. Y. Kuan "Is this review believable? A study of factors affecting the credibility of online consumer reviews from an ELM perspective," *J. Assoc. Inf. Syst.*, vol. 13, no. 8, pp. 618–635, Aug. 2012, doi: 10.17705/1jais.00305.
[129] S. Ketron, "Investigating the effect of quality of grammar and mechanics (QGAM) in online reviews: The mediating role of reviewer crediblity," *J. Bus. Res.*, vol. 81, pp. 51–59, Dec. 2017, doi: 10.1016/j.jbusres.2017.08.008.
[130] B. Lis, "In eWOM we trust: A framework of factors that determine the eWOM credibility," *Bus. Inf. Syst. Eng.*, vol. 5, no. 3, pp. 129–140, Jun. 2013, doi: 10.1007/s12599-013-0261-9.
[131] C. Luo, X. Luo, L. Schatzberg, and C. L. Sia, "Impact of informational factors on online recommendation credibility: The moderating role of source credibility," *Decision Support Syst.*, vol. 56, pp. 92–102, Dec. 2013, doi: 10.1016/j.dss.2013.05.005.
[132] H. Baek, J. Ahn, and Y. Choi, "Helpfulness of online consumer reviews: Readers' objectives and review cues," *Int. J. Electron. Commerce*, vol. 17, no. 2, pp. 99–126, Dec. 2012, doi: 10.2753/JEC1086-4415170204.
[133] C. Luo, X. Luo, Y. Xu, M. Warkentin, and C. L. Sia, "Examining the moderating role of sense of membership in online review evaluations," *Inf. Manage.*, vol. 52, no. 3, pp. 305–316, Apr. 2015, doi: 10.1016/j.im.2014.12.008.
[134] M. Y. Cheung, C. Luo, C. L. Sia, and H. Chen, "Credibility of electronic word-of-mouth: Informational and normative determinants of on-line consumer recommendations," *Int. J. Electron. Commerce*, vol. 13, no. 4, pp. 9–38, 2009, doi: 10.2753/JEC1086-4415130402.
[135] V. L. Rubin and T. Lukoianova, "Truth and deception at the rhetorical structure level," *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 5, pp. 905–917, May 2015, doi: 10.1002/asi.23216.
[136] V. L. Rubin and T. Vashchilko, "Identification of truth and deception in text: Application of vector space model to rhetorical structure theory," *Proc. EACL Workshop Comput. Approaches Decept. Detect.*, 2012, pp. 97–106.

[137] M. S. Dogo, P. Deepak, and A. Jurek-Loughrey, "Exploring thematic coherence in fake news," in *Communications in Computer and Information Science*, vol. 1323. Ghent, Belgium: Springer, 2020, pp. 571–580, doi: 10.1007/978-3-030-65965-3_40.

[138] O. Popoola, "Detecting fake Amazon book reviews using rhetorical structure theory," in *Proc. Misinformation Misbehavior Mining Web (MIS)*, 2018, pp. 2–7. [Online]. Available: http://snap.stanford.edu/mis2/files/MIS2_paper_20.pdf

[139] O. Popoola, "Using rhetorical structure theory for detection of fake online reviews," in *Proc. 6th Workshop Recent Adv. RST Rel. Formalisms*, 2017, pp. 58–63, doi: 10.18653/v1/w17-3608.

[140] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: Predicting deception from linguistic styles," *Personality Social Psycho. Bull.*, vol. 29, no. 5, pp. 665–675, May 2003, doi: 10.1177/0146167203029005010.

[141] R. L. Oliver, "Expectancy theory predictions of Salesmen's performance," *J. Marketing Res.*, vol. 11, no. 3, pp. 243–253, Aug. 1974, doi: 10.1177/002224377401100302.

[142] G. W. Allport and L. Postman, "An analysis of rumor," *Public Opin. Q.*, vol. 10, no. 4, pp. 501–517, 1946, doi: 10.1093/poq/10.4.501.

[143] M. Deutsch and H. B. Gerard, "A study of normative and informational social influences upon individual judgment," *J. Abnormal Social Psychol.*, vol. 51, no. 3, pp. 629–636, 1955, doi: 10.1037/h0046408.

