IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 1, 2021, accepted April 21, 2021, date of publication April 26, 2021, date of current version May 6, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3075573

Fake Reviews Detection: A Survey

RAMI MOHAWESH[1], **SHUXIANG XU**[1], **SON N. TRAN**[1], **ROBERT OLLINGTON**[1], **MATTHEW SPRINGER**[1], **YASER JARARWEH**[2], **AND SUMBAL MAQSOOD**[1]

[1]School of Information and Communications Technology, University of Tasmania, Hobart, TAS 7005, Australia
[2]Computer Science Department, Jordan University of Science and Technology, Irbid 22110, Jordan

Corresponding author: Rami Mohawesh (rami.mohawesh@utas.edu.au)

ABSTRACT In e-commerce, user reviews can play a significant role in determining the revenue of an organisation. Online users rely on reviews before making decisions about any product and service. As such, the credibility of online reviews is crucial for businesses and can directly affect companies' reputation and profitability. That is why some businesses are paying spammers to post fake reviews. These fake reviews exploit consumer purchasing decisions. Consequently, the techniques for detecting fake reviews have extensively been explored in the past twelve years. However, there still lacks a survey that can analyse and summarise the existing approaches. To bridge up the issue, this survey paper details the task of fake review detection, summing up the existing datasets and their collection methods. It analyses the existing feature extraction techniques. It also summarises and analyses the existing techniques critically to identify gaps based on two groups: traditional statistical machine learning and deep learning methods. Further, we conduct a benchmark study to investigate the performance of different neural network models and transformers that have not been used for fake review detection yet. The experimental results on two benchmark datasets show that RoBERTa performs about 7% better than the state-of-the-art methods in a mixed domain for the deception dataset with the highest accuracy of 91.2%, which can be used as a baseline for future studies. Finally, we highlight the current gaps in this research area and the possible future directions.

INDEX TERMS Fake review, fake review detection, feature engineering, machine learning, deep learning.

I. INTRODUCTION

In this era of the internet, customers can post their reviews or opinions on several websites. These reviews are helpful for the organizations and for future consumers, who get an idea about products or services before making a selection [19]–[21]. In recent years, it has been observed that the number of customer reviews has increased significantly. Customer reviews affect the decision of potential buyers [33], [34]. In other words, when customers see reviews on social media, they determine whether to buy the product or reverse their purchasing decisions. Therefore, consumer reviews offer an invaluable service for individuals.

Positive reviews bring big financial gains, while negative reviews often exert a negative financial effect [47], [48]. Consequently, with customers becoming increasingly influential to the marketplace, there is a growing trend towards relying on customers' opinions to reshape businesses by enhancing products, services, and marketing [52]–[54]. For example, when several customers who purchased a specific model of Acer laptop posted reviews complaining about the low display quality, the manufacturer was inspired to produce a higher-resolution version of the laptop.

The way consumers openly express and use their feedback has contributed to issues with websites containing customer reviews. Social media (Twitter, Facebook, etc.) allows anyone to freely post feedback or critiques of any company at any time with no obligations or limits. The lack of restrictions, in turn, leads certain companies to use social media to unfairly promote their goods, brands or shops, or to unfairly criticise those of their rivals. For example, suppose a few consumers who bought a specific digital camera posted negative reviews on image quality. These reviews portray the digital camera unfavourably to the public. Thus, the camera manufacturer might employ an individual or team to post fake positive reviews about the camera. Similarly, in order to promote the company, the producer might ask the hired persons to post negative comments about competitors' products. Reviews published by people who have not personally encountered the items being reviewed are considered fake reviews [11]. Accordingly, a person who posts fake reviews is called a spammer [11]. When the spammer works with other

The associate editor coordinating the review of this manuscript and approving it for publication was Wentao Fan.

spammers to achieve a specific goal, the spammers are called a group of spammers [11].

Many studies have investigated the fake review detection problem and its challenges. The main task associated with fake review detection is classifying the review as fake or genuine. In this survey paper, we have presented a comprehensive survey of the literature to further identify existing problems for future directions in this research area.

It provides traditional statistical machine learning and deep learning techniques which will assist researchers, who are interested in fake review detection, to choose the best machine learning method. To help the reader easily understand the field of fake review detection, relevant publications from Google Scholar, Web of Sciences, and some high-profile conferences are presented in this paper to demonstrate the challenges in the field. Finally, papers from 2007 to 2021 have been identified for summary and analysis.

A. MAJOR DIFFERENCES AND CONTRIBUTIONS OF THIS SURVEY

This survey paper is not the first study conducted on fake review detection. Several others summarised the existing techniques for fake review detection [11], [23], [31], [39], [49], [58], [63], [73], [74]. However, these surveys have some limitations, as shown in Table 1. For example, they did not cover all aspects of fake reviews, such as all existing datasets and all recent deep learning algorithms. They did not provide insights about the impact of features on the detection models' performance. They did not provide a deep investigation for each existing model to identify efficient features in fake reviews detection. Furthermore, this survey paper provides the performance details of some promising models and gives some promising future directions for further study. This is an up-to-date survey paper related to fake reviews detection, which has tried to add all related datasets. The primary objective of this paper is to provide detailed, in-depth literature, existing methodologies, available datasets which may assist future work and improvements in this research domain. The key contributions of this paper can be summarised as follows:

- An exposition of features extraction techniques and how are they calculated. We also analyse the impact of features for the existing methods to determine the most appropriate features in fake reviews detection.
- Provide the existing datasets and their collection methods for future study. Furthermore, we summarise the necessary information of the datasets in Table 4, including the construction methods, the number of reviews in each dataset and related papers.
- We investigate the efficiency and accuracy of each method to find the most appropriate methods to detect fake reviews. We also critically analyse and summarise the existing techniques to identify the gaps.
- We investigate the performance of some promising models such as character-level convolutional -LSTM, convolutional -LSTM, HAN, convolutional HAN, BERT, DistilBERT, and RoBERTa that have not been used in fake review detection yet to the best of our knowledge.
- Summarise the main challenges facing fake review detection and key implications stemming from this study in section 6 before concluding the survey paper in section 7.

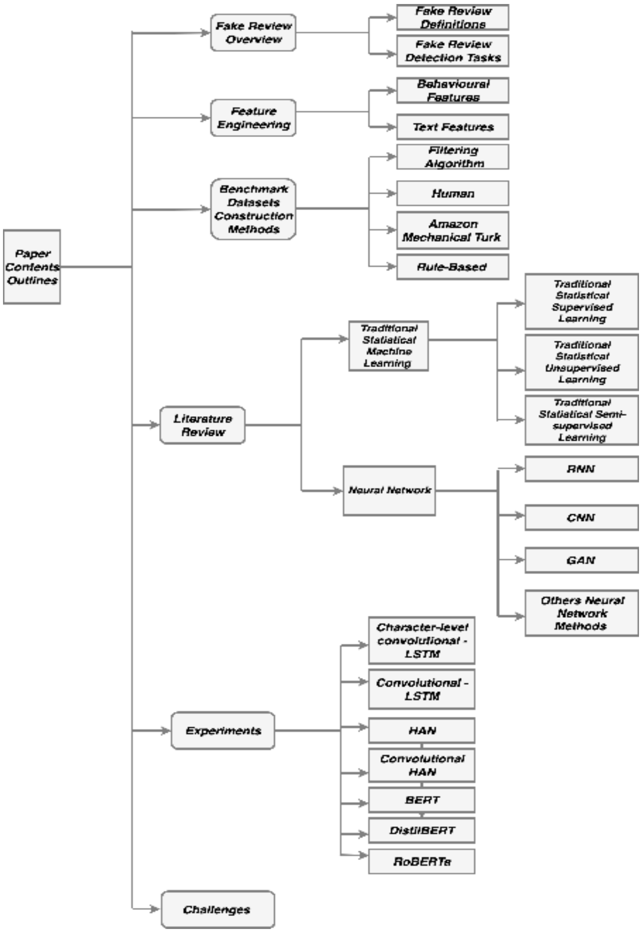
Paper Contents Outlines
Fake Review Overview
Fake Review Definitions
Fake Review Detection Tasks
Feature Engineering
Behavioural Features
Text Features
Benchmark Datasets Construction Methods
Filtering Algorithm
Human
Amazon Mechanical Turk
Rule-Based
Literature Review
Traditional Statistical Machine Learning
Traditional Statistical Supervised Learning
Traditional Statistical Unsupervised Learning
Traditional Statistical Semi-supervised Learning
Neural Network
RNN
CNN
GAN
Others Neural Network Methods
Experiments
Character-level convolutional -LSTM
Convolutional -LSTM
HAN
Convolutional HAN
BERT
DistilBERT
RoBERTa
Challenges

FIGURE 1. Outline of this survey.

This paper is organised as follows: Section 2 describes the existing feature extraction techniques. Section 3 describes detailed descriptions of the publicly available datasets with descriptions in addition to a summary Table. Section 4 presents the existing methods for fake review detection and the limitations for each technique, including traditional classical machine learning and neural network models. Section 5 presents the experiments of promising approaches for fake review detection. Section 6 presents the current gaps in this research area and the possible future direction. Section 7 presents the conclusion. Fig.1 shows the outline of this survey.

II. FAKE REVIEW OVERVIEW

Fake reviews are often defined as deceptive opinions, spam opinion, spam reviews, and their authors can be known as spammers. The spam opinion or what is known as a fake review can be categorized into three types [4]:

TABLE 1. Summary of recent surveys and reviews of fake review detection.

Ref.	Year	FE	DA	TM	DL	EX	FD	Description
Da Xu [2]	2014	**	-	***	**	-	**	• Provided limited future directions. • Summarised limited existing features engineering. • Summarised limited existing techniques in fake review detection.
Heydari [11]	2015	-	*	**	-	-	**	• Provided limited future directions in fake review detection. • Summarised limited existing techniques in fake review detection. • Summarised limited existing features without providing the feature engineering.
Patel [23]	2018	-	-	**	-	-	-	• Summarised limited exiting techniques.
Aslam [31]	2019	*	-	*	*	-	*	• Summarised limited exiting techniques without providing the existing datasets and existing features.
Visani [39]	2017	*	*	**	-	-	**	• Summarised limited exiting techniques. • Summarised limited existing features. • Provided limited future directions.
Rodrigu [49]	2020	-	-	**	**	-	-	• Summarised limited exiting techniques.
Vidanag [58]	2020	-	**	**	**	-	-	• Summarised limited exiting datasets. • Summarised limited exiting techniques. • Lack of providing future directions.
Wu [63]	2020	-	**	**	**	-	***	• Provided future directions. • Summarised the exiting datasets. • Summarised limited exiting techniques.
Our Survey	2021	***	***	***	***	***	***	• An exposition of features engineering techniques and how are they calculated. We also analyse the impact of features for the existing methods to determine the most appropriate features in fake reviews detection. • Provide the existing datasets and their collection methods for future research guide. Furthermore, we summarize the necessary information of the datasets including the construction methods with their pros and cons, and the number of reviews in each datasets and related papers. • We investigate the efficiency, accuracy for each method to find the most appropriate methods to detect fake reviews. We also analyse and summarize critically the existing techniques to identify the gaps. • investigated the performance of some promising models such as character-level convolutional -LSTM, convolutional -LSTM, HAN, convolutional HAN, BERT, DistilBERT, and RoBERTa. • Summarise the main challenges facing fake review detection.

(-) no discussion, (*) Low, (**) Medium, (***) High,

FE: Feature Extraction, DA: Datasets, TM: Traditional Machine Learning, DL: Deep Learning, EX: Experiments, FD: Future Directions.

- Untruthful opinions describe users who post negative reviews to damage a product/business's reputation or post positive reviews to promote a product/business. These reviews are called fake or deceptive reviews, and they are tough to detect simply by reading, as real and fake reviews are similar to each other [4].
- Reviews of a brand only describe those who are commenting on the brand of the products.
- Non-reviews that are irrelevant and offer no genuine opinion or are simply advertisements.

The last two types are called disruptive spam opinions, cause little threat and can be quickly identified to anyone by reading them [4]. To explain and understand the nature of fake reviews, we must consider the following two examples of reviews taken from a Yelp Chi real-life public dataset [8]. The first review is genuine, while the second is fake.

- Review 1: "*I like this hotel. The staff very friendly, you will feel like in home. Great location, great hotel to spend the night*:)".
- Review 2: "*What an awesome place to stay. The staff is amazing and so friendly. The perks, such as free bike rental, are nice. The history (and restoration) of the building is really cool. Thanks for making my stay so memorable.*".

We can conclude that it is challenging for humans to distinguish between two reviews by merely reading them. They are just too similar. Many researchers have manually annotated the reviews to classify them, and their model has only been able to achieve an accuracy of 60% [77]. As such, introducing efficient models to recognize fake reviews automatically is imperative [73].

A. FAKE REVIEW DETECTION TASKS

Whether it is known as spam review detection, fake opinion detection, and spam opinion detection, the main problem associated with fake review detection is classifying the review as either fake or genuine. Machine learning plays a significant role in fake review detection [73]. For example, supervised learning is one of the popular tasks in fake review detection, which requires labelled data to classify the fake review from genuine review based on specific features. Distinguishing a fake review from a truthful one by reading a large number of reviews is very difficult. Machine learning methods can separate fake reviews from genuine ones by revealing text hidden patterns that the human eye cannot recognize. Existing work of fake review detection can be classified according to their detecting an individual spammer, a group of spammers or fake reviews in one, mix and cross-domain [11]. It is worth mentioning that this paper covers various techniques for fake review detection in natural language processing. As such, it is mainly focused on English language reviews, their related problems, their datasets, and their applications.

III. FEATURE EXTRACTION

In this section, we analyse the existing features used in the literature. These features can be classified into two main categories: behavioural features and textual features.

1) BEHAVIOURAL FEATURE

Behavioural features represent the statistical significance of a user's review and behaviour based on his past and current review. For the statistical analysis, there is a need to understand some symbols and formulas. Table 2 displays the symbols utilized in this section. These features can be summarised as follows.

- **MAXIMUM NUMBER OF REVIEWS -F1.**

Most existing studies showed that 75% of spammers created more than five reviews on some specific days [81]–[83]. However, it was also observed that 90% of the normal users have never written more than one review within one day. Thus, the number of reviews created by each user can identify the normal or spammer reviewers. The maximum number of reviews can be computed as follows.

$$F_1(a) = \frac{MaxRev(a)}{max(MaxRev)} \tag{1}$$

- **PERCENTAGE OF POSITIVE REVIEW -F2.**

The high percentage of positive reviews written by a spammer about products may indicate fake reviews [84]. The percentage of positive reviews with the score of 4 and 5 rating to weed out certain reviewers who appeared to encourage businesses can be calculated as follow.

TABLE 2. List of symbols.

Symbols	Descriptions
a	*Author*
r_a	*Review r by an author a*
r	*Review*
MaxRev(a)	*Maximum number of reviews posted by an author 'a'*
R_a	*List of reviews posted by an author 'a'*
F_a, L_a	*Time to the author 'a' first review, Time to the author 'a' last review*
t	*Current review time*
**(r)*	*Review rating*
$Len(r_a)$	*Number of characters in the review*
r_{ij}	*Review Rank order*
$\grave{a}$	*Parameter greater than one, and it shows the decay rate.*
r_i	*Current review*
y	*Total number of reviews posted by the author*
r_{ap}	*The rating is given by an author towards product p*
$\overline{r}_p$	*The average rating is given by all authors than a towards product p*

$$F_2(a) = \frac{\sum_{x=1}^{|R_a|} |\{\star(r_x) \in \{4, 5\}\}|}{|R_a|} \tag{2}$$

- **AVERAGE REVIEW LENGTH -F3.**

Most existing research works showed that spammers do not write detailed reviews about a service or a product, which might help to detect spammers [81]–[83], 85]–[88]. Since the spammers are trying to create fake reviews, they typically spend little time writing their reviews [72]. However, 90% of reliable reviewers write longer reviews with an average length of more than 200 words. The authors [72] classified the reviews as fake, which has less than X length. The average review length can be calculated as follow where X=135.

$$F_3(a) = \begin{cases} 1, & len(r_a) < X \\ 0, & otherwise. \end{cases} \tag{3}$$

- **BUSRSTNESS (BST) -F4.**

The majority of spammers appear to explode the ratings to produce immediate results. In a short time, posting so many reviews is considered an irregular practice and could identify that the user might be a spammer [91]. This method is proposed to analyse the number of writers' reviews created by the user in the previous 24 hours. If the total number of reviewers has crossed a threshold, then the reviewer could be a spammer. The threshold value is set to X = 28 by the experimental dataset studies. The burstiness feature can be computed as follows.

$$F_4(a) = \begin{cases} 1, & \sum_{x=1}^{|Ra|} |\{r_x \in R_a\} \\ & \cap(t_x \text{ is in last } 24 \text{ hours}| > X) \\ 0, & otherwise \end{cases} \tag{4}$$

- **RREVIEWER DEVIATION -F5.**

An impartial reviewer is expected to rate the products based on the average review rating. However, when the spammers try to demote or promote some products, their given ratings can vary considerably from that product's average ratings. Spammers will usually provide a high rating about a product, which may help us detect fake reviews [81], [83], [92].

