Artificial Intelligence Review
https://doi.org/10.1007/s10462-019-09697-5

Deceptive consumer review detection: a survey

Dushyanthi U. Vidanagama[1] · **Thushari P. Silva**[1] · **Asoka S. Karunananda**[1]

Abstract
Consumer reviews are considered to be of utmost significance in the field of e-commerce, for they have a stronghold in deciding the revenue of a business. When arriving at a purchasing decision, a majority of online consumers rely on reviews since they offer credible means of mining opinions of other consumers regarding a particular product. The trustworthiness of online reviews directly affects a company's reputation and profitability, which is why certain business owners pay fraudsters to generate deceptive reviews. Such generation of deceptive reviews which manipulate the purchasing decision of consumers is a persistent and harmful issue. Hence, developing methods to assist businesses and consumers by distinguishing between credible reviews and deceptive reviews remains to be a crucial, yet challenging task. In view of that, this paper unravels prominent techniques that have been proposed to solve the issue of deceptive review detection. Accordingly, the primary goal of this paper is to provide an in-depth analysis of current research on detecting deceptive reviews and to identify the characteristics, strengths, and bottlenecks of those methodologies which may need further improvements.

Keywords User reviews · Review detection methodologies · Deceptive review detection · Comparative evaluation

1 Introduction

The advancement of Web 2.0 has augmented the trend of online purchasing via e-commerce websites. Online reviews, generated through e-commerce websites and social media denote consumer perception and these reviews play a significant role in e-business, for they could indirectly affect future purchasing decisions. Product manufacturers use reviews to identify product problems as well as to find market intelligence information about their competitors. Leveraging the judging ability of the quality of products/services through online reviews, fraudsters generate fabricated reviews called 'deceptive reviews' to reshape business. Higher numbers of positive reviews induce a customer to purchase a product and fortify financial gains of manufacturers while negative reviews prompt consumers to look

✉ Dushyanthi U. Vidanagama
udeshika@kdu.ac.lk

[1] Department of Computational Mathematics, Faculty of Information Technology, University of Moratuwa, Katubedda, Sri Lanka

Published online: 30 March 2019

for alternatives, thereby causing financial losses. Consequently, a large number of reviews are added through different tools on social media, bringing more challenges in mining opinions and deriving truthful conclusions effectively. Further, there is no control over the quality and content of the web which facilitates the generation of many low-quality consumer reviews out of which deceptive reviews are considered the worst. Thus, the availability of deceptive reviews makes the situation graver as they could influence the potential to reverse a purchase decision. Also worth noting is the ability vendors possess to generate deceptive reviews as the means of either promoting their products or unfairly tarnishing the image of their competitors.

Deceptive review detection has a great impact on customer satisfaction. In a scenario where a customer is cheated or deceived by a review, he/she is most unlikely to use that e-commerce site again for purchases. This brings into focus the requirement of identifying and filtering deceptive reviews from e-commerce sites and social media in order to alleviate problems related to the credibility of online opinion mining. This paper, therefore attempts to review the state-of-the-art methods in deceptive review detection and to present an in-depth analysis of the effectiveness of these methods in order to improve future research in deceptive review detection.

By means of reviewing scholarly studies (Christopher and Rahulnath 2016; Jindal and Liu 2007), review spam is divided into two main categories as to deceptive reviews—the main focus of the paper and destructive reviews. Destructive reviews are non-reviews, comprising mainly of advertisements and other irrelevant reviews containing no opinion (Ren et al. 2014). However, deceptive reviews cause more damage, for they deliberately mislead readers by giving undeserving positive reviews to certain target objects in order to promote the objects, or by giving negative reviews to some target objects in order to damage their reputation (Ren et al. 2014). Deceptive reviews can negatively impact on business due to the loss in consumer trust as well (Crawford et al. 2015). While positive reviews increase the revenue, the negative reviews may decrease the conversions by as far as 67% (Campbell 2014). According to Campbell (2014), 66% of customers place more trust in e-commerce websites that have both positive and negative reviews.