[144] S. Chaiken, "Heuristic versus systematic information processing and the use of source versus message cues in persuasion," *J. Personality Social Psychol.*, vol. 39, no. 5, pp. 752–766, Nov. 1980, doi: 10.1037/0022-3514.39.5.752.

[145] R. E. Petty and J. T. Cacioppo, "The elaboration likelihood model of persuasion," *Adv. Exp. Soc. Psychol.*, vol. 19, pp. 123–205, Jan. 1986, doi: 10.1016/S0065-2601(08)60214-2.

[146] W. C. Mann and S. A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization," *Text Interdiscipl. J. Study Discourse*, vol. 8, no. 3, pp. 243–281, 1988, doi: 10.1515/text.1.1988.8.3.243.

[147] T. R. Levine, "Truth-default theory (TDT): A theory of human deception and deception detection," *J. Lang. Social Psychol.*, vol. 33, no. 4, pp. 378–392, Sep. 2014, doi: 10.1177/0261927X14535916.

[148] R. H. Fazio and M. P. Zanna, "Direct experience and attitude-behavior consistency," *Adv. Exp. Soc. Psychol.*, vol. 14, pp. 161–202, Jan. 1981, doi: 10.1016/S0065-2601(08)60372-X.

[149] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2010, pp. 939–948, doi: 10.1145/1871437.1871557.

[150] A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, and R. Ghosh, "Spotting opinion spammers using behavioral footprints," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2013, pp. 632–640, doi: 10.1145/2487575.2487580.

[151] S. Xie, G. Wang, S. Lin, and P. S. Yu, "Review spam detection via temporal pattern discovery," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2012, pp. 823–831, doi: 10.1145/2339530.2339662.

[152] K. Goswami, Y. Park, and C. Song, "Impact of reviewer social interaction on online consumer review fraud detection," *J. Big Data*, vol. 4, no. 1, pp. 1–9, Dec. 2017, doi: 10.1186/s40537-017-0075-6.

[153] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," *J. Big Data*, vol. 2, no. 1, pp. 1–24, Dec. 2015, doi: 10.1186/s40537-015-0029-9.

[154] F. Li, M. Huang, Y. Yang, and X. Zhu, "Learning to identify review spam," in *Proc. IJCAI Int. Joint Conf. Artif. Intell.*, 2011, pp. 2488–2493, doi: 10.5591/978-1-57735-516-8/IJCAI11-414.

[155] S.-J. Ji, Q. Zhang, J. Li, D. K. W. Chiu, S. Xu, L. Yi, and M. Gong, "A burst-based unsupervised method for detecting review spammer groups," *Inf. Sci.*, vol. 536, pp. 454–469, Oct. 2020, doi: 10.1016/j.ins.2020.05.084.

[156] R. Filieri, "What makes an online consumer review trustworthy?" *Ann. Tourism Res.*, vol. 58, pp. 46–64, May 2016, doi: 10.1016/j.annals.2015.12.019.

[157] S. Banerjee, S. Bhattacharyya, and I. Bose, "Whose online reviews to trust? Understanding reviewer trustworthiness and its impact on business," *Decis. Support Syst.*, vol. 96, pp. 17–26, Apr. 2017, doi: 10.1016/j.dss.2017.01.006.

[158] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proc. 20th Int. Conf. World Wide Web (WWW)*, 2011, pp. 675–684, doi: 10.1145/1963405.1963500.

[159] E. Abedin, A. Mendoza, and S. Karunasekera, "What makes a review credible? Heuristic and systematic factors for the credibility of online reviews," in *Proc. Australas. Conf. Inf. Syst.*, Jan. 2019, pp. 701–711. Accessed: May 17, 2022. [Online]. Available: https://aisel.aisnet.org/acis2019/75

[160] Z. Liu and S. Park, "What makes a useful online review? Implication for travel product Websites," *Tourism Manage.*, vol. 47, pp. 140–151, Apr. 2015, doi: 10.1016/j.tourman.2014.09.020.

[161] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of Twitter accounts: Are you a human, bot, or cyborg?" *IEEE Trans. Dependable Secure Comput.*, vol. 9, no. 6, pp. 811–824, Aug. 2012, doi: 10.1109/TDSC.2012.75.