Therefore, the ranking deviation is a potential activity to identify spam reviewers [94]. This feature can be calculated as follows.

$$F_5\left(a\right) = avg\frac{\left|r_{ap} - \bar{r}_p\right|}{4} \tag{5}$$

- **WEIGHTED RATING DEVIATION -F6.**

Early deviation captures the actions of a spammer who spams an evaluation soon after the product is released. Such spams would certainly draw the attention of other spammers to exploit the views of subsequent spammers [72], [95]. It takes the victimized products to recover from these low early reviews by another legitimate reviewer. Rating weight showing how the rating has been issued early. The weight of the rating can be calculated as follows.

$$F_6\left(a\right) = \frac{1}{\left(r_{ij}\right)^{\grave{a}}} \tag{6}$$

- **RATIO OF NEGATIVE REVIEWS -F7.**

Since calculating the ratio of positive reviews is substantial, the ratio of negative reviews by a reviewer is also significant. In order to determine the percentage of negative reviews with 1 and 2 score ratings, the proposed method filtered those reviewers who were more likely to demote businesses by calculating the percentage of their negative reviews [72]. The ratio of negative reviews is calculated as follow.

$$F_7\left(a\right) = \frac{\sum_{x=1}^{|R_a|}\left|\{\star\left(r_x\right) \in \{1,2\}\}\right|}{|R_a|} \tag{7}$$

- **MAXIMUM CONTENT SIMILARITY-F8.**

Having reviews with similar content about distinct products is a strong indication of a spammer [10], [96], [97]. They usually write the same reviews about various products to support them because they do not spend time creating new fake reviews [98]. Therefore, to detect the author's spamming behaviour, it is essential to find the author's reviews' content similarity. Existing works used cosine similarity to capture the maximum and average similarity between the reviews' contents. The maximum content similarity feature can be calculated as follows.

$$F_8\left(a\right) = \underset{x}{\overset{y}{max}}\ \left[cosine\left(r_i, r_x\right)\right] where\ r_i, r_x \in R_a, x < y \tag{8}$$

- **REPATED REVIEWS.**

For the same product, multiple and repetitive reviews posted by the users is an indication of abnormal behaviour [4]. Though, there is still a need to address the configuration settings of this feature Jindal and Liu [4] suggested that because of internet connectivity problems or operating faults, it might be possible that the same user posted the reviews multiple time should not be treated as a fake review.

- **BOTTOM RANKED REVIEWS RATIO -F9 AND F-10.**

Genuine reviewers would probably rate a product or service after they have encountered it and thus take time compared to spammers who rate early to alter customer decisions [95]. The bottom ranked reviews feature can be calculated as follows.

$$F_{10}\left(a\right) = \frac{\left|\{r : r \in R_a \& F_{BRR}(r) = 1\}\right|}{|R_a|} \tag{9}$$

where F_{BRR} indicates the botttom rank of review and can be calculated:

$$F_{TRR}\left(a\right) = \begin{cases} 1, & r_{ij} \leq \gamma_1 \\ 0, & otherwise \end{cases} \tag{10}$$

where γ_2 is a threshold indicating whether the review is bottom ranked or not.

- **TOP RANKED REVIEWS RATIO -F11 AND F-2.**

As we mentioned, posting early reviews is a sign of fake reviews. If a reviewer has most of their reviews as top-ranking reviews, their behaviour might be considered suspicious [95]. The top ranked review ratio feature can be computed as follows.

$$F_{11}\left(a\right) = \frac{\left|\{r : r \in R_a \& F_{TRR}\left(r\right) = 1\}\right|}{|R_a|} \tag{11}$$

where F_{TRR} indicates the top rank of review and can be computed:

$$F_{TRR}\left(a\right) = \begin{cases} 1, & r_{ij} \leq \gamma_2 \\ 0, & otherwise \end{cases} \tag{12}$$

where γ_2 is a threshold indicating whether the review is top ranked or not.

- **EEXTREME RATING BEHAVIOR -F13.**

Consumers can deliberately glorify or disgrace the products by the highest or lowest ranking scores. Similarly, the spammers attempt to give high or low ratings for praising or damaging some products intentionally [83]. In the five-star rating system, the rating of one star or five stars is known as extreme rating behaviour, which can be calculated as follows.

$$F_{13}\left(a\right) = \begin{cases} 1, & \star\left(r_a\right) \in \{1,5\} \\ 0, & \star\left(r_a\right) \in \{2,3,4\} \end{cases} \tag{13}$$

- **FIRST REVIEW RATIO -F14.**

Early reviews on service and products may have a significant impact on the sales of companies. Spammers are also seeking to become early reviewers to be more influential [101] in order to mislead the buyers. The first review ratio can be computed as follow.

$$F_{14}\left(a\right) = \frac{\sum_{x=1}^{|R_a|}\left|\{r_x \in R_a \cap \left(r_x\ is\ a\ first\ review\right)\right|}{|R_a|} \tag{14}$$

A. TEXT FEATURES

Text features include semantic, grammar, lexicon, and metadata features about the review, which can help identify the fake reviews. There are different types of features related to this category which are further discussed below.

TABLE 3. Comparison of exiting feature extraction methods.

Ref	Algorithm	Features	Dataset	Metric	Results
[4]	Logistic Regression	Behavioral & review	Amazon	AUC	78%
[4]	Logistic Regression	Review	Amazon	AUC	63%
[77]	SVM	Bigram & LIWC	AMT	Accuracy	89.8%
[77]	SVM	Bigrams	AMT	Accuracy	89.6%
[123]	SVM	Unigram & Deep syntax	AMT	Accuracy	91.2%
[91]	SVM	Bigrams & behavioral	Yelp	Accuracy	86.1%
[99]	SVM	N-gram	AMT	Accuracy	86%
[107]	SVM	Stylometric	AMT	F1-measure	84%
[124]	SVM	Unigram	AMT	Accuracy	83.21%
[84]	Naïve Bayes	Behavioral & review	Trip-Advisor	F1-measure	63.1%
[100]	SAGA	Unigram & POS & LIWC	AMT	Accuracy	65%
[46]	Unsupervised multi-iterative graph-based method	Content-based features, Behavior-based features and Relation-based features	AMT dataset	Accuracy	95.3%.
[43]	Logistic Regression	(Doc2vec) and (Node2vec)	Yelp Chi YelpNYC and Yelp Zip	AUC	80.71%. 81.29%. 83.18%.
[16]	CNN	Word2vec and word order.	AMT dataset	Accuracy	70.02%.
[37]	Recurrent Convolutional Neural network	Word2vec	AMT dataset and Deceptive dataset	Accuracy	82.9%. 80.8%.
[3]	Bi-GRU with attention	Integrated features (word embedding) and Discrete features (Unigram & POS & LIWC)	Deceptive dataset	Accuracy	One domain:81.3% for Hotel, 87% for Restaurant and 76.3% for Doctor. Cross-domain 83.7% on Restaurant domain, 57.3% in Doctor domain.
[70]	Combination of long short-term memory and convolutional neural network	Character-level	Spam review	Accuracy	99.5%.
[7]	Fake Generative Adversarial Network (FakeGAN)	Glove2vec	AMT dataset	Accuracy	89.2%.

- **META-DATA.**

These features include actual given reviews and review information such as review ID, feedback, review length, rating, data, reviewer ID, and store ID [84], [97]. Meta-data features have shown themselves to be useful for fake review detection. Unusual or abnormal reviews can be detected using meta-data information. When a reviewer is identified as someone who writes fake reviews, all reviews linked to this reviewer can easily be categorised as fake. However, these features may not be available in many data sources, restricting their usefulness for this fake review detection task. For example, experts can identify certain spammer by the reviewer's identity, such as IP address to detect the location of the reviewer's computer, users review time, feedback given by the reviewer. There are some scenarios that can be analysed:

- Some users sometimes create several negative or positive reviews of a specific product using the same computer; that can be suspicious activity.
- If we analyse the competing product brands, we evaluate the ratings given by a reviewer on a particular product. We can notice that a specific reviewer has created many positive reviews for that particular brand's products. Further, the same reviewer has posted many negative reviews for products of other competing brands.
- The reviewer's location also clearly indicates the significance of the given review. It might be possible to find positive reviews created from locations near to hotel; these feedbacks are not genuine because the hotel reviewers should be at distant locations.

- **PART OF SPEECH (POS).**

The frequency of each POS (Part of Speech) in the text is used as a POS feature. Existing research works related to computational linguistics showed that a certain degree of distinction in POS could be found in various types of texts [102], [103]. However, while POS feature achieves good results in cross-domain, it is not effective in detecting fake reviews when compared to other features, such as BoW [77], [100].

- **BAG OF WORD (BoW).**

These features have been used by several tasks related to natural language processing and are also known as n-gram features. These features represent text as a con-

TABLE 4. Detailed information of public existing datasets in the literature.

Dataset & authors	Data Construction Method	Total Reviews (users)	Domain	Publications (data used by)	Review			Product	Reviewer
					Text	Rating	Image		
Yelp CHI [8]	Filtering algorithm	67,365 (38,063	Restaurants and Hotel	[1],[22], [14],[36], [42],[50], [57],[43], [65],[66], [61]	√	√		√	√
Yelp NYC [72]	Filtering algorithm	359,052 (160,25)	Restaurants	[44],[44], [43]	√	√		√	√
Yelp ZIP [72]	Filtering algorithm	608,598 (260,277)	Restaurants	[41],[44], [43],[75], [76], [78]	√	√		√	√
Yelp Consumer Electronic [79]	Rule-based Technique	18.912	Consumer electronic	[79]	√	√		√	√
Dianping [80]	Filtering algorithm	9,765 (9,067)	Restaurants	[81]	√			√	√
Amazon [4]	Rule-based Technique	5,8 million (2,15 M)	Products, DVD, Music, Book	[27], [5]	√			√	√
Amazon [89]	Rule-based Technique	6,819 (4,811)	Books	[29]	√			√	√
TripAdvisor [90]	Rule-based Technique	2.848	Hotels	[28], [64]	√				
TripAdvisor [93]	Human	3,000	Hotels	[93]	√			√	
Opinions [84]	Human	6,000	Products	[29]	√			√	
TripAdvisor [77]	Amazon Mechanical Turk	800	Hotels	[37],[16], [28].	√			√	
TripAdvisor [99]	Amazon Mechanical Turk	1,600	Hotels	[37],[16], [7], [78].	√			√	
TripAdvisor [100]	Amazon Mechanical Turk	3,032	Hotels, Restaurants, Doctor	[4],[27], [32],[37], [12],[3], [29], [65]	√			√	

tinuous number of words or a single word. BoW features such as unigram, bigram, and trigram (n= 1, 2, and 3), has been used in various fake review detection methods. Features related to BoW gives different results on multiple datasets [4], [77], [84], [99], [100]. For example, it obtained 89.6% accuracy by using AMT datasets, while it achieved poor performance on the Yelp dataset with 67.8% accuracy. This is because real-life reviews have different features compared with reviews collected based on a crowdsourcing platform. However, BoW has a significant limitation which is the scalability (e.g., "this great" and "is this great" have the same vector representation). Moreover, it cannot capture the semantic meaning of reviews.

- **LINGUITSTIC INQUIRE AND WORD COUNT (LIWC).**

This is a popular text analysis software tool that can be used in text to analyse the linguistic features from different aspects [104]–[106]. In fake reviews detection, this is not as effective compared to other features, such as unigram, bigram, and trigram [77], [100]. However, the classification model performance could be improved by integrating n-gram features and LIWC. Specifically, it counts and groups the keyword instances into meaningful psychological dimensions, which further divides into four main categories: spoken, personal, linguistics, and psychological features. However, these features cannot be used for natural language which the framework has not been updated to assist and designed for a specific purpose which is spoken language.

- **STOLYMETRIC.**

These features contain syntactic features or word-based and character-based features [107]. Typically, word-based include average word length and number of upper-case characters, indicating the types of characters and words that the reviewer uses. Syntactic features include features such as the number of punctuations that represent the reviewer's writing style.

- **SEMANTIC FEATURES.**

These features present the concepts or underlying meaning of words. These features build a semantic meaning method for fake review detection [10]. Li *et al.* [6] found that semantic features are better than LIWC, POS and n-gram in cross-domain. Later, Kim *et al.* [78] proposed a technique established with FrameNet-based semantic features showing that the classification performances have improved effectively. However, these features cannot capture the semantic connection between documents and words.

- **WORD EMBEDDING.**

One of the most common and significant feature extraction methods for text data is word embedding. It is a vector representation of words and low dimensional proposed in natural language processing [108]. Integrating Word embedding into neural networks model has achieved state of-the-art performance in natural language processing [109]–[111]. Word embedding differs from the traditional vector space method, in which every word can represent a vector of a fixed length of the vocabulary of the documents in the corpus [112]. Word vectors are produced by learning from their surrounding words using neural network architecture. Moreover, unlike traditional methods, it does not suffer from the curse of dimensionality. The word vectors contain condensed continuous numbers as elements with much smaller lengths than the number of documents in the corpus [113]. The word2vec is usually based on predictive methods. It can be learned by using Continuous Bag-of-Words (CBOW), which can predict the word based on its position in the context, or the Continuous Skip-gram method (CSG), which can predict the nearest words to a given word [114]. For simplicity and computation efficiency, the Skip-gram is performed very well in natural language processing [37]. However, these methods cannot be learned from words with a small number of co-occurrence information. To overcome this limitation, character2vec (C2V) [115] was proposed for unseen words. Recently, Pennington *et al.* [116] proposed Glove methods based on count-based models [117], this algorithm has some limitation; cannot capture the polysemy, require high memory for storage and cannot capture the out of vocabulary words. More recently, Joulin *et al.* [118] proposed FastText embedding, which may learn vectors for character n-gram and in a faster way than word2vec. Inspired by the work on word vectors' representation, the authors [114], [119] proposed an unsupervised Pragraph2vec. It derives sentence vectors for large text such as document, sentence, and paragraph based on the proposed skip-gram algorithm. This method requires the user to train vectors for word groups that often occur. However, this method cannot be used for streaming data because retraining data is needed for unseen word-groups during the testing period [120]. Node2vec are unsupervised learning algorithms proposed by Grover and Leskovec [121] to generate node embedding from the network data. In brief, word embedding has been widely used in natural language processing tasks. Methods such as FastText and Glove achieved more accurate and fast results in the natural language processing tasks [122].

SUMMARY: Feature extraction is a method of extracting features from data. We summarised and analysed the most commonly used features in the fake review detection domain. Using a combination of features to train the classifier has been used and achieved better performance than using a single type of feature [85], [86], [97]. Further, Mukherjee *et al.* [97] found that using combination features (BOW, LIWC, and POS) is better than using BOW alone. In another study, the authors [91] found that using behavioural features is efficient than linguistic features. We can conclude that using behavioural and text features plays a significant role in improving fake reviews detection model performance.

IV. BENCHMARK DATASETS

In this section, we shall summarise the existing datasets of the literature shown in Table 4. We classify these datasets based on the construction methods into four types, as shown in Fig 2:

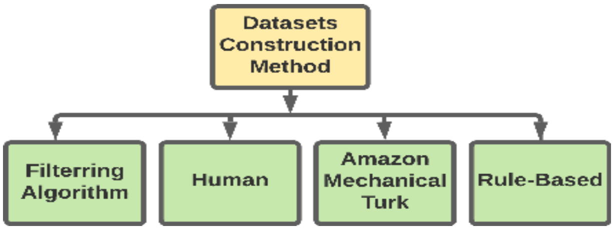

FIGURE 2. Datasets Construction Method Types.

A. FILTRING ALGORITHM METHOD

Yelp CHI is a real-world dataset, from 2004 to 2012, collected by Mukherjee *et al.* [8], which contains 67,365 reviews of both restaurants and hotels in Chicago city. Reviews were labelled as either fake or genuine by the Yelp spam filter. The authors used behavioural features and lexical features to learn classifiers. User behaviour features were explicitly collected by analysing website ads and internal data, such as geographic location information, user IP address, session logs, and network. By following the same method, two more real-world datasets, Yelp NYC and Yelp ZIP, were collected from Yelp.com by Rayana and Akoglu [72] from 2004 to 2015 where Yelp NYC contains 359.052 reviews and Yelp Zip contains 608.598. Similarly, each review was labelled as fake or genuine by the Yelp spam filter. The average length for Yelp datasets is 130.6. Later, Li *et al.* [80] constructed datasets in the Chinese language using a Dianping filtering algorithm where the average review length is 85.5. This dataset includes 9.765. reviews. However, these datasets were built based on an unknown filtering algorithm to label each review as fake or genuine, and these algorithms are not available publicly.

B. HUMAN METHOD

Li *et al.* [84] constructed a dataset using 30 rules written by humans. In particular, three volunteers (undergraduates) were invited to annotate fake reviews. Each student classified a review individually to decide whether the review is fake or not. The majority voting rule, where the individual human judges were partial, was used to predict "fake review," when two out of three human judges claimed that the review was fake. Lastly, they obtained the dataset, which contains 6,000 reviews, where 1,398 reviews were labelled as fake. Similarly, Ren *et al.* [93] constructed a dataset contains 3,000 reviews, where 712 reviews were labelled as fake reviews. However, manual annotation requires a lot of manpower. Moreover, there were several inaccurate labels in this kind of dataset, so artificial recognition accuracy is still very low [99]. Thus, these datasets still have a lot of mislabelled reviews.

C. AMAZON MECHINICAL TURK METHOD (AMT)

The datasets in this section are collected based on crowdsourcing platforms. Crowdsourcing services can carry out massive data collection. It mainly describes the network website's task and pays for anonymous online employees to carry out the task. Human beings cannot exactly differentiate between fake and genuine ones, but they can generate fake reviews as part of the dataset. From their part, Ott *et al.* [77] constructed a dataset containing 800 reviews of Chicago hotels by AMT. They collected 400 genuine reviews and 400 fake reviews from TripAdvisor. In a similar way, Ott *et al.* [99] constructed datasets containing 1,600 reviews, where 800 were labelled as fake reviews. Later, Li *et al.* [100] created datasets using the same method and having 3,032 reviews. However, the distribution of this data is distinct from the distribution of a real-life dataset.

D. RULR-BASED METHOD

Another study conducted by Jindal and Liu [4] constructed a dataset based on the rule-based method from Amazon. They found that three kinds of repeated reviews appear to be fake reviews: various reviewer IDs on the same product, same reviewer IDs on various products, and various reviewer IDs on various products. The authors used the Jaccard distance method to calculate the three types of repeated reviews concerning the review text's similarity. They considered the review is fake if the similarity is greater than 0.9. This dataset contained 5.8 million reviews, where 55,000 were labelled as fake. Using different methods with a defining set of rules, the authors [89] and [90] constructed datasets contain 6,819 and 2.848 review for book and hotel domain, respectively. Later, Barbado *et al.* [79] crawled through review datasets based on the web-scraper process from Yelp.com. They labelled them based on content and user behavioural features, where 9653 reviews were labelled as fake and 20828 genuine reviews. These datasets were annotated depends on the rule-based method. The method based on rules does not rely on manual annotations, and the cost of annotations is relatively low. A large number of annotation data is simple to create, but some noise is present. For example, using the dataset constructed by Jindal and Liu [4], they considered the review as fake if different users posted reviews of the same product, or the same users posted reviews of the same product, or different users posted reviews on different products. Considering these review rules as fake is not reliable because there is a phenomenon where the same consumer has a few assessments with a high likelihood for the same product due to the mismanagement or network link, yet they labelled them as fake. So, this annotation approach needs to be discussed.