Due to the importance of reviews in profit making, certain business owners offer incentives to those who write favourable reviews about their merchandise, and they even pay fraudsters to write adverse reviews about their competitor products or services (Crawford et al. 2015). Deceptive reviews of that nature have created adverse effects, so much so that researchers strive to develop methods to assist both producers and consumers to distinguish between credible reviews and deceiving reviews (Christopher and Rahulnath 2016). It is expected that effective deceptive review detection methods will ultimately boost consumer trust on products or services which are sold via e-commerce websites.

In terms of the characteristics of a review, it may contain any number of sentences, comment on different aspects of an object, and embody negative, positive or neutral sentiments. Further, online reviews can be considered as big data which exhibits main 4Vs; Volume—the amount of data, Velocity—the arrival of real-time or near real-time review data at a high rate in social media tools, Variety—different structures of review data and Veracity—different quality levels of data. The data volume is regarding the amount and size of reviews. The data velocity is about how frequently reviews are generated by the consumers in real-time or in near real-time. The data variety includes different types of data that are structured, semi-structured and unstructured which are generated through different sources. There is a great variety across industry sectors in terms of reviews (such as hotels, restaurants, e-commerce, and home services), along with the multiplicity of languages that reviews are written in Crawford et al. (2015). Moreover, reviews in social media are quite

unstructured as they consist of emoticons and emoji in addition to the text. A closer observation on e-commerce sites, Twitter and customer rating sites, such as Yelp and Amazon shed light on the volume and velocity of online reviews (Crawford et al. 2015). Veracity, the quality level of data, is a substantive factor in assessing online reviews since the vast majority of reviews are not categorized as trustworthy or not (Crawford et al. 2015).

The rest of this paper is organized as follows: deceptive review detection (see Sect. 2) provides an overview on the nature and functionality of deceptive review detection, while techniques of deceptive review detection (see Sect. 3) presents methods of discovering deceptive reviews with a comprehensive evaluation of each method regarding their accuracy. Data collection (see Sect. 4) elaborates on the parameters of selected research papers for the survey and the graphical representations of data sources used by other researchers. The next section on comparative evaluation (see Sect. 5) summarizes all approaches and the final section (see Sect. 6) provides the conclusion of the paper.

2 Deceptive review detection

Review spams are of three types and they are, namely: untruthful reviews (type 1), reviews on particular brands (type 2) and non-reviews (type 3) (Jindal and Liu 2008). It has been found out that type 2 and type 3 reviews are relatively easy to identify merely by a content analysis of the review (Farooq and Khanday 2016). Furthermore, traditional classification learning methods are quite user-friendly as they utilize manually labeled training data to detect these two types of spam. Nevertheless, recognizing whether a review is an untruthful opinion spam (type 1) is extremely difficult by manually reading the review because one can carefully craft a spam review similar to any other original/real review (Farooq and Khanday 2016). Due to financial benefits offered for generating deceptive reviews, fraudsters attempt to generate fake reviews for the purpose of promoting or de-promoting a product/service. However, these spammers can easily disguise themselves in cyberspace, which makes it relatively difficult to identify them. Due to these phenomena, there is a lacuna of reliable data related to fake and genuine reviews that can be treated as training data.

3 Techniques of deceptive review detection

Deceptive review detection methods can be classified into three major categories as to machine learning approaches, network-based approaches, and pattern-mining approaches. These three types of approaches are elaborated in the subsequent subsections.

3.1 Machine learning approaches for deceptive review detection

Machine learning techniques are widely used in deceptive review detection where supervised learning requires labeled training data, unsupervised learning requires a set of unlabeled data and semi-supervised learning requires a relatively small set of labeled data supplemented with a large amount of unlabeled data. Semi-supervised learning is ideal for cases such as review spam detection where vast amounts of unlabeled data exist (Crawford et al. 2015). Nonetheless, the most commonly used learning method is supervised learning (Crawford et al. 2015). The content of reviews, authors who add reviews and combinations of these two have been considered in applying machine learning techniques.