[162] C. Chang, Y. Zhang, C. Szabo, and Q. Z. Sheng, "Extreme user and political rumor detection on Twitter," in *Advanced Data Mining and Applications* (Lecture Notes in Computer Science), vol. 10086. Gold Coast, QLD, Australia: Springer, 2016, pp. 751–763, doi: 10.1007/978-3-319-49586-6_54.

[163] S. Rayana and L. Akoglu, "Collective opinion spam detection: Bridging review networks and metadata," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2015, pp. 985–994, doi: 10.1145/2783258.2783370.

[164] J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, "Sentiment analysis: A review and comparative analysis of Web services," *Inf. Sci.*, vol. 311, pp. 18–38, Aug. 2015, doi: 10.1016/j.ins.2015.03.040.

[165] A. U. Akram, H. U. Khan, S. Iqbal, T. Iqbal, E. U. Munir, and M. Shafi, "Finding rotten eggs: A review spam detection model using diverse feature sets," *KSII Trans. Internet Inf. Syst.*, vol. 12, no. 10, pp. 5120–5142, Oct. 2018, doi: 10.3837/tiis.2018.10.026.

[166] D. Plotkina, A. Munzel, and J. Pallud, "Illusions of truth—Experimental insights into human and algorithmic detections of fake online reviews," *J. Bus. Res.*, vol. 109, pp. 511–523, Mar. 2020, doi: 10.1016/j.jbusres.2018.12.009.

[167] M. L. KNAPP and M. E. Comaden, "Telling it like it isn't: A review of theory and research on deceptive communications," *Hum. Commun. Res.*, vol. 5, no. 3, pp. 270–285, Mar. 1979, doi: 10.1111/j.1468-2958.1979.tb00640.x.

[168] A. C. Graesser, D. S. McNamara, and J. M. Kulikowich, "Coh-metrix: Providing multilevel analyses of text characteristics," *Educ. Researcher*, vol. 40, no. 5, pp. 223–234, Jun. 2011, doi: 10.3102/0013189X11413260.

[169] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*. Lake Tahoe, NV, USA: Curran, 2013, pp. 3111–3119.

[170] P. Hajek, A. Barushka, and M. Munk, "Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining," *Neural Comput. Appl.*, vol. 32, no. 23, pp. 17259–17274, Dec. 2020, doi: 10.1007/s00521-020-04757-2.

[171] S. G. Tesfagergish, R. Damaševičius, and J. Kapočiūtė-Dzikienė, "Deep fake recognition in Tweets using text augmentation, word embeddings and deep learning," in *Computational Science and Its Applications* (Lecture Notes in Computer Science), vol. 12954. 2021, pp. 523–538, doi: 10.1007/978-3-030-86979-3_37.

[172] P. Ekman, *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. New York, NY, USA: Norton, 1985.

[173] T. R. Levine and S. A. McCornack, "Theorizing about deception," *J. Lang. Social Psychol.*, vol. 33, no. 4, pp. 431–440, Sep. 2014, doi: 10.1177/0261927X14536397.

[174] P. Ekman, "Deception, lying, and demeanor," in *States of Mind: American and Post-Soviet Perspectives on Contemporary Issues in Psychology*. New York, NY, USA: Oxford Univ. Press, 1997, pp. 93–105. [Online]. Available: http://books.google.com/books?hl=en&lr=&id=3Iqu_0OC8d0C&oi=fnd&pg=PA93&dq=Deception,+Lying,+and+Demeanor&ots=vZlhzUMwzd&sig=rvRtwSjJJsX1sB_bgeBC3aPDW94

[175] P. Ekman, "Lying and nonverbal behavior: Theoretical issues and new findings," *J. Nonverbal Behav.*, vol. 12, no. 3, pp. 163–175, 1988, doi: 10.1007/BF00987486.

[176] A. Nortje and C. Tredoux, "How good are we at detecting deception? A review of current techniques and theories," *South Afr. J. Psychol.*, vol. 49, no. 4, pp. 491–504, Dec. 2019, doi: 10.1177/0081246318822953.