V. LITERARTURE REVIEW FOR FAKE REVIEW DETECTION

During the past few decades, a wide range of problems, including face recognition, speech recognition, font recognition, fraud detection, and disease diagnosis, have been addressed by machine learning methods [125]–[131]. In recent years, machine learning has also been investigated to combat spam, an issue that is expanding to various online applications such as SMS, email, and blogs [132]–[135]. Below we provided the detection techniques used for detecting fake reviews with their pros and cons.

A. TRADITIONAL STATISTICAL MACHINE LEARNING IN DETECTING FAKE REVIEWS

Machine learning plays a significant role in detecting fake reviews and is commonly divided into supervised, semi-supervised, and unsupervised learning [4].

1) TRADITIONAL STATISTICAL SUPERVISED LEARNING IN DETECTING FAKE RREVIEWS

Supervised learning techniques are used to predict if reviews are fake or not. This sub-section shall sum up the existing supervised learning techniques in the literature shown in Table 5. For example, Jindal and Liu [4] introduced a supervised learning algorithm to detect fake reviews by studying duplicate reviews. The proposed model consisted of two phases. The first phase used unigram and bigram as features, with Naïve Bayes, Random forest, and support vector machine utilized as a classification algorithm. The second phase used two ensemble methods (stacking and voting) to enhance the classification methods performance. The results on the AMT dataset [77] showed that the ensemble techniques gave better results than the Naïve Bayes random forest and SVM classification algorithms. Using the simple feature and ensemble methods can enhance the accuracy in detecting fake reviews. However, it can be unreliable if duplicate reviews are considered to be fake reviews.

TABLE 5. Summary of supervised traditional statistical machine learning models in One domain, mix domain and cross-domain for fake reviews detection.

Ref	Dataset	Method	Features	Results	Comments
[4]	• The gold standard dataset contains data collected from three domains (Hotel, Restaurant, and Doctor).	• Ensemble machine learning	• Unigram and bigram	• Naïve Bayes Accuracy: 87.1%. • Ensemble techniques Accuracy: 87.68%.	•Considering duplicate reviews as fake is unreliable.
[12]	• The gold standard dataset consists of three domains review data (Hotel, Restaurant, and Doctor)	• Sparse Additive Generative Model (SAGE)	• Linguistic query and word account (LIWC). • Part of Speech (POS). • Unigram.	• Unigram accuracy: 76.1%. • Cross-domain accuracy on restaurant domain using unigram: 77%. • Cross-domain accuracy on restaurant domain using POS: 74.6%. • Cross-domain accuracy on restaurant domain using LIWC: 74.2%. • Cross-domain accuracy on doctor domain using unigram: 52%. • Cross-domain accuracy on doctor domain using POS: 63.4%. • Cross-domain accuracy on doctor domain using LIWC: 64.7%.	•The proposed model failed in extracting the semantic information of the sentence. •The proposed method is efficient in detecting fake reviews in a single domain only.
[29]	• DeRev dataset • OpSpam dataset • Opinion's dataset	• Support vector network (SVN)	• Linguistic query and word account (LIWC). • Word space model (WSM) • Latent Dirichlet Different (LDA).	• One domain, WSM & LDA accuracy: 90.9% on OpSpam dataset, 94.9% on DeRev dataset, 87.5% on Abortion dataset, 87% on Best Friend dataset and 80% on Death Penalty dataset. • Mix Domain, LDA & WSM accuracy: 76.3%. • Cross-domain, WSM and LDA accuracy: 59.3% on DeRev datasets and 64% on Best Friend dataset.	•A deep neural network is probably more appropriate to improve the performance in cross-domain.
[27]	• Dianping Real-life dataset	• Hybrid supervised machine learning	• Behavioral features • Content features.	• Combination features accuracy: 98%. • Behavioral features accuracy: 74%. • Content features accuracy: 69%.	•The findings of this study showed that the behavior of the reviewer is temporal dynamic.
[15]	• Reviews from Yelp.com • The gold standard dataset collected data from three domains (Hotel, Restaurant and Doctor)	• Decision tree method	• Feature selection method	• F-measure: 76.91% on Yelp dataset • Hotel domain F-measure: 78.3%. • Restaurant domain F-measure: 81.8%. • Doctor domain F-measure: 75.0%.	•The performance can be improved by considering the data correlation in selecting the appropriate features. •The proposed model is not good as compared to neural network models.
[42]	• Yelp Chi dataset	• Naive Bayes • Random forest. • JRip. • AdaBoost • J48 classifiers.	• TFIDF • Feature selection.	• AdaBoost accuracy: 73.4.	•It was found that the results were not stable during different configurations. •Compared with traditional machine learning algorithms is not enough to determine the effectiveness of the proposed model.

TABLE 5. ***(Continued.)*** **Summary of supervised traditional statistical machine learning models in One domain, mix domain and cross-domain for fake reviews detection.**

[1]	• Yelp Chi dataset • TripAdvisor collection	• Naïve Bayes. • KNN • Multinomial Naïve Bayes • MDLText • RF • Rocchio • SVM •	• N-gram with TFIDF for text representation	• F-measure on TripAdvisor negative reviews using SVM: 87.3%. • F-measure on TripAdvisor positive: 89.9%. • F-measure on TripAdvisor (negative and positive review): 89.9%. • F-measure on Yelp using MDLText: 71.7%. •	• Model performance dropped over time. • Sentiment polarity affected performance. • Diversity of services and products affected the performance. • MDLText achieved the best performance on the Yelp datasets. • SVM achieved the best performance on the TripAdvisor dataset.
[18]	• Yelp Chi dataset • Semi-real dataset	• Ensemble Learning Model	• TFIDF. • Feature selection •	• Ensemble learning F1 measure: 81.7%. • Ensemble learning F1 measure: 76.1% on an artificial dataset. •	• The chi-Squared feature plays a significant role in improving performance. • Compared with traditional machine learning algorithms is not enough to determine the effectiveness of the proposed model.
[28]	• AMT dataset	• Ensemble learning model	• Unigram & bigram features	• Naïve Bayes accuracy: 87.12%. • Random forest accuracy: 84.87%. • Support Vector Machine accuracy: 83%. • Stacking accuracy: 87.68%. • Voting accuracy: 87.43%.	• The proposed model did not outperform deep learning algorithms. This suggest using deep learning with review embedding to the proposed model can enhance the results.
[38]	• The gold standard dataset consists of three domains (Hotel, Restaurant and Doctor)	• Adaptation model for detecting fake reviews in cross-domain	• Character n-gram	• Cross-domain accuracy on restaurant domain: 79.3%. • Cross-domain accuracy on Doctor domain: 63.8%.	• Difficulty in detecting fake reviews in cross-domain, so deep neural network is probably more appropriate to improve the performance in cross-domain.
[51]	• Yelp Chi. • Yelp NYC. • Yelp ZIP • Yelp Consumer Electronic	• Analysis of concept drift (SVM, LR and PNN)	• TF-IDF	• Accuracy on Yelp Chi: 68.17%. • Accuracy on Yelp ZIP: 91.35%. • Accuracy on Yelp NYC: 84.85%. • Accuracy on Yelp Consumer Electronic: 76.72%.	• They found that the performance dropped significantly over time due to changing the reviews' characteristics over time. • There is a strong relation between concept drift and classification performance which negatively affects the prediction algorithm performance.
[60]	• Dataset collected from Yelp.com	• SVM. • NB. • RF. • MLP	• Lexicon-based method (SentiWordNet)	• RF accuracy:92.9%. • SVM accuracy: 84.9%. • NB accuracy: 73.5%. • MLP accuracy: 83.6%.	• Rating sentiment inconsistency features plays significant roles in improving the performance for fake review detection. • The proposed model works on limited data. • Using word embedding representation with deep learning can enhance the performance.
[66]	• Yelp Chi	• Ensemble model (RF, Xgboost, Lightgbm, Catboost and GBDT).	• Review centric. • Reviewers' centric	• Hotel domain F1-score using stacking: 72.06%. • Hotel domain F1-score using majority voting: 71.51%. • Restaurant domain F1-score using stacking: 79.46%. • Restaurant domain F1-score using majority voting: 78.97%.	• Stacking method performed better majority voting. • The proposed model didn't outperform the state-of-the-art method. • The proposed model suffers from time complexity.

Similarly, Lin *et al.* [12] introduced a classification model to detect fake reviews in a cross-domain environment based on a Sparse Additive Generative Model (SAGE), which is created based on the Bayesian generative model [136]. The model is a combination of a generalized additive model and topic modelling [137]. They used linguistic query and word account (LIWC), POS, and unigram techniques as features to detect fake reviews in cross-domains. The proposed model could capture different aspects such as fake vs. truthful and positive vs. negative. They used the AMT dataset [77] which consisting of three domain reviews (Hotels, Doctors, and Restaurants) to evaluate the proposed model. The experimental results showed that the accuracy of the classification using unigram was 65%. The accuracy of two class classifications (Turker and Employee reviews) using unigram was 76.1%. The accuracy on cross-domain using unigram, POS, and LIWC separately were 77%, 74.6%, and 74.2%, respectively, on the restaurant domain. The accuracy on cross-domain using unigram, POS, and LIWC separately using Doctor domain were: 52%, 63.4%, and 64.7%. However, the proposed model failed in capturing the semantic information of the sentence. In related work, Hernández-Castañeda *et al.* [29] investigated the efficiency of using SVN (Support Vector Network) in classification tasks to detect fake reviews in one, mixed and cross-domains. They used the LIWC, Word space model (WSM), and latent Dirichlet Allocation (LDA) techniques as a feature extraction method. They evaluated the proposed model on three datasets; the DeRev dataset [89], OpSpam dataset [77] and Opinions dataset [138]. The results compared to the previous works [77], [89], [138] showed that a combination of WSM and LDA achieved the best results in one domain with an accuracy of 90.9% on the OpSpam dataset, 94.9% on DeRev dataset, 87.5% on Abortion dataset, 87% on Best Friend dataset and 80% on Death Penalty dataset. There was also an accuracy of 76.3% in a mixed domain compared to the Naïve Bayes classifier. However, the proposed model did not achieve the best results on cross-domain compared to state-of the-art methods. The performance was good in one domain and mix domain and poor in cross-domain because they used the dataset for testing and combined the remaining dataset for training. This suggests that a deep neural network is probably more appropriate to improve fake review detection in a cross-domain by improving the learning presentation.

From their part, Sedighi *et al.* [15] proposed a decision tree method to detect fake reviews. They used traditional feature selection techniques to select suitable features and evaluate them. The proposed model can be improved by taking into account the data correlation in choosing the appropriate features. In the study by.Khurshid *et al.* [42], the authors proposed a supervised machine learning model to detect fake reviews based on content features and primal features. The proposed model used five classifiers to classify the reviews: Naive Bayes, Random forest, JRip, AdaBoost, and J48. The results on a real-life dataset [8], showed that the AdaBoost with combined features performed better than other classifiers with an accuracy of 73.4%. Further, using Primal features has a significant impact on improving performance. However, the proposed model did not perform well with an imbalanced dataset.

Motivated by this, Khurshid *et al.* [18] extended their previous work and proposed an ensemble learning model to detect fake reviews based on selected features. The proposed model consisted of two tiers: Tier 1 used three classifiers (Discriminative Multi-nominal Naive Bayes, a library for Support Vector Machine and J48), and Tier 2 used Logistic Regression classifier to introduce an accurate result. They also used the following feature selections to extract structural and linguistic features: Particle swarm optimization used to explore the feature space, Cuckoo Search used to explore the attribute space, Greedy stepwise, carried out in vector space and Chi-Squared utilized to evaluate the worth of an attribute by calculating the value of Chi-Squared statistic value. They evaluated the proposed model on a real-life dataset [8], and a semi-real dataset [77]. The experimental results showed that the chi-squared feature plays a significant role in improving the proposed model's performance with an 84.1% accuracy on the Yelp restaurant dataset and 81.7% semi-real dataset. However, the proposed model performance could be improved by integrating the chi-squared feature into the deep learning model.

In the study by Cardoso *et al.* [1], the authors performed a comparison analysis of distinctive content-based classification models to investigate if the data characteristics change over time or not. The experimental results on real-world datasets from Yelp [8] showed that the models' performance dropped significantly over time. This is because the spammers continuously tried to avoid the spam filter. Further, in the real-world application, most recent reviews contain features not demonstrated by a model trained with past reviews. Furthermore, they discovered that the performance of the models dropped significantly over-time. Hence, the need for new models that can work with dynamic changes of fake review characteristics over time. Moreover, the performance of the methods was affected by the polarity of the reviews. So, they recommended using a specialised method for each type of polarity. Further, they found that the techniques' performance could be affected by the diversity of products and services. They recommended using a specific model for each type of product and service.

Sánchez-Junquera *et al.* [139] proposed a fake review detection model based on the character n-gram feature. They used a support vector machine and Naïve Bayes as classification algorithms. The proposed model was evaluated on a dataset consists of 'Death penalty', 'Abortion' and 'Best Friend' domains [140]. The experimental results showed that the proposed model performed better than SVM with LIWC, LDA& words, and Deep syntax & words [138] in identifying fake reviews. However, the results were inferior when compared to other methods [29], [123], [141]. This suggests that using a combination feature could improve the classification model performance.

By taking advantage of using the ensemble model, Mani *et al.* [28] introduced a supervised learning model to detect fake reviews based on unigram, and bigram features model contained two phases. In the first, Random forest, Naïve Bayes, and Support Vector Machine were used as classification algorithms. In the second phase, stacking and voting ensemble methods were used to enhance the classification model performance. The experimental results on the gold standard dataset [99] showed that the Naïve Bayes achieved the best accuracy (87.21%) in the first phase. In contrast, the stacking ensemble method performed better than voting with 87.68% accuracy. The proposed model showed the importance of using an ensemble method for detecting fake reviews. However, the proposed model did not outperform deep learning algorithms.

Motivated by previous work, Nilizadeh *et al.* [142] proposed OneReview fake review detection model based on textual and metadata features. OneReview concentrates on separating anomalous changes in business profiles via multiple review websites to discover harmful activity without depending on particular patterns. OneReview used change point analysis techniques on each review from numerous websites. Then, they evaluated the change point to identify any reviews which did not match through the webpages. After classifying them as suspicious with the introduced change point analyser, they used the Random Forest classifier to identify the fake reviews. They evaluated the proposed method on two datasets; Yelp Data Challenge dataset (https://www.yelp.com/dataset) and dataset crawled from TripAdvisor. The experimental results showed that the proposed method performed well in fake reviews detection, with an accuracy of 97% with combining all features and 86% with textual features. However, the change point analyser evaluated time series data in one month; this may create some latency issues between the OneReview classification and the posted review.

Spammers posting reviews in an aggregate way within short periods is called Co-bursting. Based on that, a hybrid supervised machine learning method was proposed by Li *et al.* [81] to identify spammers. They found that reviewers' behaviour is temporal dynamic; for this reason, they proposed a labelled hidden Markov method to identify spamming through single reviewer posting time. Then expanded the technique to multi-hidden Markov to determine posting signals and behaviour with Co-bursting. They introduced the Co-bursting method to assist in detecting spammers. They used Dianping [80] real dataset, though these methods did not use any metrics to evaluate the model.

More recently, Sánchez-Junquera *et al.* [38] proposed an adaptation model for detecting fake reviews in cross-domain. The proposed model frequently used Co-occurring Entropy to find the domain features and then used a mismatch method to mask them. The gold standard dataset results using naïve Bayes classifiers showed that the proposed model had difficulty in detecting fake reviews in cross-domain. While the authors [51] highlighted a concept drift problem in fake reviews where the characteristics of reviews change-over time. The authors utilized two methods, statistical machine learning technique, and benchmark concept drift detection methods, to investigate and prove their argument. The authors utilized two methods, statistical machine learning technique and benchmarking concept drift detection methods to investigate and prove their argument. They tested four real-life Yelp datasets [8], [72], [79] and they found that the classifier performance dropped significantly due to the changing of fake review characteristics over time. Furthermore, they stated a strong relation between concept drift and classification performance which negatively affects the prediction algorithm performance. This study indicates the importance of developing fake review detection models that can handle this issue. On the other side, The authors [60] proposed a framework to investigate the review inconsistency based on different features (content, language, and rating) in fake reviews detection. The extracted features are fed into different machine learning classifiers (SVM, NB, RF, and MLP) in order to identify whether the review is fake or genuine. They collected datasets from Yelp.com to evaluate the proposed model. The experimental results show that the review inconsistency features can boost the performance in fake review detection. However, the proposed model works on limited data, and using word embedding representation with deep learning can enhance the performance. From their part, Yao *et al.* [66] proposed an ensemble fake review detection model based on review content and reviewer features. The author handled the unbalance data by combining the grid search method and resampling by finding the best sampling ratio for each classifier. Then, the extracted features are fed separately to each classifier. Finally, they utilized majority voting and stacking methods to enhance the classification model performance. The experimental results on the Yelp dataset [8] showed that the proposed model did not outperform the state-of-the-art methods. Further, the proposed model suffers from time complexity.

2) TRADITIONAL STATISTICAL UNSUPERVISED LEARNING IN DETECTING FAKE RREVIEWS

Depending on the difficulty of creating accurately labelled datasets, supervised learning is not always appropriate. This sub-section sums up the existing unsupervised learning techniques in the literature shown in Table 6. Unsupervised learning can handle this issue because it does not need labelled data. Lau *et al.* [10] proposed an unsupervised model and introduced a Semantic Language Model (SLM) to detect fake reviews. The proposed model followed the assumption proposed by Jindal and Liu [4] that two duplicate reviews were labelled as fake reviews. The cosine similarity method was used to identify fake reviews and then manually confirmed them. Conversely, the reviews that did not have a cosine similarity above a certain threshold with any other reviews were kept as truthful reviews and not manually reviewed. The dataset from Amazon.com contains 54,618 reviews, of which 6% were labelled as fake. SLM method was used to

TABLE 6. Summary of unsupervised traditional statistical machine learning models in one domain, mix domain and cross-domain for fake reviews detection.

Ref	Dataset	Method	Features	Results	Significant Outlines
[10]	• Amazon reviews	• Unsupervised model and developed SLM	• Considering duplicate reviews as fake	• AUC: 87%	• Outperformed SVM • SLM was effective in detecting fake reviews • Considering duplicate reviews as fake is unreliable.
[15]	• Reviews from Yelp.com • Gold standard dataset consists of three domains (Hotel, Restaurant and Doctor).	• Decision tree method	• Feature selection method	• F-measure: 76.91% on Yelp dataset • Hotel domain F-measure: 78.3%. • Restaurant domain F-measure: 81.8%. • Doctor domain F-measure: 75.0%.	• The performance could be improved by considering the data correlation in selecting the appropriate features.
[30]	• Yelp Chi dataset	• Unsupervised topic sentiment joint probabilistic method.	• Latent Dirichlet allocation (LDA)	• Restaurant F1 measure: 83.92% • Hotel: F1 measure: 85.03%.	• Incorporating behavioural features with LDA can improve the performance. • Integrating the proposed model with work [40] could be improve the performance
[46]	• AMT dataset • Collected dataset from Amazon.com	• Unsupervised multi-iterative graph-based method	• Content-based features • Behavior-based features • Relation-based features	• AMT dataset accuracy: 95.3%. • Crowdsourced dataset accuracy: 93%.	• Combined features improved the performance compared to a single model. • The performance could be enhanced by Integrating network structure with iterative algorithm.
[59]	• Dataset collected from JD.com	• Unsupervised learning	• LDA	• Accuracy: 96.42%.	• Considering duplicate content as fake is unreliable.

give each review a spamming score. The experimental results of the proposed model achieved a 0.9987 AUC score, which outperformed SVM. Further, SLM was effective in detecting fake reviews. However, considering duplicate reviews as fake can be unreliable. Later, Dong *et al.* [30] introduced an unsupervised topic sentiment model to identify fake reviews. The proposed model consisted of four layers: document, topic, sentiment, and word. They enhanced the LDA model, which was used to find topic information from documents to acquire the reviews' topic sentiment [143]. Sentiment and topic features are fed into random forest and support vector machine classifiers. Gibbs sampling algorithm [144] was used to obtain the probabilistic distribution between topics and words as well as sentiment and topics. The results on the real-life dataset from Yelp.com showed that the proposed model with document level was better than other models with features such as POS, LDA, character n-gram, and unigram. However, the proposed model compared only with content-based methods, which was insufficient to determine its effectiveness and ignored the reviewer behaviour features. Motivated by this, Li *et al.* [59] proposed a method to identify a group of fake reviews based on nominated topics. The proposed model consists of three stages; defining the equivalent groups and their target topics, then they used the K-means algorithm to cluster reviews. Finally, they used content duplication and time burstiness to label suspicious group as fake. The experimental results on the dataset collected from JD.com showed the effectiveness of the proposed model. However, considering duplicate content as fake is unreliable. More recently, to extract the semantic meaning from reviews text, Noekhah *et al.* [46] proposed an unsupervised graph-based model in order to detect fake reviews detection by using implicit and explicit features. The experimental results based on the crowdsourced dataset and AMT dataset [77] showed that using combined features enhanced the fake reviews detection model performance. However, the proposed model did not compare with neural network models to show it is effectiveness.

3) TRADITIONAL STATISTICAL SEMI-SUPERVISED LEARNING IN DETECTING FAKE RREVIEWS

Semi-supervised learning is a machine learning method to perform fake review detection using unlabelled data due to the difficulty of obtaining labelled reviews. This sub-section shall sum up the existing semi-supervised learning techniques in the literature shown in Table 7.

Positive and unlabelled learning method has been extensively utilized in text classification and achieved good results [145], [146]. Typically, Yafeng *et al.* [9] proposed a novel Positive and Unlabelled learning method (PU), called mixing population, and individual nature PU learning method to detect fake reviews. Some reliable negative examples were identified from the unlabelled dataset. The integration of latent Dirichlet and K means provided some representative

TABLE 7. Summary of semi-supervised traditional statistical machine learning models in one domain, mix domain and cross-domain for fake reviews detection.

Ref	Dataset	Method	Features	Results	Comments
[9]	• Gold standard dataset	• Novel PU method (MPIP-UL)	• Latent Dirichlet Allocation	• Accuracy:79.2%	• The proposed model outperformed previous PU learning models.
[13]	• Gold standard dataset consists of three domains (Hotel, Restaurant and Doctor).	• Multi-task method (MTL-LLR)	• Unigram and bigram features	• Doctor accuracy: 85.4%. • Hotel accuracy: 88.7%, • Restaurant accuracy: 87.5%.	• The difficulty of detecting fake reviews in cross-domain. This suggests using transfer learning for detecting fake reviews in cross-domain [26].
[35]	• Dataset from JD.com	• PU semi-supervised learning	• Review content features • Metadata features	• Accuracy on 200 test data with 600 training data is 87.6%. • Accuracy on 100 test data with 700 training data is 89.3%.	• The proposed model did not perform well in the short text (less than 20 words).
[43]	• Yelp Chi • Yelp NYC • Yelp Zip	• Semi-Supervised learning framework (SPR2EP)	• Textual content (Doc2vec). • Reviewer items network features (Node2vec).	• AUC on Yelp Chi: 80.71%. • AUC on Yelp NYC: 81.29%. • AUC on Yelp Zip: 83.18%.	• The proposed model works with long text only.
[55]	• Yelp Chi • AMT	• Ramp one-class SVM.	• TF-IDF	• AMT dataset accuracy: 92.3%. • Yelp Chi dataset accuracy: 74.34%.	• The proposed model sensitive to the presence of outliers and noises which can affect the decision boundary of OC-SVM classifier.
[61]	• Yelp CHI	• Semi-supervised (SVM, NB RF, LR, KNN, LDA and DT).	• Review text. • Reviewer features	• Co-training multi fusion features precision:83.97%. • Co-training multi fusion features recall:84.45%. • Co-training multi fusion features F1-score:81.89%.	• Adding more reviewer feature can improve the performance. • Using deep learning to build end to end fake review detection is properly make it robust and accurate.
[68]	• AMT dataset • Yelp dataset	• Investigated the effectiveness of semi-supervised learning method.	• Bigram.	• Co-training method accuracy on the AMT: 88%. • Self-training accuracy on the AMT: 93%. • TSVM accuracy on the AMT: 83%. • Co-training method accuracy on the Yelp: 69%. • Self-training accuracy on the Yelp: 73%. • TSVM accuracy on the Yelp: 64%	• Using the metadata information about the reviews can improve the performance.

positive examples and negative examples. All fake reviews were clustered into distinct groups based on the Dirichlet process mixture model. Then, they mixed two schemes - individual nature and population nature to identify the group label of fake reviews. The final classifier was built using multiple kernel learning. The experimental results showed that the proposed model outperformed previous PU learning models in terms of accuracy Deng *et al.* [35] proposed a PU semi-supervised learning model to detect fake reviews based on review content and metadata features. They used the similarity features of the review (duplicate or near duplicate) to detect fake reviews. They then used the K-Means algorithm to classify the reviews by calculating the percentage of the fake review in each group. They classified each group depending on its threshold value. They labelled the review positively if the review is far away from the trusted negative case. In contrast, the reviews are negative if the reviews are close to the true negative case. They collected the electronic products dataset from JD.com. The results showed that the proposed model performed well in identifying fake reviews with an 88.1% average accuracy. Yet, the proposed model did not perform well with short texts of less than 20 words. Furthermore, they did not compare their model with other models to prove the effectiveness. Another research was conducted by Hai *et al.* [13], who introduced a multi-task method (SMTL-LLR) in order to detect fake reviews. Laplacian Logistic regression (LLR) was used to leverage the unlabelled data and introduced a semi-supervised multi-task method via Laplacian Regularized Logistic Regression (SMTL-LLR). The proposed model improved the learning for the single task by using the knowledge covered inside the training of another similar task. They selected 10,000 unlabelled reviews randomly, containing three domains (Doctor, Hotel, and Restaurant) from the datasets created by Ott *et al.* [77]. The experimental results showed that SMTL-LLR outperforms the state-of-the-art methods [4], [77], [147]–[149] in three domains (Doctor, Hotel, and Restaurant) with an accuracy of 85.4%,88.7%, and 87.5%, respectively. However, the proposed model ignored the reviewer information, which could improve the classification model performance.

Recently, Yilmaz and Durahim [43] introduced a semi-supervised learning framework (SPR2EP) to detect fake reviews based on textual content and reviewer items network features. Two unsupervised learning algorithms (Doc2vec and node2vec) proposed by [119] and [121], respectively, were used. Doc2vec was used to generate document embedding from the review content, while node2vec generate node embedding from the network data. A reviewer item feature was produced by generating a link between items (hotel and restaurant). Once the reviewer creates a review on an item, then running node2vec to learn the vector representation for items and reviewers. Afterwards, these representations are fed to a logistic regression algorithm to classify reviews as spam or not. They evaluated the proposed model on three real-life Yelp datasets [8], [72]. The results showed that the proposed model with combined features outperformed the state of art methods [72], [150] on the three datasets, with an 80.71% AUC, 81.29% AUC, and 83.18% AUC, respectively. Node2vec performed better than Doc2ve. However, the proposed method was not compared with other methods, such as a neural network, to show its effectiveness.

More recently, to handle the scarcity of labelled datasets, a semi-supervised approach called "Ramp One-Class SVM" was utilized [55] in order to identify fake reviews. The experimental results based on the Yelp dataset [8], and AMT dataset [77] showed that the proposed model achieved good results on the AMT dataset with a 92.3% accuracy and 74.37% accuracy on the Yelp dataset. However, the proposed model did not outperform the state-of-the-art methods. In another study, the author [61] proposed a fake review detection model based on combining multiple features, review text, and reviewer features. First, they proposed a method to analyse whether the reviewer's emotion can improve the performance. Second, they used the rolling decision-making method in order to use unlabelled data by collaborating the training data to dynamically update the extracted features. Then they used seven machine learning models (SVM, NB RF, LR, KNN, LDA, and DT) to identify whether the review is fake or genuine. Yelp dataset's experimental results show that the proposed model achieved good performance in terms of precision and recall. More recently, Ligthart *et al.* [68] investigated the effectiveness of several. Semi-supervised methods for fake reviews detection. They used three semi-supervised algorithms, self-training, co-training and Trasnductive SVM (TSVM) algorithm [151]. The experimental results on the AMT datasets and Yelp Chi showed that self-training with naïve Bayes classifier achieved the best performance on both datasets.

SUMMARY: The traditional machine learning methods learn from data with significant predefined features for the prediction values. Further, it is easy to implement, and doesn't require high computational resources. Furthermore, traditional machine learning usually achieves good results with small datasets compared to the deep learning models. However, feature engineering is a challenging task that needs to collect knowledge for the original dataset's feature extraction. Further, It doesn't achieve good results with a large dataset compared to the deep learning model.

B. NEURAL NETWORK IN DETECTING FAKE REVIEWS

Neural network methods provide great results in data classification projects for natural language processing tasks [114], [152]–[155]. Compared to traditional machine learning, most representative neural networks, which are deep learning methods, can quickly extract useful data features. Deep learning can also capture the text's semantic meaning using a word embedding method. For fake review detection, much work has been done using Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM).

1) CONVOLUTIONAL NEURAL NETWORK (CNN) IN DETECTING FAKE REVIEWS

CNN is a specific type of neural network used in the computer vision field. CNN plays a significant role in capturing the local features that are important for the classification of natural language processing tasks. We display a CNN algorithm for fake reviews detection with a simple example, as shown in Fig. 3. First, the word vectors of the input review are split into a matrix. This matrix is fed to the convolutional layer that consists of numerous filters with distinct dimensions. Second, Passing the results from the convolutional layer to the pooling layer. Then, concatenate the pooling results to achieve the final representative vector. The final vector predicts the review label.

In this sub-section, we shall sum up the existing CNN methods in the literature shown in Table 8. Li *et al.* [6] introduced a neural network model to learn document representation in order to detect deceptive spam opinions using Convolutional Neural Networks. The proposed model used the words vector as an input for training and testing. A sentence weights neural network model is introduced

TABLE 8. Summary of CNN models in one domain, mix domain and cross-domain for fake reviews detection.

Ref	Dataset	Methods	Features	Results	Comments
Li, et al. [6]	• Gold standard dataset consists of three domains (Hotel, Restaurant and Doctor).	• Sentence Weight Neural Network	• Word2vec (Skip-gram).	• Accuracy: 79.5 % • Precision: 76.1 % • Recall: 89.8 % • F1: 82.3 %	• The results showed the effectiveness of CNN in cross-domain. • CNN performed better than LSTM in a mixed domain. • Considering sentence and document representation to extract the semantic meaning of text improved the performance. • The proposed model works only with limited datasets.
[16]	• AMT dataset • 1,000 hand-annotated reviews.	• Word Order-Preserving CNN	• Word2vec and word order.	• CNN accuracy: 70.02%.	• CNN is not efficient for long text. • The hand-annotated method requires a lot of manpower.
[22]	• Yelp Chi dataset	• Unsupervised neural network model	• Word2vec (CBOW). • Behavioural Features.	• Hotel domain accuracy:65.4%. • Restaurant domain accuracy: 62%.	• Learning review embedding with encodes behavioral and linguistic features is effective. • They found the contextual information is similar to that of the new reviewers.
[37]	• AMT dataset • Deceptive dataset	• Recurrent Convolutional Neural network and word contexts (DRI-RCNN)	• Word2vec (Skip-gram).	• AMT dataset accuracy: 82.9%. • Deceptive dataset accuracy: 80.8%.	• The proposed model ignored the behavioural features that can boost the performance.
[44]	• Yelp NYC • Yelp Zip	• CNN	• Extracted Real behaviour features. • Pretrained Glove algorithm).	• F1-measure: 85% for normal reviews and 27% for fake reviews.	• They found that the social relation of the user plays a significant role in enhancing the classification performance.
[56]	• Yelp NYC • Yelp Zip	• Unsupervised model	• Extracted Real behaviour features	• Hotel domain: 60% F1 measure. • Restaurant domain: 70%.	• Behaviour representation by dynamic link re-weighting plays a significant role in enhancing the performance. • Ignored the review text features that can enhance the performance.
[64]	• Hotel reviews from TripAdvisor	• local outlier factor (LOF) algorithm.	• Aspect rating. • Review text feature (TF-IDF).	• Accuracy:79.6%. • Precision:79%. • Recall:80.7%. • F1-score:79.8.	• Aspect rating plays a significant role in identifying fake reviews. • Adding more features can boost the performance in fake review detection. • Works only with limited datasets.

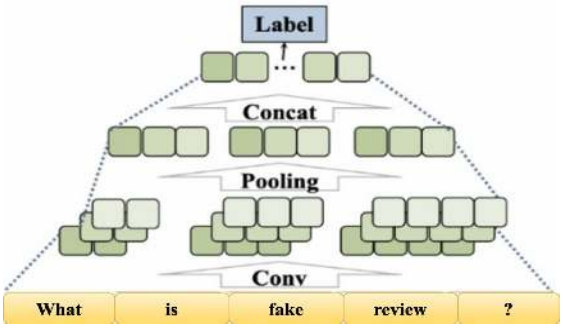

FIGURE 3. The simple architecture of convolutional neural network (CNN).

to represent every sentence and document in the review. The proposed model's architecture includes two convolutional layers: the sentence layer to create a composition of the sentence and the document layer to transform the sentence vector towards a document vector. The proposed model is evaluated based on a dataset used in the study by Li *et al.* [6], which contained a hotel, restaurant, and doctor reviews. The results showed the effectiveness of CNN in cross-domain. Furthermore, in mixed-domain, CNN performed better than LSTM. Motivated by this, Zhao *et al.* [16] introduced a word order-preserving CNN method for detecting fake reviews. They used word 2vec and the word order reserving pooling method rather than the original max pooling to generate a word vector. As an output layer, the obtained features are concatenated from the pooling layer. They evaluated the proposed model on AMT dataset [77] in addition to 10,000 reviews were annotated by using the data annotation method proposed by Li *et al.* [84]. The experimental results showed that the proposed model outperformed the state-of-the-art methods [156]–[158] with an 70.02% accuracy. In contrast, CNN was a more efficient model for classifying short text reviews. CNN had a shorter training time, while RNN was more efficient for long texts [16]. However, the hand-annotated technique requires a lot of manpower.

To enhance the classification model performance, the attention neural network method, introduced by Wang *et al.* [75] to indicate whether a review is behavioural misleading or linguistically misleading, or both. The proposed model used dynamic weight as a form of measure by observing behavioural and linguistic patterns for training. Multi-layer perceptron was used to extract the behavioural features: a CNN to extract linguistic features. Then, the attention method was used to learn the dynamic weight for linguistic and behavioural features. The experimental results on Yelp dataset [8] showed that the proposed model outperformed the state of art methods [8], [76] with an 88.8% accuracy on the Hotel domain and 91% on the Restaurant domain. Furthermore, attention mechanism plays significant role in enhancing the classification model performance. However, the proposed model focused more on linguistic features than behavioural features, which are not enough to identify fake reviews.

Later, an unsupervised neural network model was introduced by.Wang *et al.* [22] to handle the cold-start problem (a new reviewer posts a new review) for fake reviews detection based on behaviour and text features. CNN was used to model the review text, which can catch the complex semantic information that is very extremely difficult to express with traditional features such as unigram and LIWC [3]. The proposed model learned how to distinguish between the reviews by integrating textual information as well as behavioural information. Further, TransE is a method that can code the structure of a graph, representing nodes and edges used to encode behavioural information [159]. The experimental results on the Yelp dataset [8] showed that the proposed model achieved better results than SVM with an accuracy of 65.4% on the hotel domain and a 62% accuracy on the restaurant domain. [8]. However, learning review embedding with encode behavioural and linguistic information is more effective. Yet, the proposed model was compared neither to other embedding methods nor to dimension reduction methods. More recently, Zhang *et al.* [37] introduced the DRI-RCNN identification model for fake reviews by utilizing a recurrent convolutional neural network with word contexts. The proposed model consists of four layers; a convolutional layer implemented to train the overall vector towards representing a word; a recurrent neural layer implemented to learn right and left for a fake and real context vector of a word. The proposed model was evaluated on two datasets (AMT and Deception dataset). The results showed that the proposed model achieved the best results with 82.9% accuracy on AMT datasets compared to the state of art methods such as LIWC and unigram with SVM [77], LIWC feature and four n-grams with SVM [140], recurrent convolutional neural network [160], profile alignment compatibility method [161], sparse additive generative model [100], lexicalized production rules with SVM [123], convolutional neural network and gated recurrent neural network [154]. Furthermore, the proposed model performed well with 80.8% accuracy on deceptive dataset compared to the state of art method [77]. However, the proposed model suffers from time complexity.