[177] C. F. Bond, T. R. Levine, and M. Hartwig, "New findings in nonverbal lie detection," in *Detecting Deception: Current Challenges and Cognitive Approaches*. Chichester, U.K.: Wiley, 2015, pp. 37–58, doi: 10.1002/9781118510001.ch2.

[178] J. Masip, S. L. Sporer, E. Garrido, and C. Herrero, "The detection of deception with the reality monitoring approach: A review of the empirical evidence," *Psychol., Crime Law*, vol. 11, no. 1, pp. 99–122, Mar. 2005, doi: 10.1080/10683160410001726356.
[179] G. Köhnken, "Statement validity analysis and the 'detection of the truth," in *The Detection of Deception in Forensic Contexts*. Cambridge, U.K.: Cambridge Univ. Press, 2004, pp. 41–63, doi: 10.1017/CBO9780511490071.003.
[180] N. Smith, "Reading between the lines?: An evaluation of the scientific content analysis technique (SCAN)," Home Office Policing Reducing Crime Unit Res. Develop. Statist. Directorate, London, U.K., Tech. Rep. Police Research Series Paper 135, 2001.
[181] L. N. Driscoll, "A validity assessment of written statements from suspects in criminal investigations using the scan technique," *Police Stud. Int. Rev. Police Dev.*, vol. 17, p. 77, Jan. 1994.
[182] G. Nahari, A. Vrij, and R. P. Fisher, "Does the truth come out in the writing? Scan as a lie detection tool," *Law Hum. Behav.*, vol. 36, no. 1, pp. 68–76, 2012, doi: 10.1037/h0093965.
[183] G. Nahari, A. Vrij, and R. P. Fisher, "The verifiability approach: Countermeasures facilitate its ability to discriminate between truths and lies," *Appl. Cognit. Psychol.*, vol. 28, no. 1, pp. 122–128, Jan. 2014, doi: 10.1002/acp.2974.
[184] A. Vrij, P. Taylor, and I. Picornell, "Verbal lie detection," in *Communication in Investigative and Legal Contexts*. Hoboken, NJ, USA: Wiley, 2015, pp. 259–286, doi: 10.1002/9781118769133.ch12.
[185] A. Vrij and G. Ganis, "Theories in deception and lie detection," in *Credibility Assessment: Scientific Research and Applications*. Oxford, U.K.: Academic, 2014, pp. 301–374, doi: 10.1016/B978-0-12-394433-7.00007-5.
[186] G. Nahari, "The applicability of the verifiability approach to the real world," in *Detecting Concealed Information and Deception: Recent Developments*, J. P. Rosenfeld, Ed. London, U.K.: Elsevier, 2018, pp. 329–349, doi: 10.1016/B978-0-12-812729-2.00014-8.
[187] S. A. Mccornack and M. R. Parks, "Deception detection and relationship development: The other side of trust," *Ann. Int. Commun. Assoc.*, vol. 9, no. 1, pp. 377–389, Jan. 1986, doi: 10.1080/23808985.1986.11678616.
[188] T. R. Levine, *Duped: Truth-Default Theory and the Social Science of Lying and Deception*. Tuscaloosa, AL, USA: Univ. of Alabama Press, 2020.
[189] P. A. Granhag and L. A. Strömwall, "Repeated interrogations–stretching the deception detection paradigm," *Expert Evid.*, vol. 7, pp. 163–174, Sep. 1999.
[190] J. K. Burgoon and T. Qin, "The dynamic nature of deceptive verbal communication," *J. Lang. Social Psychol.*, vol. 25, no. 1, pp. 76–96, Mar. 2006, doi: 10.1177/0261927X05284482.
[191] L. Smith. (2018). *The State of Deception Detection Research: Two Perspectives Used to Uncover Deception Detection Methods*. Kansas State University. [Online]. Available: http://hdl.handle.net/2097/39148
[192] H. S. Park and T. R. Levine, "Base rates, deception detection, and deception theory: A reply to burgoon (2015)," *Hum. Commun. Res.*, vol. 41, no. 3, pp. 350–366, Jul. 2015, doi: 10.1111/hcre.12066.
[193] P. Grice, *Studies in the Way of Words*. Cambridge, MA, USA: Harvard Univ. Press, 1989.
[194] S. A. McCornack, K. Morrison, J. E. Paik, A. M. Wisner, and X. Zhu, "Information manipulation theory 2: A propositional theory of deceptive discourse production," *J. Lang. Social Psychol.*, vol. 33, no. 4, pp. 348–377, Sep. 2014, doi: 10.1177/0261927X14534656.
[195] R. Banerjee, S. Feng, J. S. Kang, and Y. Choi, "Keystroke patterns as prosody in digital writings: A case study with deceptive reviews and essays," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1469–1473, doi: 10.3115/v1/d14-1155.
[196] M. Al-Mosaiwi and T. Johnstone, "In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation," *Clin. Psychol. Sci.*, vol. 6, no. 4, pp. 529–542, Jul. 2018, doi: 10.1177/2167702617747074.
[197] B. Kleinberg, G. Nahari, and B. Verschuere, "Using the verifiability of details as a test of deception: A conceptual framework for the automation of the verifiability approach," in *Proc. 2nd Workshop Comput. Approaches Deception Detection*, 2016, pp. 18–25, doi: 10.18653/v1/w16-0803.