An embedding method to influence the user reviewer behaviour and social relations to handle the cold start problem in fake reviews detection was proposed by Li *et al.* [44]. The proposed model jointly embedded the user-item social relations and user behaviour into an inferable user item review rating representation. The proposed model consists of four parts: item embedding layers, rating embedding layers, review embedding networks, and user embedding layers. They embedded a co-occurrence-based user behaviour by maximizing the success rate of existing behaviour under a designated measure. They also embedded user/item social relation according to the context information generated by random walks in the user-item network produced by reviewing activities. CNN was used for text embedding by using CBOW.

The proposed model was evaluated based on Yelp NYC and Yelp Zip dataset [72]. The proposed model achieved better results than SVM with linguistic features and behavioural features [8]. Similarly, Li *et al.* [56] extended their previous work and proposed an unsupervised model to address the cold start problem in fake reviews detection. Instead of reviewing content and social relations between users with other existing users, they considered behaviour representation by dynamic links re-weighting. The proposed model was evaluated based on Yelp NYC and Yelp Zip datasets of [72]. The proposed model achieved poor results with a 60% F1 score on the hotel domain and a 70% F1 score on the restaurant domain. However, the proposed model did not outperform the state-of the-art method and ignored the review text features that could boost the classification model performance. More recently, the authors [64] proposed an aspect-rating local outlier factor in order to identify fake reviews. They considered fake review detection as outlier detection. First, they utilize the lexicon-based method to compute the aspect rating of the review. Then tensor factorization method was used for completeness. After that, the local outlier factor (LOF) algorithm was used to classify the reviews. The experimental results on a dataset from TripAdvisor.com show that aspect rating improved the performance for fake review detection. However, integrating more reviewer's features can boost performance.

2) RECURRENT NEURAL NETWORK (RNN) IN DETECTING FAKE REVIEWS

RNN is utilized for processing sequential data that use internal memory for the input sequence. Theoretically, RNN can save information for long sequences. However, in practice, RNN can only run for a few steps due to exploding gradient or vanishing gradient problems. Consequently, the researchers have developed new models to overcome the limitations of

TABLE 9. Summary of RNN models in one domain, mix domain and cross-domain for fake reviews detection.

Ref	Dataset	Method	Features	Results	Comments
[3]	• Gold standard dataset	• Bi-GRU with attention	• Integrated features (word embedding) • Discrete features (Unigram & POS & LIWC)	• One domain accuracy:81.3% for Hotel, 87% for Restaurant and 76.3% for Doctor. • Cross-domain accuracy: 83.7% on Restaurant domain, 57.3% in Doctor domain.	• Outperform the state-of-the-art methods. • The proposed model requires high computational resources.
Wang , et al. [24]	• Reviews moblil01.com in Taiwan	• LSTM	• Dictionary	• LSTM accuracy: 89.4%.	• Long short-term memory algorithm detected deceptive reviews more effectively than Support Vector Machine. • Didn't compare with neural network methods.
[32]	• Gold standard dataset	• Bidirectional LSTM	• Part-of-speech and First -Person Pronoun features. • Glove.	• On cross-domain restaurant accuracy: 81.3%, doctor accuracy: 66.8%. • On mix domain accuracy: 83.9%. • On each domain hotel: 83.9%, restaurant domain: 85.8%, doctor domain: 83.8%.	• First-person feature plays a significant role in identifying fake review. • The proposed model requires high computational resources.
[41]	• Deceptive Spam Corpus dataset • Four-City dataset • Yelp Zip dataset • large movie dataset • Drug dataset	• Hierarchical CNN-GRN deep learning and Multi instant learning methods	• Pretrained word2vec	• Deceptive Spam Corpus dataset accuracy: MIL: 90.1%, CNN-GRU: 91.9%. • Four-City dataset accuracy: MIL: 82.8%, CNN-GRU: 84.7%. • Yelp Zip dataset accuracy: MIL: 64.6%, CNN-GRU: 66.4%. • Large movie dataset accuracy: MIL: 87.1%, CNN-GRU: 88.9%. • Drug dataset accuracy: MIL: 78.2%, CNN-GRU: 83.8%.	• Outperformed the classical CNN and RNN algorithms. • The proposed model works only with short text. • Adding metadata feature to the proposed model can boost the classification model performance.
[50]	• Gold standard dataset	• Bidirectional LSTM with a self-attention mechanism	• Pretrained word embedding on Wikipedia corpus	• One domain accuracy: 85.7%, 84.7% and 85.5% on Hotel, Doctor, and restaurant domains, respectively. • Mix domain accuracy: 83.4%. • Cross-domain accuracy: 71.6% on restaurant domain and 60.5% on doctor domain.	• They found that fake reviews expressed stronger emotions than real reviews. • The model failed to achieve good results in cross-domain. This suggests using domain adaption methods [62], [67] can boost the performance in cross-domain.
[70]	• Spam email. • Spam review. • political statements	• Combination of long short-term memory and convolutional neural network	• Character-level	• Binary test accuracy of 99.5%.	• Transfer learning is a promising technique for detecting inauthentic behaviour of products. • The proposed model used a straightforward method such as n-gram.

RNN, such as Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional LSTM, Stacked LSTM, and LSTM with attention method. We display an RNN algorithm for fake reviews detection with a simple example, as shown in Fig. 4. Firstly, a particular vector using a word embedding technology is defined for each input word. The built-in word vectors are then fed one by one into RNN cells. The RNN cells with the input vector are fed into the next hidden layer. The RNN has the same weight of each input word and shares the parameters between different parts. Lastly, the reviews label can be predicted by the last hidden layer output. In this subsection, we shall sum up the existing RNN techniques in the literature shown in Table 9.

Ren and Zhang [3] utilized a gated recurrent neural network model to learn document representation for detecting fake reviews. CNN was used to construct sentence representation from word representation that gave the best results for sentiment analysis. In contrast, gated RNN was used with the attention method to produce document representation as a feature for fake review detection [162]. They used datasets from the study by Li *et al.* [6] that consisted of three domains (doctor, hotel, restaurant reviews). In one domain,

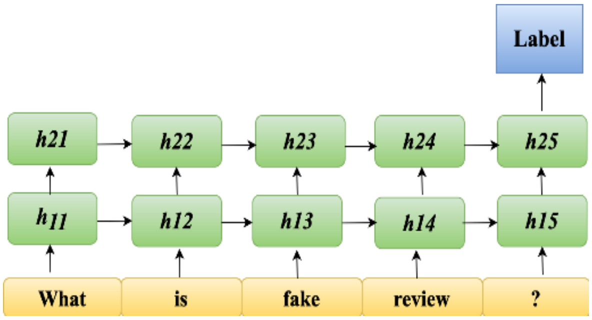

FIGURE 4. The simple architecture of recurrent neural network (RNN).

the results showed improvements in the hotel and restaurant domains over the doctor domain due to the high percentage of unidentified vocabulary. To overcome this issue, they created a discrete method using logistic regression with the same features from the study of [100] with neural features. Then, they concatenated the neural features to the discrete feature before the SoftMax layer. The results in one domain were 81.3% for Hotel, 87% for restaurant domain and 76.3% for doctor domain. The results in cross-domain, when the classifier was trained on hotel reviews, were 83.7% in the restaurant domain and 57.3% in doctor domain. The proposed model outperformed the state of art method [100] and neural network methods, such as RNN, CNN, GRNN, and Bi-directional GRNN in one and cross-domain. Furthermore, integrating more features could be improved the classification model performance. However, the proposed model suffers from time complexity.

Later, Wang *et al.* [24] utilized a long short-term memory recurrent neural network to detect spammers based on the dictionary. They produced a multilayer perceptron consisting of three layers: an input layer, which received data as a neuron, an LSTM layer, the hidden layer for dimension reduction, and an output layer for one neuron. The neuron's value determines if the reviewer is a regular (0) or spammer (1). They collected the dataset from the product reviews webpage and moblil01.com in Taiwan. They annotated the data based on internal confidential documents. The proposed model discovered that the long short-term memory detected deceptive reviews more effectively than SVM with an 89.4% accuracy. Moreover, LSTM is considered better than RNN due to long-term memory. However, the proposed model did not compare with other neural network methods. Further, the proposed model focused on text-only and ignored the behavioural and metadata feature that can improve performance.

To overcome the RNN limitation, Liu *et al.* [32] introduced the bidirectional LSTM model to learn the reviews' document level representation to detect fake reviews based on combined features. The features are added to the proposed model by merging feature representation (POS), first -person pronoun features, and document representation (word embedding (Glove)). The model was evaluated based on the AMT dataset [6], which contained three domains (doctor, hotel, and restaurant). The experimental results showed that the proposed model outperformed the state of art methods such as paragraph average, SWNN, SWNN+POS+I, BiLSTM, and basic CNN+POS+I. In mixed domain, the proposed model outperformed the state of art methods (SWNN, Deep CNN, CNN-LSTM, and CLSTM) with an 83.9% accuracy. Finally, the results in one domain outperformed the state of art methods with an 83.9% accuracy on the hotel domain, 85.8% accuracy on the restaurant domain, and an 83.8% accuracy on the doctor domain. Based on the model results, we can notice that the first-person pronouns feature plays a significant role in identifying deceptive reviews. However, the proposed model requires high computational resources.

Recently, Jain *et al.* [41] proposed hierarchical CNN-GRN deep learning methods, and Multi instant learning (MIL) methods were proposed to handle the variable lengths of reviews in fake reviews detection. A three-layer CNN was utilized to extract localized n-gram features. In contrast, GRN was employed to learn semantic dependencies between the extracted features from CNN. In several instances, the input text is divided into multiple instances, and the last instance is discarded if the word length is less than fifteen. The proposed model was evaluated on multiple benchmark datasets, including four-city dataset [163], Yelp Zip dataset [72], Deceptive Spam Corpus [77], Drug Review dataset [164] and Large Movie Review dataset [165]. The experimental results showed that MIL and CNN-GRN performed better than classical CNN and RNN on all datasets. However, the proposed model works well with short text only.

More recently, Zeng *et al.* [50] introduced an ensemble model to detect fake reviews dependent on the review structure. They found the followings; fake reviews expressed stronger emotions than genuine reviews. The first and last sentences contained stronger emotions than the middle sentence, and fake reviews start or end with similar sentences. The proposed model consists of four separate bidirectional LSTM to encode the middle, beginning, and end of the review. Then the concatenation of four representations is used to detect fake reviews. The self-attention method was utilized to integrate the three-local representation into one representation. In contrast, the attention method was applied to incorporate the two representations into the final representation. The results on the AMT dataset [100] showed that the proposed model achieved better results than other methods such as SWNN and SAGA in one domain (Hotel, Doctor, and restaurant) with an 85.7%, 84.7%, and 85.5% accuracy, respectively. Furthermore, in the mixed domain. It achieved 83.4% accuracy. However, the proposed model failed to achieve good results in cross-domain with 71.6% on the restaurant domain and 60.5% on the doctor domain.

Dhamani *et al.* [70] introduced neural network and transfer learning to tackle social media disinformation. They proposed an ensemble method combined with long short-term memory and character-level convolutional neural network. The proposed method showed the ability to transfer knowl-

TABLE 10. Summary of GAN models in One domain, mix domain and cross-domain for fake reviews detection.

Ref	Dataset	Method	Features	Results	Comments
[7]	• AMT dataset	• Fake Generative Adversarial Network (FakeGAN)	• Glove2vec	• Accuracy: 89.2%.	• Didn't outperform the state-of-the-art methods. • GAN is not good enough for text classification due the stability of GAN that makes hyper-tuning challenging task. This suggests using conditional GAN with better hyper tunning.
[14]	• Yelp Chi	• GAN	• Word2vec (CBOW). • Extracted attribute and behavioural Features.	• Hotel domain accuracy: 80% • Restaurant domain accuracy: 75.6%.	• The review embedding representation can be enhanced by exploiting more available information.
[25]	• Yelp Chi	• Behaviour features generative adversarial network	• Extracted Real behaviour features: • Extracted easily accessible features.	• Hotel domain accuracy: 83%. • Restaurant domain accuracy: 75.7%.	• The proposed model didn't perform well on restaurant domain. This suggests building specific model for each domain.

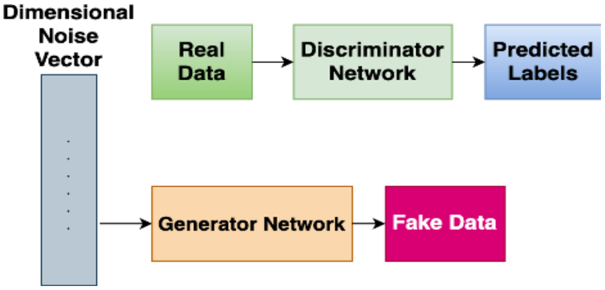

FIGURE 5. The simple architecture of Generative Adversarial Network (GAN).

edge from labelled data in one domain to another domain. Moreover, transfer learning is a promising technique for detecting the inauthentic behaviour of products. However, the proposed model used a more straightforward method, such as n-grams.

3) GENERATIVE ADVERSARIAL NETWORK (GAN) IN DETECTING FAKE REVIEWS

GAN has been achieving remarkable efficiency in various research fields, such as image processing [166]. GAN's framework consists of two models: a generator that produces reviews and a discriminator that estimates the probability that a review is genuine rather than fake. We display a simple architecture of the GAN algorithm in Fig. 5. Firstly, the generator produces a data sample and a discriminator classifying the data as real (training) or false (produced by a generator). The generator aims to produce samples similar to the true data to delude the discriminator. The purpose of the discriminator is to distinguish all types of data samples correctly. In this sub-section, we shall sum up the existing GAN methods in the literature shown in Table 10.

Aghakhani *et al.* [7] proposed a semi-supervised Fake Generative Adversarial Network (FakeGAN) model to handle the scarcity of dataset for fake reviews detection. A model for generating samples with the same distribution consists of one generative and one discrimination model [167]. Two discriminators were proposed to solve the generator's convergence and create a much stronger generator. The first one distinguishes between fake reviews and truthful reviews. The second one separates between samples from fake reviews distribution and reviews generated by the LSTM generative model. Maximum likelihood estimation is utilized for training the generator on fake reviews. The results on the AMT dataset [77] showed that the proposed model achieved an 89.2% accuracy. However, the proposed model did not outperform the state-of the-art method. Furthermore, the result indicates that using GAN is not good enough for text classification due to GAN's stability, making hyper-tuning is a challenging task for the proposed model. Similarly, You *et al.* [14] utilized deep learning techniques for incorporating inherent attributes from various domains to handle the cold start problem in fake reviews detection. They proposed a model that encoded items, reviewers, and reviews, along with their attributes such as date, price ranges, and location. Furthermore, to adapt the knowledge from one domain to another, they proposed a domain classifier. The proposed method included three layers: the first layer integrated various attributes into the model, the second layer captured the three relations such as (entity-entity), (entity-attribute), and (attribute-attribute), and the third layer implemented a domain classifier to capture the domain correlation. The proposed model is evaluated based on the Yelp Chi dataset created by Mukherjee *et al.* [8]. The proposed model achieved better results than SVM, with an 80% accuracy on the hotel domain and 75.6% on the restaurant domain. The generative adversarial network is instrumental in handling cold start problems. However, the proposed model was not compared to other embedding methods.

Later, Tang *et al.* [25] proposed a generative adversarial network model to handle the cold start problem in fake reviews detection. The synthetic behaviour features are generated for new users with no features. Firstly, six real features were extracted for regular users and three kinds of easily accessible features that already existed for new users and regular users. As they took easily accessible features as an input, synthetic behaviour features are generated using a GAN generator. The generator of GAN consists of six layers. The first three layers were used for normalization purposes and get easily accessible features. The other three layers were used to transform the readily accessible features into synthetic behaviour features. After that, the generator is trained using GAN's discriminator and applied to the new user to get synthetic behaviour features. The proposed model was evaluated on the Yelp Chi dataset [8] that contain two different domains (Hotel and Restaurant). The proposed model outperformed the state-of-the-art methods [8], [22], [14] with an accuracy 83% on hotel domain and 75.7% on restaurant domain. Further, combined features improved the classification model performance. However, the proposed model failed in detecting fake reviews in cross-domain.

4) OTHER NEURAL NETWORK METHODS

In this sub-section, we shall sum up the other neural network models in the literature shown in Table 11.

Instead of relying strongly on expert knowledge to recognize fake reviews with a new perspective, Wang *et al.* [76] introduced a new spam detection model based on the relations between reviewers and products. They constructed a 3-mode tensor based on the relations generated from two entities; tensor factorization algorithms called RESCAL [168] were utilized to learn the vector representation of product and reviewers automatically. Lastly, the final concatenated representation of the review is fed into a support vector machine classifier. The Yelp Chi dataset [8] was used to evaluated the proposed model. The results showed that the proposed model outperformed the state-of-the-art method [8], [72] with an 85.9% accuracy on the hotel domain and 87.8% on the restaurant domain. The proposed model showed that the relations between reviewer and product are critical to enhance the classification model performance. To detect a single fake review, Wang *et al.* [17] introduced a multi-dimensional time series model. A unique index was introduced to determine reviewers' credibility by taking trustworthiness and expertise together. The ranking method was introduced to summarize all spammers in various dimensions to detect abnormal time series aspects. When a single fake review happens in the time series, the window size time is reduced. The results showed that the proposed model was useful with human assessment compared to the average RHR on the datasets, consisting of 408,469 reviews from different websites. They discovered that many reviews were posted at the same time on different days between 2009-2010. However, the proposed model did not use metrics like recall, precision, F1 measure, and accuracy to evaluate the method. Heydari *et al.* [5] introduced a pattern recognition method to identify deceptive reviews fallen in suspicious periods based on metadata and rating deviation features. A time series is constructed to determine oscillations in several reviews for each product. A sliding window is utilized to capture the suspicious periods and find the patterns. The results on real datasets from Amazon.com showed that the proposed method performed well in fake review detection with an 86% F-measure. Despite the advantages of the proposed method, it focused on suspicious periods rather than reducing expensive computations in the scoring phase. Furthermore, Hand annotated techniques need many human resources, and adding metadata such as IP address can boost the proposed model's performance.