[198] A. Sepehri, D. M. Markowitzand, and M. Mir, "PassivePy: A tool to automatically identify passive voice in big text data," PsyArXiv, Feb. 2022. [Online]. Available: http://psyarxiv.com/bwp3t
[199] A. Sepehri, D. M. Markowitz, and R. Duclos, "The location of maximum emotion in deceptive and truthful texts," *Soc. Psychol. Personal. Sci.*, vol. 12, no. 6, pp. 996–1004, 2021, doi: 10.1177/1948550620949730.
[200] S. Rayana and L. Akoglu, "Collective opinion spam detection using active inference," in *Proc. SIAM Int. Conf. Data Mining*, Jun. 2016, pp. 630–638, doi: 10.1137/1.9781611974348.71.
[201] A. Rastogi and M. Mehrotra, "Impact of behavioral and textual features on opinion spam detection," in *Proc. 2nd Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, Jun. 2018, pp. 852–857, doi: 10.1109/ICCONS.2018.8662912.
[202] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011. [Online]. Available: http://scikit-learn.sourceforge.net
[203] H. Zhang, "The optimality of Naive Bayes," in *Proc. 17th Int. Florida Artif. Intell. Res. Soc. Conf.*, vol. 2, 2004, pp. 562–567.
[204] L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees*. London, U.K.: Chapman & Hall, 1984.
[205] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[206] K. Gajowniczek, T. Ząbkowski, and R. Szupiluk, "Estimating the ROC curve and its significance for classification models' assessment," *Quantum Methods Econ.*, vol. 15, no. 2, pp. 382–391, 2014.
[207] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.
[208] A. Altmann, L. Tolosi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, May 2010, doi: 10.1093/bioinformatics/btq134.
[209] T. Vanta and M. Aono, "Fake review detection focusing on emotional expressions and extreme rating," in *Proc. 25th Annu. Conf. Lang. Process. Soc. (NLP)*, 2019, pp. 422–425. Accessed: Sep. 8, 2022. [Online]. Available: http://www.ijarsct.co.in/Paper96.pdf
[210] M. F. Uchoa. (2018). *Detecting Fake Reviews With Machine Learning*. Accessed: Sep. 8, 2022. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:du-28133
[211] A. Rastogi, M. Mehrotra, and S. S. Ali, "Effective opinion spam detection: A study on review metadata versus content," *J. Data Inf. Sci.*, vol. 5, no. 2, pp. 76–110, Apr. 2020, doi: 10.2478/jdis-2020-0013.
[212] C. Yuan, W. Zhou, Q. Ma, S. Lv, J. Han, and S. Hu, "Learning review representations from user and product level information for spam detection," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 1444–1449, doi: 10.1109/ICDM.2019.00188.
[213] G. Wang, S. Xie, B. Liu, and P. S. Yu, "Review graph based online store review spammer detection," in *Proc. IEEE 11th Int. Conf. Data Mining*, Dec. 2011, pp. 1242–1247, doi: 10.1109/ICDM.2011.124.
[214] X. Wang, K. Liu, S. He, and J. Zhao, "Learning to represent review with tensor decomposition for spam detection," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 866–875, doi: 10.18653/v1/d16-1083.
[215] H. Li, G. Fei, S. Wang, B. Liu, W. Shao, A. Mukherjee, and J. Shao, "Bimodal distribution and co-bursting in review spam detection," in *Proc. 26th Int. Conf. World Wide Web*, Apr. 2017, pp. 1063–1072, doi: 10.1145/3038912.3052582.
[216] S. K. Maity, K. C. Santosh, and A. Mukherjee, "Spam2Vec: Learning biased embeddings for spam detection in Twitter," in *Proc. Companion The Web Conf. Web Conf. (WWW)*, 2018, pp. 63–64, doi: 10.1145/3184558.3186930.
[217] L. Li, B. Qin, W. Ren, and T. Liu, "Document representation and feature combination for deceptive spam review detection," *Neurocomputing*, vol. 254, pp. 33–41, Sep. 2017, doi: 10.1016/j.neucom.2016.10.080.
[218] X. Wang, K. Liu, and J. Zhao, "Detecting deceptive review spam via attention-based neural networks," in *Natural Language Processing and Chinese Computing* (Lecture Notes in Computer Science), vol. 10619. Cham, Switzerland: Springer, 2017, pp. 866–876, doi: 10.1007/978-3-319-73618-1_76.
[219] Y. Liu, L. Wang, T. Shi, and J. Li, "Detection of spam reviews through a hierarchical attention architecture with N-gram CNN and bi-LSTM," *Inf. Syst.*, vol. 103, Jan. 2022, Art. no. 101865, doi: 10.1016/j.is.2021.101865.
[220] M. S. Javed, H. Majeed, H. Mujtaba, and M. O. Beg, "Fake reviews classification using deep learning ensemble of shallow convolutions," *J. Comput. Social Sci.*, vol. 4, no. 2, pp. 883–902, Nov. 2021, doi: 10.1007/s42001-021-00114-y.
[221] M. S. Jacob and P. S. Rajendran, "Fuzzy artificial bee colony-based CNN-LSTM and semantic feature for fake product review classification," *Concurrency Comput., Pract. Exper.*, vol. 34, no. 1, Jan. 2022, doi: 10.1002/cpe.6539.

MUJAHED ABDULQADER received the bachelor's degree in computer engineering from Al-Balqa' Applied University, Amman, Jordan, in 2012. He is currently pursuing the master's degree in data science with the Islamic University of Madinah. His research interests include fake reviews, natural language processing, artificial intelligence, machine learning techniques, deep learning, and embedded systems.

ABDALLAH NAMOUN (Member, IEEE) received the bachelor's degree in computer science and the Ph.D. degree in informatics from The University of Manchester, U.K., in 2004 and 2009, respectively. He is currently an Associate Professor of intelligent interactive systems and the Head of the Information Systems Department, Faculty of Computer and Information Systems, Islamic University of Madinah. He has authored more than 60 publications in research areas spanning intelligent systems, human–computer interaction, software engineering, and technology acceptance and adoption. He has extensive experience in leading complex research projects (worth more than 21 million Euros) with several distinguished SMEs, such as SAP, BT, and ATOS. He has investigated user needs and interaction with modern interactive technologies, design of composite software services, and methods for testing the usability and acceptance of human-interfaces. His research interests include integrating state of the art artificial intelligence approaches in the design and development of interactive systems.

YAZED ALSAAWY received the Ph.D. degree in computer science from De Montfort University, U.K., in 2014. He is currently an Associate Professor (Ph.D.). He worked as the General Manager of the Digital Transformation Department, Ministry of Education, and a leader of a package of projects and initiatives at different levels in the Ministry of Education. Before that, he was the Dean of Information Technology Deanship at the Islamic University of Madinah. He works as an Assistant Professor at the Faculty of Computer and Information Systems and spent more than a year as the Vice Dean for Academic Affairs at the Faculty. He participated in Stanford Executive Program (SEP), in 2018. He works in the fields of security, privacy, and the IoT. He is the coauthor of many articles in software engineering, eLearning, networking, blockchain, security, and privacy. He participated in many founded projects. He submitted many research projects for funding at the King Abdul-Aziz City of Science and Technology (KACST).