Later, Li *et al.* [27] introduced a sentence weighted neural network model (SWNN) for review representation to detect fake reviews. The proposed model converted the sentence into a document vector; every sentence is linked to the weight. A sentence consisted of distinct reviewer's words. They then added POS and First-Person Pronoun features to determine if the review was fake or not. The proposed model is evaluated on the AMT dataset [6], which contained Hotel, Restaurant, and Doctor domains. The results showed that the unigram feature achieved the best results on the restaurant domain with an accuracy of 78.5%. In comparison, combined features achieved the best results on the doctor domain with an accuracy of 61.5%. On the mixed domain, SWNN outperformed the state-of-the-art methods [119], [169] with an accuracy of 80.1%. On one domain, the F1 score is used as a metric that yielded the following results: 83.7% on the hotel domain,87.6% on the restaurant domain and 82.9% on the doctor domain. However, it was not able to predict exact results in mix and cross-domains. In order to detect a single fake review, group of reviewers, and reviewer simultaneously, Noekhah *et al.* [45] introduced an unsupervised Multi-Iteration Network Structure based on behavioural and structural features. The proposed model used the inter-relationship (relationships among reviewers) and intra-relationships (the relationship between product, reviewers, and reviews) as feature extraction. The results on the dataset from Amazon.com showed that the proposed model achieved a 98% accuracy with combined features, 74% accuracy with behavioural features, and 69% accuracy with structural features. However, they did not compare it with other methods to show the effectiveness of the proposed model. They did not use all the metadata features, which can improve the classification model performance.

Recently, Yuan *et al.* [57] introduced a hierarchical fusion attention network to learn representations from product and user level for fake reviews detection. User-product multi attention unit is introduced to extract the user-product features from the sentence representation. Then, fusion attention units and orthogonal decomposition were applied to learn the user-product representation. Lastly, they defined the reviews as relations between product and user. They used TransH; a model used to embed a knowledge graph in vectors to encode the product-review-user relationship [170]. The proposed model was evaluated based on the Mobile01 Review dataset [171], and Yelp datasets [72]. The results showed that the proposed model outperformed

TABLE 11. Summary of other neural network models in One domain, Mix domain and cross-domain for fake reviews detection.

Ref	Dataset	Method	Features	Results	Comments
[5]	• Real datasets from Amazon.com	• Pattern recognition method	• Product feature and rating behavior feature	• F1-measure: 86%.	• It focused on suspicious periods rather than reducing the problem of expensive computations in the scoring phase. • Hand annotated techniques need a lot of human resources. • Adding metadata such as IP addresses could boost the performance of the proposed model.
[17]	• Reviews from different websites	• Multi-dimensional time series	• -	• -	• They found that many reviews were posted at the same time on different days. • The proposed model did not use metrics like recall, precision, F measure and accuracy to evaluate the method.
[27]	• Consists of three domains (hotel, restaurant, and doctor)	• Sentence weighted neural network Model (SWNN)	• Behavioral features • Content features.	• Hotel domain accuracy: 83.7%. • Restaurant domain accuracy: 87.6%. • Doctor domain accuracy: 82.9%.	• The proposed model did not perform well in cross-domain. • Utilizing soft alignment attention method to the proposed model could be enhanced the results.
[22]	• Yelp Chi	• Spam detection model	• Word2vec (CBOW). • Behavioral Features.	• Hotel domain accuracy:65.4% • Restaurant domain accuracy: 62%.	• Relations between reviewer and product are significant to enhance performance.
[45]	• Reviews from Amazon.com	• Unsupervised Multi-Iteration Network	• Unsupervised Multi-Iteration Network Structure	• Combination features accuracy: 98%. • Behavioral features accuracy: 74%. • Content features: 69%.	• Adding all the metadata features can boost the performance.
[57]	• Mobil01_first post dataset. • Mobil01_Reply dataset. • Yelp Chi • Yelp NYC • Yelp Zip	• Hierarchical fusion attention network	• User-product features • Pre-trained embedding (300-dimensions).	• 86.96% F1 measure on Mobil01_first post dataset • 48.37% F1 measure on Mobil01_Reply dataset. • 83.24% AUC on Yelp Chi • 84.78% AUC on Yelp NYC • 87.28% AUC on Yelp Zip	• The product level and user level are critical in detecting fake reviews. • The proposed model could capture the coarse-grained feature.
[65]	• Gold standard dataset • Yelp Chi	• Deceptive reviews detection framework	• LDA • Word2vec	• Accuracy on Yelp Chi: 84.5%. • Accuracy on hotel: 85.9%. • Accuracy on restaurant: 81.5%. • Accuracy on Doctor: 82.7%.	• Combination of fine-grained and coarse-grained features. • Course-grained features improved the performance better than fine-grained features. • The proposed model suffers from time complexity compared to a single model.
[69]	• Gold standard dataset	• Hybrid deep learning models	• Paragraph vector method • BOW	• Accuracy: 92.5%. • F1-measue: 92.4%.	• Ignored some other features such as emotional aspects that can improve the performance.
[71]	• Twitter dataset and Weibo dataset	• Deep graph neural network	• Occasional relations and stable relations	• Twitter dataset accuracy: 93.95%. • Weibo dataset accuracy: 90.74%	• The proposed model lost some useful information during the training.

the state of art methods such as: SVM with content and behavioural features [97], [171], Graph-based model (RSD) [150], SpEagle [72], Tensor decomposition model (TDSD) [76], Couple Hidden Markov model (CHMM) [81], Spam2Vec [172], CNN-GRNN [3], Sentence weight neural network (SWNN) [27], Attention-based neural network

(ABNN) [75], AEDA [14]. It achieved an 86.96% F1 score on the Mobil01_first post dataset, with a 48.37% F1 score on the Mobil01_Reply dataset, 83.24% AUC on the Yelp Chi, 84.78% AUC on the Yelp NYC, and 87.28% AUC on the Yelp Zip. The proposed model showed that the product level and user level are critical in fake review detection.

More recently, Cao *et al.* [65] introduced a deceptive reviews detection framework based on combination fine-grained and coarse features to implicit the semantic information from reviews. The extract features were learned with a coarse-grained concatenation of 2- neural network layer and Latent Dirichlet Allocation (LDA). The fine-grained features were learned parallelly by using deep learning techniques. Lastly, a combination of features is used to train a support vector machine algorithm for classifying whether the review is genuine or not. The results on a gold standard dataset [100] and a real-life dataset [8] showed that the proposed model could improve the performance in one and mix domain. Moreover, LDA with Text CNN achieved the best results on the two datasets in one and mix domain. Further, using coarse-grained features improved the performance better than fine-grained features. However, the proposed model suffers from time complexity compared to the single model. Similarly, a hybrid deep learning model was proposed [69] to capture the semantic meaning of reviews to identify fake reviews. The proposed model consisted of three phases; First, they utilized two neural network architectures (Paragraph Vector Distributed Bag of Words and the Denoising Autoencoder) to extract the review embedding. Then, the feature representation embedding from the two models is concatenated and fed to a fully connected layer to determine whether the review is fake or genuine. The results on a gold standard dataset [77] showed that the proposed model outperformed the state-of-the-art methods with 92.5% accuracy. However, adding other features, such as an emotional aspect, could improve the performance.

From their part, Guo *et al.* [71] proposed graph neural network method to identify spammer by jointly embedding the occasional relations and stable relations. The parametric random walk method [173] was used to extract the occasional relations, while a direct vectorized encoding method was used to model the stable relation. Graph deep learning was developed to model the features of interaction. The experimental results on two real-world datasets showed that the proposed model outperformed the baseline approaches such as CNN, MLP, SVM and LSTM.

SUMMARY: Neural networks are one of the most effective machine learning methods. For this area of research, deep learning is used and achieved significant outcomes. Further, there is no need for feature extraction for deep learning; these can be extracted directly from the input dataset without any learned knowledge or interventions. However, these models also have some limitations when applied in fake review detection. One of the main problems associated with deep learning algorithms is that deep learning models don't provide a comprehensive understanding of learning. Deep learning can be considered a ''black box'' model that does not have declarative knowledge to explain the outcomes. Moreover, it requires a large amount of data compared with traditional machine learning, which means we cannot use deep learning algorithms with a small dataset. Furthermore, deep learning models require extremely computational resources.

VI. EXPERIMENTS

In this section, a first-hand evaluation of the performances of seven promising deep learning algorithms on two datasets is presented. These algorithms are character-level convolutional -LSTM, convolutional -LSTM, HAN, convolutional HAN, BERT, DistilBERT, and RoBERTa. The main goal is investigating to what extent such algorithms are able to detect fake reviews. Note that, some of these algorithms have been used by researchers in different domains [174]–[179]. However, as of yet, they have not been used in the fake review detection field. Therefore, this study demonstrated the efficiency of such algorithms in detecting fake review, which can pose as a baseline for further research. For the initial experiments in this study, we used two datasets. The first dataset is the ''**Yelp Consumer Electronic dataset''** [79] that crawled through review datasets based on the web-scraper process from Yelp.com. They labelled them based on content and user behavioural features. This dataset was annotated based on the rule-based method. For example, the dataset was constructed on some rules that considered the review as a fake if different/ same users posted reviews of the different/ same product. This dataset presents a real-life dataset which is preferred as this will help the researchers to build a fake review detection model that can be used efficiently in the real world. A second dataset is the ''**deception dataset''**[100] constructed from TripAdvisor and Amazon Mechanical Turk websites from Chicago city, which contains 3,032 reviews from different domains (Hotel, Restaurant, and Doctor) by crowdsourcing platform. This dataset has extensively used in literature, and it is semi-real dataset [3], [4], [12], [27], [29], [32], [37], [65]. For simplicity, we combined these three-domain reviews at current stages, and we leave the investigation of each domain separately (i.e., multi-domain detection model) for future work. As can be seen from the previous section, to design a fake review detection model, the following steps are performed.

A. DATASET PRE-PROCESSING

During this phase, datasets were pre-processed in order to eliminate the noise, such as stop words, URLs, emojis, etc. The pre-processing has been carried out with the NLTK toolkit,[1] an open-source library commonly used. First, we used tokenization to divide the text into a list of tokens; then, we removed the stop words that cause noise in text classification. Finally, we used the stemming method to reduces the words to their root. Table 12 shows the information of reviews in the deception dataset and yelp consumer electronic

[1] https://www.nltk.org/

TABLE 12. The basic information of reviews in Yelp dataset and deception dataset.

Dataset	Subject	Category	Number of reviews	Number of unique words	Number of sentences
Deception dataset	Doctor	Fake reviews	356	5128	2369
		Genuine reviews	200	5098	1151
	Hotel	Fake reviews	1080	16,635	8463
		Genuine reviews	1080	17,328	9258
	Restaurant	Fake reviews	201	5136	1827
		Genuine reviews	201	5126	1892
Yelp Consumer Electronic dataset	Restaurant review	Fake reviews	9653	38916	30481
		Genuine reviews	20828		

datasets. For simplicity, we combined the three-domains reviews in the deception dataset.

B. FEATURE EXTRACTION

Feature extraction is an essential part of getting the most accurate and useful information from the given data to improve performance and results. For Neural network models, we used pre-trained GloVe embedding methods with 100-dimensions [116]. The GloVe is an unsupervised learning method trained on large datasets of one billion words used for obtaining vector representation of words and have achieved good results in fake review detection, as we mentioned earlier. GloVe is very straightforward and used to enforce the word vectors to capture sub-linear relationships in the vector space. Thus, it proves to perform better than Word2vec in the word analogy tasks. Moreover, Glove adds some more practical meaning into word vectors by considering the relationships between word pairs rather than words. Furthermore, Glove gives lower weight for highly frequent word pairs, so to avoid meaningless stop words like "the", "an" will not dominate the training progress.

C. ALGORITHMS

In this section, we describe the neural network model and transformers used in our experiments.

1) C-LSTM

The C-LSTM extracts a sequence of higher-level phrase representations using CNN and feeds the sequence into a long short-term memory recurrent neural network (LSTM) to obtain the sentence representation [174]. For each word in a given sentence, the convolutional layer applies a matrix-vector function. LSTM propagates historical information over the neural network chain. In our work, first, CNN is built to learn the higher representation of n-grams on top of the pre-trained word vector. Then, the feature maps CNN that are designed as sequential window features to serve as the input of LSTM in order to learn sequential correlations from high sequence representations. This turns each sentence into a succession window (n-gram) features to activate factors in sentences. In our work, we used one LSTM layer and one convolutional layer with 128 filters. Then, we fed it into LSTM architecture with dropout 0.2 and 100 output dimensions. Finally, we used the sigmoid function for the output layer to classify the review as fake or genuine.

2) CHARACTER LEVEL C-LSTM

A sequence of encoded characters is recognised as input in this model [175]. The encoding is achieved by prescribing for the input language, a fixed length of alphabets, and measuring each character with one-hot encoding. Then, the characters are converted into vectors with a fixed length. The quantification order for the character is reversed, allowing the last character reading near the beginning of the output, making it easy to compare weights with the latest reading for completely connected layers. In our work, we designed a character-level embedding layer by retrieving characters from the review dataset. One layer of convolutional units was followed by two layers of convolutional filters 3 and 5. We used two max-pooling and a dropout of 0.2. Then, we created bi-LSTM with fully connected layers and ReLU function. Finally, we used the sigmoid function for the output layer with an ADAM optimizer.

3) HIERARCHAL ATTENTION NETWORK (HAN)

Hierarchical Attention Network (HAN) is an algorithm proposed to capture the whole document structure. The HAN model consists of hierarchal structures (word encoder, word attention, sentence encoder, and sentence attention) using Bidirectional GRU [176]. In our work, we set the maximum length to 200, then, Bi-GRU with 100 output dimensions was fed to the attention layer. We utilized the word encoder as input to generate sentence encoder time distributed layer. Lastly, we used an ADAM optimizer with a 0.001 learning rate to optimize our model.

4) CONVOLUTIONAL HAN

In this model, in addition to HAN architecture, we included a 1-dimensional convolution layer before each two-way GRU layer in HAN to extract high-level input features. This layer takes the feature of the text review before being fed to the attention layer. Similarly to HAN architecture, we set the maximum length to 200, then, Bi-GRU with 100 output dimensions was fed to the attention layer. Further, we used an ADAM optimizer with a 0.001 learning rate to optimize our model.

5) BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT)

BERT is a transformer-based pre-trained model to pre-train deep bidirectional representations from the unlabelled text by learning right and left word context [177]. BERT is pretrained on English Wikipedia text paragraphs of 2500 million words and books corpus with 800 million words. In contrast to the directional model, which read the text sequentially from right to left or left to right, BERT read the entire sequence of words at once, which allows the model to learn the context of the word based on its surroundings (right and left of the word). In our work, we used the BERT model consisted of 12 layered transformer blocks, where each block contained 12 self-attention layers and 768 hidden layers. One sentence at a time was fed into the model. The input sentences were divided into tokens and mapped with the BERT library as input IDs. At the beginning and end of each sentence, both the Classification Token and SEP (separate Segment Token) were added. A fixed-length input mask of 0 was applied, indicating padded tokens, and 1 shows unpadded tokens. The token embedding lists were given to each transformer, and a feature vector of the same length was generated at the output. The CLS output on the 12th transformer layer containing prediction probability vector transformations was used as a combined sequence representation from which classification was made.

6) DistilBERT

DistilBERT is a light version of BERT [178] proposed to mitigate BERT limitations, such as computational complexity, fixed input length size and word piece embedding problem. DistilBERT has the same architecture as BERT, but with additional steps as the number of layers is reduced, token type embedding, and the pooler are removed. In our work, we used the DistillBERT model consisted of 6 layered transformer blocks, where each block contained 12 self-attention layers and 768 hidden layers. We tokenized the input texts and convert the tokens into input IDS. Then, we padded and fed the input IDs into the DistilBERT model for a binary classification task.

7) ROBUSTLY OPTIMIZED BERT APPROACH (RoBERTa)

RoBERTa is an extended version of BERT that can exceed the BERT transformer's performance [179] by training the model longer, training on longer sequences, and removing the next sentence prediction. In addition to English Wikipedia and books corpus, RoBERTa is pre-trained on one more dataset; Common Crawl News datasets containing 63 million news articles in the English language. Similarly, in this research, to encode the inputs in tokens and designate them as input ids, the RoBERTa tokenizer was used. These IDs have been padded to a fixed length to prevent row variation. The characteristics of the sentence pair classification were then extracted from the tokens.

D. RESULTS AND DISCUSSION

In this section, we specifically discuss the performance analyses of deep learning models and transformers architectures. To do these experiments, we used the same parameters according to the original proposed architecture. We divided each dataset into training, validation and testing to perform the experiments. Based on these predefined parameters, evaluate these algorithms performance in fake review detection in terms of performance accuracy, precision, recall, and F1-score as described in Table 13.

As it can be noticed that RoBERTa achieved the best performance for both datasets compared with peer algorithms where it obtained 70.2%, 65%, 61%, and 61.5% for accuracy, precision, recall, and F1-score, respectively. It also achieved 91.02%, 92.5%, 90%, and 90.5% for accuracy, precision, recall, and F1-score, respectively, on the deception dataset. Interestingly, its performance on the deception dataset is much better than Yelp datasets. This is because fake reviews on the Yelp website are more realistic (70.2% accuracy), and fake review detection is more challenging with such sort of dataset where there is overlapping between legitimate and fake review data. In contrast, the deception dataset is representing semi-real data. BERT, another transformer model, also achieved a considerable performance for both datasets. As such, it can be concluded from such results that transformer models are much better in detecting fake reviews, and

TABLE 13. Performance of neural network models and transformers.

	Datasets							
	Yelp Consumer Electronic				**Deception**			
Model	**Acc**	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**
C-LSTM	68.2%	50%	50%	56%	58%	58%	51%	46%
HAN	64.8%	56%	55%	54%	75.1%	74.5%	75.5%	75%
Convolutional HAN	65.9%	58.5%	56.5%	57%	68.1%	67.5%	66.5%	66.5%
Char-level C-LSTM	65%	56.5%	54.5%	54%	79.4%	79%	79.5%	79%
BERT	70%	**65%**	57%	56%	86.2%	88.5%	84.5%	85.5%
DistilBERT	68.8%	62%	57.5%	57%	83.2%	85%	81.5%	82%
RoBERTa	**70.2%**	64%	**61%**	**61.5%**	**91.02%**	**92.5%**	**90%**	**90.5%**

this is because they are trained on large datasets. This could be a good starting point for utilizing such models and developing new ones in the future to improve fake review detection.

On the other hand, deep learning algorithms such as C-LSTM, HAN, convolutional HAN, and char-level C-LSTM have showed poor performances. This can be explained in two folds: First, such algorithms need a huge amount of data to learn and achieve a good performance. In our experiments, both datasets have few thousands of reviews that may not be sufficient to learn the boundary between legitimate and fake reviews. The second reason is such algorithms need extensive parameters tunning process to obtain better results. In our experiments, we used the predefined parameters of such models in literature which may not appropriate for fake review data. This study also provided deep analysis for improving these algorithms' performance in the future to enhance the fake review detection accuracy.

VII. CHALLENGES

In recent years, much work has been done to increase the reliability of online contents. Despite the progress that has been made, there are still challenges that need to be addressed. In this section, we highlight the current gaps in this research area and the possible future directions.

1) **Group of spammers detection.** Literature indicates that identification group of spammers is a significant part of fake reviews detection (Mukherjee, *et al.*, 2012). The plenty of group of spammers leads to the propagation of fake reviews in specific real-time intervals. Therefore, by considering the studies that focus on burst patterns to detect fake reviews, they discovered high accuracy in fake reviews detection. The study of burst patterns using new techniques to detect spammers needs more investigation for future research.
2) **Explainable Fake Review Detection Model**. Deep learning performed a significant role in natural language processing with excellent outcomes. However, it is considered as ''Black Box'' that does not have declarative knowledge for further explanations of the outcomes. All of the deep learning models for fake review detection are not interpretable. Due to this, it is difficult to trust the model performance and results. For instance, what do some of the deep learning models underperform other models on one dataset but outperform another dataset? What deep learning models learn? Interpretability can be conducted by relying on fundamental theories. So far, there has been no research to explain the fake review detection model. Based on that, there is a need for explainable fake review detection models [180].
3) **Handling Concept drift problem**. Existing methods may not be appropriate for fake reviews detection in the real-world application where the reviews' features changed over time regarding the dynamic nature of the reviews [51]. Furthermore, the prediction model needs to be updated frequently in real world applications [51]. So, there is a need for an efficient model that can handle the concept drift problem in the real-world scenarios.
4) **One class classification model.** One class classification method can provide a solution to deal with unlabeled datasets in a real-world application. For example, one class condition, Randoms field, has been applied to Twitter datasets for anomaly information analysis [181]. Other one-class classification algorithms can be used, such as one-class support vector machine (OSVM) [182], and Non-OSVM models [183]–[185], which can deal with unlabeled real-world data. There is a need to study for unlabeled fake review dataset to solve the lack of dataset issue in fake review detection.
5) **Cross domain fake review detection.** The cross-domain problem needs to be effectively addressed. The issue of lack of annotation datasets is disappointing in fake review detection. Applying a model trained in the source domain and tested in the target domains is a significant research direction. The existing literature focused only on one domain of fake review detection, so these proposed models failed when training in the domain and tested in other domains. For example, the authors [80] trained the model in the domain and tested the other domain. The experimental results show the performance has significantly dropped compared with the performance in the same domain. More research and investigation are needed for cross-domain fake reviews detection [67], [186].
6) **Multilingual fake reviews detection.** Fake review detection is turned into the multilingual analysis. Users can post a review in a different language, such as English, Chinese, Malay or Arabic. So far, few studies have used fake review datasets from different languages [97], [187]. As spammer write quickly, and they copy text from another dataset. The spammer can also use a language translation tool to convert the English review to any other language. So, there is still a need to address this issue for detecting multilingual fake reviews.

VIII. CONCLUSION

This paper presented an extensive survey of the most notable works to date on machine learning-based fake review detection. Firstly, we have reviewed the feature extraction approaches used by many researchers. Then, we detailed the existing datasets with their construction methods. Then, we outlined some traditional machine learning models and neural network models applied for fake review detection with summary tables. Traditional statistical machine learning enhances text classification model performance by improving the feature extraction and classifier design. In contrast, deep learning improves performance by enhancing the presentation learning method, algorithm's structure and additional knowledge. We also provided a comparative analysis of some neural network model-based deep learning and transformers that have not been used in fake review detection. The outcomes showed that RoBERTa achieved the highest accuracy on both datasets. Further, recall, precision, and F1 score proved the efficacy of using RoBERTa in detecting fake reviews. Finally, we summarised the current gaps in this research area and the possible future direction to get robust outcomes in this domain.

We can conclude that most of the existing works focused on supervised machine learning to detect fake reviews. However, supervised machine learning needs a labelled dataset to predict whether the review is fake or not, which can be hard to obtain in a fake review detection area. According to the difficulty of obtaining labelled dataset, we observed that the most commonly used datasets in the current works are constructed based on a crowdsourcing framework. Evaluating the machine learning techniques on these datasets is not preferred as these datasets do not present the fake review in a real-world application. Consequently, assessing the classifiers on the real-world application is preferred as this will help us developing algorithms that can work efficiently in the real world.

We believe this survey will be valuable for researchers with a comprehensive understanding of this field's key aspects. It elucidates the most notable advances and sheds some light on expected future directions.

REFERENCES

[1] E. F. Cardoso, R. M. Silva, and T. A. Almeida, ''Towards automatic filtering of fake reviews,'' *Neurocomputing*, vol. 309, pp. 106–116, Oct. 2018.

[2] L. Da Xu, W. He, and S. Li, ''Internet of Things in industries: A survey,'' *IEEE Trans. Ind. Informat.*, vol. 10, no. 4, pp. 2233–2243, Nov. 2014.

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

[4] N. Jindal and B. Liu, ''Opinion spam and analysis,'' in *Proc. Int. Conf. Web Search Web Data Mining (WSDM)*, 2008, pp. 219–230.

[5] A. Heydari, M. Tavakoli, and N. Salim, ''Detection of fake opinions using time series,'' *Expert Syst. Appl.*, vol. 58, pp. 83–92, Oct. 2016.

[6] 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*. Nanjing, China: Springer, 2015, pp. 393–404.

[7] 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.

[8] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, ''Fake review detection: Classification and analysis of real and pseudo reviews,'' Univ. Illinois Chicago, Chicago, IL, USA, Tech. Rep. UIC-CS-03-2013, 2013.

[9] R. Yafeng, J. Donghong, Z. Hongbin, and Y. Lan, ''Deceptive reviews detection based on positive and unlabeled learning,'' *J. Comput. Res. Develop.*, vol. 52, no. 3, p. 639, 2015.

[10] R. Y. K. Lau, S. Y. Liao, R. C.-W. Kwok, K. Xu, Y. Xia, and Y. Li, ''Text mining and probabilistic language modeling for online review spam detection,'' *ACM Trans. Manage. Inf. Syst.*, vol. 2, no. 4, pp. 1–30, Dec. 2011.

[11] 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, May 2015.

[12] Y. Lin, T. Zhu, X. Wang, J. Zhang, and A. Zhou, ''Towards online review spam detection,'' in *Proc. 23rd Int. Conf. World Wide Web (WWW Companion)*, 2014, pp. 341–342.

[13] 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.

[14] 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. 1884–1895.

[15] 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.

[16] S. Zhao, Z. Xu, L. Liu, and M. Guo, ''Towards accurate deceptive opinion spam detection based on word order-preserving CNN,'' 2017, *arXiv:1711.09181*. [Online]. Available: http://arxiv.org/abs/1711.09181

[17] Y. Wang, S. C. F. Chan, H. V. Leong, G. Ngai, and N. Au, ''Multi-dimension reviewer credibility quantification across diverse travel communities,'' *Knowl. Inf. Syst.*, vol. 49, no. 3, pp. 1071–1096, Dec. 2016.

[18] 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, pp. 387–394, 2018.

[19] B. Pang and L. Lee, ''Opinion mining and sentiment analysis,'' *Comput. Linguist*, vol. 35, pp. 311–312, Mar. 2009.

[20] E. Fitzpatrick, J. Bachenko, and T. Fornaciari, ''Automatic detection of verbal deception,'' *Synth. Lectures Hum. Lang. Technol.*, vol. 8, no. 3, pp. 1–119, Sep. 2015.

[21] B. Liu, ''Sentiment analysis and opinion mining,'' *Synthesis Lectures Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, 2012.

[22] X. Wang, K. Liu, and J. Zhao, ''Handling cold-start problem in review spam detection by jointly embedding texts and behaviors,'' in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 366–376.

[23] N. A. Patel and R. Patel, ''A survey on fake review detection using machine learning techniques,'' in *Proc. 4th Int. Conf. Comput. Commun. Autom. (ICCCA)*, Dec. 2018, pp. 1–6.

[24] 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.

[25] 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.

[26] Z. Li, H. Yao, and F. Ma, ''Learning with small data,'' in *Proc. 13th Int. Conf. Web Search Data Mining*, Jan. 2020, pp. 3539–3540.

[27] 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.

[28] S. Mani, S. Kumari, A. Jain, and P. Kumar, ''Spam review detection using ensemble machine learning,'' in *Proc. Int. Conf. Mach. Learn. Data Mining Pattern Recognit.* New York, NY, USA: Springer, 2018, pp. 198–209.

[29] Á. Hernández-Castañeda, H. Calvo, A. Gelbukh, and J. J. G. Flores, ''Cross-domain deception detection using support vector networks,'' *Soft Comput.*, vol. 21, no. 3, pp. 585–595, Feb. 2017.

[30] 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.

[31] U. Aslam, M. Jayabalan, H. Ilyas, and A. Suhail, ''A survey on opinion spam detection methods,'' *Int. J. Sci. Technol. Res.*, vol. 8, no. 9, pp. 1–10, 2019.

[32] W. Liu, W. Jing, and Y. Li, ''Incorporating feature representation into BiLSTM for deceptive review detection,'' *Computing*, vol. 102, pp. 701–715, Nov. 2019.

[33] R. Xu, Y. Xia, K.-F. Wong, and W. Li, ''Opinion annotation in on-line Chinese product reviews,'' in *LREC*, vol. 8, 2008, pp. 26–30.

[34] A. K Samha, Y. Li, and J. Zhang, ''Aspect-based opinion extraction from customer reviews,'' 2014, *arXiv:1404.1982*. [Online]. Available: http://arxiv.org/abs/1404.1982

[35] 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. Appl. IEEE Int. Conf. Ubiquitous Comput. Commun. (ISPA/IUCC)*, Dec. 2017, pp. 1278–1280.

[36] X. Tang, T. Qian, and Z. You, ''Generating behavior features for cold-start spam review detection,'' in *Proc. Int. Conf. Database Syst. Adv. Appl.* Chiang Mai, Thailand: Springer, 2019, pp. 324–328.

[37] 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.

[38] J. Sánchez-Junquera, L. Villaseñor-Pineda, M. Montes-y-Gómez, P. Rosso, and E. Stamatatos, "Masking domain-specific information for cross-domain deception detection," *Pattern Recognit. Lett.*, vol. 135, pp. 122–130, Jul. 2020.

[39] C. Visani, N. Jadeja, and M. Modi, "A study on different machine learning techniques for spam review detection," in *Proc. Int. Conf. Energy, Commun., Data Anal. Soft Comput. (ICECDS)*, Aug. 2017, pp. 676–679.

[40] Z. Zhao, C. Li, Y. Zhang, J. Z. Huang, J. Luo, S. Feng, and J. Fan, "Identifying and analyzing popular phrases multi-dimensionally in social media data," *Int. J. Data Warehousing Mining*, vol. 11, no. 3, pp. 98–112, Jul. 2015.

[41] N. Jain, A. Kumar, S. Singh, C. Singh, and S. Tripathi, "Deceptive reviews detection using deep learning techniques," in *Proc. Int. Conf. Appl. Natural Lang. Inf. Syst.* Salford, U.K.: Springer, 2019, pp. 79–91.

[42] 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.

[43] 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.

[44] 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.

[45] S. Noekhah, N. B. Salim, and N. H. Zakaria, "A novel model for opinion spam detection based on multi-iteration network structure," *Adv. Sci. Lett.*, vol. 24, no. 2, pp. 1437–1442, Feb. 2018.

[46] 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.

[47] N. N. Ho-Dac, S. J. Carson, and W. L. Moore, "The effects of positive and negative online customer reviews: Do brand strength and category maturity matter?" *J. Marketing*, vol. 77, no. 6, pp. 37–53, Nov. 2013.

[48] F. Zhu and X. Zhang, "Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics," *J. Marketing*, vol. 74, no. 2, pp. 133–148, Mar. 2010.

[49] J. C. Rodrigues, J. T. Rodrigues, V. L. K. Gonsalves, A. U. Naik, P. Shetgaonkar, and S. Aswale, "Machine & deep learning techniques for detection of fake reviews: A survey," in *Proc. Int. Conf. Emerg. Trends Inf. Technol. Eng. (ic-ETITE)*, Feb. 2020, pp. 1–8.

[50] 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.

[51] 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.

[52] I. Peñalver-Martinez, F. Garcia-Sanchez, R. Valencia-Garcia, M. Á. Rodríguez-García, V. Moreno, A. Fraga, and J. L. Sánchez-Cervantes, "Feature-based opinion mining through ontologies," *Expert Syst. Appl.*, vol. 41, no. 13, pp. 5995–6008, Oct. 2014.

[53] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Natural Language Processing and Text Mining*. London, U.K.: Springer, 2007, pp. 9–28.

[54] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2004, pp. 168–177.

[55] 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.

[56] Q. Li, Q. Wu, C. Zhu, J. Zhang, and W. Zhao, "Unsupervised user behavior representation for fraud review detection with cold-start problem," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*. Springer, 2019, pp. 222–236.

[57] 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.

[58] D. U. Vidanagama, T. P. Silva, and A. S. Karunananda, "Deceptive consumer review detection: A survey," *Artif. Intell. Rev.*, vol. 53, no. 2, pp. 1323–1352, Feb. 2020.

[59] J. Li, P. Lv, W. Xiao, L. Yang, and P. Zhang, "Exploring groups of opinion spam using sentiment analysis guided by nominated topics," *Expert Syst. Appl.*, vol. 171, Jun. 2021, Art. no. 114585.

[60] 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.

[61] J. Wang, H. Kan, F. Meng, Q. Mu, G. Shi, and X. Xiao, "Fake review detection based on multiple feature fusion and rolling collaborative training," *IEEE Access*, vol. 8, pp. 182625–182639, 2020.

[62] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[63] 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.

[64] L. You, Q. Peng, Z. Xiong, D. He, M. Qiu, and X. Zhang, "Integrating aspect analysis and local outlier factor for intelligent review spam detection," *Future Gener. Comput. Syst.*, vol. 102, pp. 163–172, Jan. 2020.

[65] N. Cao, S. Ji, D. K. W. Chiu, M. He, and X. Sun, "A deceptive review detection framework: Combination of coarse and fine-grained features," *Expert Syst. Appl.*, vol. 156, Oct. 2020, Art. no. 113465.

[66] 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.

[67] M. Peng, Q. Zhang, Y.-G. Jiang, and X.-J. Huang, "Cross-domain sentiment classification with target domain specific information," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 2505–2513.

[68] A. Ligthart, C. Catal, and B. Tekinerdogan, "Analyzing the effectiveness of semi-supervised learning approaches for opinion spam classification," *Appl. Soft Comput.*, vol. 101, Mar. 2021, Art. no. 107023.

[69] A. Fahfouh, J. Riffi, M. Adnane Mahraz, A. Yahyaouy, and H. Tairi, "PV-DAE: A hybrid model for deceptive opinion spam based on neural network architectures," *Expert Syst. Appl.*, vol. 157, Nov. 2020, Art. no. 113517.

[70] N. Dhamani, P. Azunre, J. L. Gleason, C. Corcoran, G. Honke, S. Kramer, and J. Morgan, "Using deep networks and transfer learning to address disinformation," 2019, *arXiv:1905.10412*. [Online]. Available: http://arxiv.org/abs/1905.10412

[71] Z. Guo, L. Tang, T. Guo, K. Yu, M. Alazab, and A. Shalaginov, "Deep graph neural network-based spammer detection under the perspective of heterogeneous cyberspace," *Future Gener. Comput. Syst.*, vol. 117, pp. 205–218, Apr. 2021.

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

[73] 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, p. 23, Dec. 2015.

[74] Y. Ren and D. Ji, "Learning to detect deceptive opinion spam: A survey," *IEEE Access*, vol. 7, pp. 42934–42945, 2019.

[75] X. Wang, K. Liu, and J. Zhao, "Detecting deceptive review spam via attention-based neural networks," in *Proc. Nat. CCF Conf. Natural Lang. Process. Chin. Comput.* Dalian, China: Springer, 2017, pp. 866–876.

[76] 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.

[77] 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: Human Lang. Technol.*, vol. 1, 2011, pp. 309–319.

[78] S. Kim, H. Chang, S. Lee, M. Yu, and J. Kang, "Deep semantic frame-based deceptive opinion spam analysis," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2015, pp. 1131–1140.

[79] 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.

[80] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao, "Spotting fake reviews via collective positive-unlabeled learning," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2014, pp. 899–904.

[81] 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.

[82] Y. Li, "Highlighting the fake reviews in review sequence with the suspicious contents and behaviours," *J. Inf. Comput.al Sci.*, vol. 12, no. 4, pp. 1615–1627, Mar. 2015.

[83] 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.

[84] F. H. Li, M. Huang, Y. Yang, and X. Zhu, "Learning to identify review spam," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 1–6.

[85] N. Jindal and B. Liu, "Review spam detection," in *Proc. 16th Int. Conf. World Wide Web*, 2007, pp. 1189–1190.

[86] 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. AAAI Conf. Weblogs Social Media*, 2013, pp. 1–10.

[87] M. Jiang, P. Cui, and C. Faloutsos, "Suspicious behavior detection: Current trends and future directions," *IEEE Intell. Syst.*, vol. 31, no. 1, pp. 31–39, Jan. 2016.

[88] S. P. Algur, N. Ayachit, and J. G. Biradar, "Exponential distribution model for review spam detection," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 3, 2017.

[89] T. Fornaciari and M. Poesio, "Identifying fake Amazon reviews as learning from crowds," in *Proc. 14th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2014, pp. 279–287.

[90] A. S. Abu Hammad, "An approach for detecting spam in arabic opinion reviews," in *An Approach for Detecting Spam in Arabic Opinion Reviews*. Amman, Jordan: The International Arab Journal of Information Technology, 2013.

[91] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," in *Proc. 21st Int. Conf. World Wide Web (WWW)*, 2012, pp. 191–200.

[92] 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.

[93] Y. Ren, D. Ji, and H. Zhang, "Positive unlabeled learning for deceptive reviews detection," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 488–498.

[94] 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.

[95] 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.

[96] 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.

[97] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "What yelp fake review filter might be doing?" in *Proc. 7th Int. AAAI Conf. Weblogs Social Media*, 2013, pp. 1–10.

[98] 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, 2018.

[99] M. Ott, C. Cardie, and J. T. Hancock, "Negative deceptive opinion spam," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2013, pp. 497–501.

[100] 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.

[101] 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.

[102] P. Rayson, A. Wilson, and G. Leech, "Grammatical word class variation within the British National Corpus sampler," in *New Frontiers Corpus Research*. Amsterdam, The Netherlands: Brill | Rodopi, 2002, pp. 295–306.

[103] D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan, *Longman Grammar of Spoken and Written English*. London, U.K.: Longman, 2000.

[104] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *J. Artif. Intell. Res.*, vol. 30, pp. 457–500, Nov. 2007.

[105] W. L. Cade, B. A. Lehman, and A. Olney, "An exploration of off topic conversation," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2010, pp. 669–672.

[106] 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.

[107] 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.

[108] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Feb. 2003.

[109] W. Y. Zou, R. Socher, D. Cer, and C. D. Manning, "Bilingual word embeddings for phrase-based machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 1393–1398.

[110] A. Bordes, S. Chopra, and J. Weston, "Question answering with subgraph embeddings," 2014, *arXiv:1406.3676*. [Online]. Available: http://arxiv.org/abs/1406.3676

[111] D. Weiss, C. Alberti, M. Collins, and S. Petrov, "Structured training for neural network transition-based parsing," 2015, *arXiv:1506.06158*. [Online]. Available: http://arxiv.org/abs/1506.06158

[112] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975.

[113] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 160–167.

[114] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[115] W. Ling, T. Luís, L. Marujo, R. Fernandez Astudillo, S. Amir, C. Dyer, A. W. Black, and I. Trancoso, "Finding function in form: Compositional character models for open vocabulary word representation," 2015, *arXiv:1508.02096*. [Online]. Available: http://arxiv.org/abs/1508.02096

[116] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[117] M. Baroni, G. Dinu, and G. Kruszewski, "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2014, pp. 238–247.

[118] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," 2016, *arXiv:1607.01759*. [Online]. Available: http://arxiv.org/abs/1607.01759

[119] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.

[120] S. Achsas and E. H. Nfaoui, "Language representation learning models: A comparative study," in *Proc. 13th Int. Conf. Intell. Syst.: Theories Appl.*, 2020, pp. 1–7.

[121] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 855–864.

[122] F. Almeida and G. Xexéo, "Word embeddings: A survey," 2019, *arXiv:1901.09069*. [Online]. Available: http://arxiv.org/abs/1901.09069

[123] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2012, pp. 171–175.

[124] Y. Ren, L. Yin, and D. Ji, "Deceptive reviews detection based on language structure and sentiment polarity," *J. Frontiers Comput. Sci. Technol.*, vol. 8, no. 3, pp. 313–320, 2014.

[125] D. Tao, Y. Guo, Y. Li, and X. Gao, "Tensor rank preserving discriminant analysis for facial recognition," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 325–334, Jan. 2018.

[126] X. Zhang, X. Liu, and Z. J. Wang, "Evaluation of a set of new ORF kernel functions of SVM for speech recognition," *Eng. Appl. Artif. Intell.*, vol. 26, no. 10, pp. 2574–2580, Nov. 2013.

[127] D. Tao, X. Lin, L. Jin, and X. Li, "Principal component 2-D long short-term memory for font recognition on single chinese characters," *IEEE Trans. Cybern.*, vol. 46, no. 3, pp. 756–765, Mar. 2016.

[128] J. Z. Lei and A. A. Ghorbani, "Improved competitive learning neural networks for network intrusion and fraud detection," *Neurocomputing*, vol. 75, no. 1, pp. 135–145, Jan. 2012.

[129] L. Nanni, "An ensemble of classifiers for the diagnosis of erythemato-squamous diseases," *Neurocomputing*, vol. 69, nos. 7–9, pp. 842–845, Mar. 2006.

[130] M. Al-Hawawreh and E. Sitnikova, "Leveraging deep learning models for ransomware detection in the industrial Internet of Things environment," in *Proc. Mil. Commun. Inf. Syst. Conf. (MilCIS)*, Nov. 2019, pp. 1–6.

[131] M. Al-Hawawreh and E. Sitnikova, "Industrial Internet of Things based ransomware detection using stacked variational neural network," in *Proc. 3rd Int. Conf. Big Data Internet Things (BDIOT)*, 2019, pp. 126–130.

[132] P. P. K. Chan, C. Yang, D. S. Yeung, and W. W. Y. Ng, "Spam filtering for short messages in adversarial environment," *Neurocomputing*, vol. 155, pp. 167–176, May 2015.

[133] R. M. Silva, T. C. Alberto, T. A. Almeida, and A. Yamakami, "Towards filtering undesired short text messages using an online learning approach with semantic indexing," *Expert Syst. Appl.*, vol. 83, pp. 314–325, Oct. 2017.

[134] C. Hua Li and J. Xiangji Huang, "Spam filtering using semantic similarity approach and adaptive BPNN," *Neurocomputing*, vol. 92, pp. 88–97, Sep. 2012.

[135] T. C. Alberto, J. V. Lochter, and T. A. Almeida, "Post or block? Advances in automatically filtering undesired comments," *J. Intell. Robotic Syst.*, vol. 80, no. S1, pp. 245–259, Dec. 2015.

[136] J. Eisenstein, A. Ahmed, and E. P. Xing, "Sparse additive generative models of text," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 1041–1048.

[137] 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, p. 74, 2003.

[138] V. Pérez-Rosas and R. Mihalcea, "Cross-cultural deception detection," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2014, pp. 440–445.

[139] J. Sánchez-Junquera, L. Villaseñor-Pineda, M. Montes-y-Gómez, and P. Rosso, "Character N-grams for detecting deceptive controversial opinions," in *Proc. Int. Conf. Cross-Lang. Eval. Forum Eur. Lang.* Avignon, France: Springer, 2018, pp. 135–140.

[140] L. Cagnina and P. Rosso, "Classification of deceptive opinions using a low dimensionality representation," in *Proc. 6th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2015, pp. 58–66.

[141] R. Mihalcea and C. Strapparava, "The lie detector: Explorations in the automatic recognition of deceptive language," in *Proc. ACL-IJCNLP Conf. Short Papers*, 2009, pp. 309–312.

[142] S. Nilizadeh, H. Aghakhani, E. Gustafson, C. Kruegel, and G. Vigna, "Think outside the dataset: Finding fraudulent reviews using cross-dataset analysis," in *Proc. World Wide Web Conf.*, 2019, pp. 3108–3115.

[143] J. Li, C. Cardie, and S. Li, "Topicspam: A topic-model based approach for spam detection," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2013, pp. 217–221.

[144] E. A. Suess and B. E. Trumbo, "Using Gibbs samplers to compute Bayesian posterior distributions," in *Introduction to Probability Simulation and Gibbs Sampling with R*. New York, NY, USA: Springer, 2010, pp. 219–248.

[145] C. Elkan and K. Noto, "Learning classifiers from only positive and unlabeled data," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2008, pp. 213–220.

[146] H. Zhao, Z. Lu, and P. Poupart, "Self-adaptive hierarchical sentence model," 2015, *arXiv:1504.05070*. [Online]. Available: http://arxiv.org/abs/1504.05070

[147] B. Liu, W. S. Lee, P. S. Yu, and X. Li, "Partially supervised classification of text documents," in *Proc. ICML*, vol. 2, 2002, pp. 387–394.

[148] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proc. ICML*, vol. 99, 1999, pp. 200–209.

[149] Y. Zhang and D.-Y. Yeung, "A convex formulation for learning task relationships in multi-task learning," presented at the 26th Conf. Uncertainty Artif. Intell., Catalina Island, CA, USA, 2010.

[150] 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.

[151] F. Gieseke, A. Airola, T. Pahikkala, and O. Kramer, "Sparse quasi-Newton optimization for semi-supervised support vector machines," in *Proc. ICPRAM*, vol. 1, 2012, pp. 45–54.

[152] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12 pp. 2493–2537, Aug. 2011.

[153] Y. Jing, "Research of deceptive opinion spam recognition based on deep learning," East China Normal Univ., Shanghai, China, Tech. Rep., 2014.

[154] Y. Ren and D. Ji, "Neural networks for deceptive opinion spam detection: An empirical study," *Inf. Sci.*, vols. 385–386, pp. 213–224, Apr. 2017.

[155] Y. Ren, R. Wang, and D. Ji, "A topic-enhanced word embedding for Twitter sentiment classification," *Inf. Sci.*, vol. 369, pp. 188–198, Nov. 2016.

[156] S. M. Asadullah and S. Viraktamath, "Classification of Twitter spam based on profile and message model using Svm," Dept. Electron. Commun., SDM College Eng. Technol., Karnataka, India, Tech. Rep., 2017.

[157] S. Seneviratne, A. Seneviratne, M. A. Kaafar, A. Mahanti, and P. Mohapatra, "Spam mobile apps: Characteristics, detection, and in the wild analysis," *ACM Trans. Web*, vol. 11, no. 1, p. 4, 2017.

[158] H.-H. Huang, Y.-W. Wen, and H.-H. Chen, "Detection of false online advertisements with DCNN," in *Proc. 26th Int. Conf. World Wide Web Companion (WWW Companion)*, 2017, pp. 795–796.

[159] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2787–2795.

[160] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 1–7.

[161] V. W. Feng and G. Hirst, "Detecting deceptive opinions with profile compatibility," in *Proc. 6th Int. Joint Conf. Natural Lang. Process.*, 2013, pp. 338–346.

[162] Y. Ren, Y. Zhang, M. Zhang, and D. Ji, "Context-sensitive Twitter sentiment classification using neural network," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 215–221.

[163] J. Li, M. Ott, and C. Cardie, "Identifying manipulated offerings on review portals," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 1933–1942.

[164] F. Gräßer, S. Kallumadi, H. Malberg, and S. Zaunseder, "Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning," in *Proc. Int. Conf. Digit. Health*, Apr. 2018, pp. 121–125.

[165] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Hum. Lang. Technol.*, vol. 1, 2011, pp. 142–150.

[166] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, and Y. Zheng, "Recent progress on generative adversarial networks (GANs): A survey," *IEEE Access*, vol. 7, pp. 36322–36333, 2019.

[167] I. Goodfellow, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[168] M. Nickel, V. Tresp, and H.-P. Kriegel, "A three-way model for collective learning on multi-relational data," in *Proc. ICML*, vol. 11, 2011, pp. 809–816.

[169] J. Li, "Feature weight tuning for recursive neural networks," 2014, *arXiv:1412.3714*. [Online]. Available: http://arxiv.org/abs/1412.3714

[170] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1112–1119.

[171] Y.-R. Chen and H.-H. Chen, "Opinion spam detection in Web forum: A real case study," in *Proc. 24th Int. Conf. World Wide Web*, 2015, pp. 173–183.

[172] S. K. Maity, K. C. Santosh, and A. Mukherjee, "Spam2vec: Learning biased embeddings for spam detection in Twitter," in *Proc. Companion Web Conf.*, 2018, pp. 63–64.

[173] F. Xia, J. Liu, H. Nie, Y. Fu, L. Wan, and X. Kong, "Random walks: A review of algorithms and applications," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 4, no. 2, pp. 95–107, Apr. 2020.

[174] X. Wang, W. Jiang, and Z. Luo, "Combination of convolutional and recurrent neural network for sentiment analysis of short texts," in *Proc. 26th Int. Conf. Comput. Linguistics: Tech. Papers (COLING)*, 2016, pp. 2428–2437.

[175] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," 2015, *arXiv:1509.01626*. [Online]. Available: http://arxiv.org/abs/1509.01626

[176] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2016, pp. 1480–1489.

[177] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[178] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*. [Online]. Available: http://arxiv.org/abs/1910.01108

[179] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*. [Online]. Available: http://arxiv.org/abs/1907.11692

[180] G. Vilone and L. Longo, "Explainable artificial intelligence: A systematic review," 2020, *arXiv:2006.00093*. [Online]. Available: http://arxiv.org/abs/2006.00093

[181] J. Zhao, N. Cao, Z. Wen, Y. Song, Y.-R. Lin, and C. Collins, "#FluxFlow: Visual analysis of anomalous information spreading on social media," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1773–1782, Dec. 2014.

[182] S. S. Khan and M. G. Madden, "A survey of recent trends in one class classification," in *Proc. Irish Conf. Artif. Intell. Cogn. Sci.* Dublin, Ireland: Springer, 2009, pp. 188–197.

[183] R. Chalapathy, A. Krishna Menon, and S. Chawla, "Anomaly detection using one-class neural networks," 2018, *arXiv:1802.06360*. [Online]. Available: http://arxiv.org/abs/1802.06360

[184] S. S. Khan and M. G. Madden, "One-class classification: Taxonomy of study and review of techniques," *Knowl. Eng. Rev.*, vol. 29, no. 3, pp. 345–374, Jun. 2014.

[185] C. Désir, S. Bernard, C. Petitjean, and L. Heutte, "One class random forests," *Pattern Recognit.*, vol. 46, no. 12, pp. 3490–3506, Dec. 2013.

[186] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," 2014, *arXiv:1409.7495*. [Online]. Available: http://arxiv.org/abs/1409.7495

[187] P. Capuozzo, I. Lauriola, C. Strapparava, F. Aiolli, and G. Sartori, "DecOp: A multilingual and multi-domain corpus for detecting deception in typed text," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 1423–1430.

RAMI MOHAWESH received the B.S. degree in computer science from Al-Albayt University and the M.S. degree in computer science from the Jordan University of Science and Technology. He is currently pursuing the Ph.D. degree with the University of Tasmania, Tasmania, Australia. In his Ph.D. research, he is the first researcher who investigated the concept drift in fake review detection. He is a reviewer of high impact factor journals, such as the *Information Processing and Management* journal, *Artificial Intelligence Review*, and *Secure Computing*. His research interests include software engineering, cloud computing, natural language processing, cybersecurity, and machine learning. His current work is on fake review.

SHUXIANG XU received the bachelor's degree in applied mathematics from the University of Electronic Science and Technology of China, China, the master's degree in applied mathematics from Sichuan Normal University, China, and the Ph.D. degree in computing from the University of Western Sydney, Australia. He is currently a Lecturer and a Ph.D. Student Supervisor with the School of Information and Communication Technology, University of Tasmania, Tasmania, Australia. Much of his work is focused on developing new machine learning algorithms and using them to solve problems in various application fields. His research interests include artificial intelligence, machine learning, and data mining.

SON N. TRAN received the Ph.D. degree in computer science from City, University of London, U.K., in 2016. Before taking the position of a Lecturer with the University of Tasmania, he was a Postdoctoral Research Fellow with CSIRO, working on the development of a novel prototype of smart homes for aged care. He has publications in flagship conferences and journals, such as *SIGIR*, *IJCAI*, *IJCNN*, *ECIR*, and IEEE Transactions on Neural Networks and Learning Systems (TNNLS). His research interest includes theoretical artificial intelligence, such as bridging the gap between connectionism and symbolism, and applications of (deep) neural networks for various tasks: the Internet of Things, music informatics, natural language processing retrieval, and computer vision.

ROBERT OLLINGTON received the Ph.D. degree in computing from the University of Tasmania, Tasmania, Australia. He is currently a Lecturer with the School of Information and Communication Technology, University of Tasmania. His research interests include neural networks, intelligent robotics and artificial life, and complex adaptive systems.

MATTHEW SPRINGER received the Ph.D. degree in information systems from the University of Tasmania, in 2010. He is currently a Lecturer with the School of Technology, Environments, and Design, University of Tasmania. His major focus has been on improving teaching within the discipline of information and communication technology. He is also an Active Member of the Industry Transformation Research Group and Games and Creative Technologies Research Group.

YASER JARARWEH received the Ph.D. degree in computer engineering from The University of Arizona, in 2010. He is currently an Associate Professor of computer science with the Jordan University of Science and Technology. He has coauthored several technical papers in established journals and conferences in fields related to cloud computing, edge computing, SDN, and big data. He is a Steering Committee Member and the Co-Chair of CCSNA 2018 with INFOCOM. He is the General Co-Chair of the IEEE International Conference on Software Defined Systems SDS-2016 and SDS 2017. He is also chairing many IEEE events, such as ICICS, SNAMS, BDSN, and IoTSMS. He is also the Steering Committee Chair of the IBM Cloud Academy Conference. He served as a guest editor for many special issues in different established journals. He is also an Associate Editor of *Cluster Computing* Journal (Springer) and *Information Processing and Management* (Elsevier).

SUMBAL MAQSOOD received the B.S. degree (Hons.) in computer science from Punjab University College of Information Technology (PUCIT) and the M.S. degree in computer science from GC University, Lahore, Pakistan. She is currently pursuing the Ph.D. degree with the University of Tasmania, Tasmania, Australia. She worked as an IT officer in one of the biggest organization of Pakistan. She is currently working on biosignals analysis using deep learning. Her research interests include machine learning, natural language cybernetics, bio-technologies, data science, and software engineering.