The review-centric approach detects spam reviews by analyzing the content of the review. Review-centric features are the features that construct using the information contained in a single review (Crawford et al. 2015). These features are categorized into linguistic features, POS tagging, N-grams, sentiments, textual features, rating-related features, and quality-related features. However, purely content-based classifiers have several limitations. Firstly, the spammers can easily manipulate the content to avoid review detection for spam. For example, if the duplicate reviews are to be considered as spam, the spammers may simply reword the content. Secondly, content-based spam detection is designed for specific application domains and it cannot be easily applied to diverse domains. Accordingly, the content-based approach compares the linguistic features of genuine and spam reviews. Thirdly, most content-based classifiers require manually-labeled real datasets.

Reviewer centric features include the characteristics and behaviors of the person who writes a particular review. Further, it was observed that about 75% of spammers write more than five reviews on any given day (Crawford et al. 2015); therefore, the number of reviews a user writes per day can help to detect spammers since 90% of legitimate reviewers never create more than one review on any given day. Approximately 85% of spammers wrote more than 80% of their reviews as positive reviews; thus, a high percentage of positive reviews can be considered an indication of an untrustworthy reviewer. The average review length could also be regarded as an important indication of reviewers with questionable intentions since about 80% of spammers have no reviews longer than 135 words, while more than 92% of reliable reviewers have an average review length of greater than 200 words (Crawford et al. 2015).

3.1.1 Supervised learning techniques

Initially, the researchers who incorporated supervised learning for deceptive review detection used duplicate reviews as positive training examples (fake), and the rest of the reviews as negative training examples (non-fake) (Jindal and Liu 2008). Duplicate and near-duplicate reviews were detected using the 2-gram method based on review content comparison and the review pairs with a similarity score of at least 90% were chosen as duplicates. They incorporated those examples to develop a model to discover non-duplicate reviews with similar characteristics, which could most probably be pseudo-reviews and the logistic regression (LR) they utilized outperformed support vector machine (SVM) and Naive Bayes (NB) classification.

Banerjee et al. (2015) analyzed the extent to which authentic and pseudo- reviews are distinguishable using supervised learning based on four linguistic clues, namely, understand-ability, level of details, writing style, and cognition indicators. Several supervised learning algorithms were used to classify authentic and pseudo-reviews based on linguistic clues and some of the algorithms used included LR, decision tree, neural network, JRip, NB, random forest, SVM, and voting.

In addition, Shojaee et al. (2013) proposed a novel method to detect review spam using Stylometric (Lexical and Syntactic) features. They built SVM and NB classifiers using a hybrid set of both lexical and syntactic features and compared that with either lexical or syntactic features. Shojaee et al. (2013) further found that SVM outperformed NB for all sets of features. Further, Li et al. (2011) tested several supervised methods including SVM, LR, and NB with all three types of features including review-centric, reviewer-centric and product-related features which NB showed outperformed results.

However, due to the lack of reliable truth label of fake/non-fake review data, existing research has relied mostly on ad-hoc or fake/non-fake labels for model building (Mukherjee et al. 2012). Accordingly, Jindal and Liu (2008) used duplicate and near-duplicate reviews as fake reviews which were restrictive and, more often than not, unreliable. It should be noted that most of the other researchers have used manually labeled datasets, which also incurred reliability issues because the accuracy of human labeling of fake reviews is shown to be quite poor (Li et al. 2011; Ott et al. 2011). Ott et al. (2011) used Amazon mechanical turk (AMT) to crowd-source pseudo hotel reviews by paying anonymous online workers (Turkers) to write fake reviews. These were not considered to be 'real' fake reviews as Turkers do not possess sufficient domain knowledge, psychological state of mind or experience to write convincing fake reviews (Mukherjee et al. 2012). Further, Li et al. (2017) proposed a Bayesian approach named 'TopicSpam', a variation of latent Dirichlet allocation (LDA), for deceptive review detection. This aimed at detecting the subtle differences between the topic-word distributions of deceptive reviews vs. truthful reviews.

3.1.2 Semi-supervised learning techniques

A model was created based on positive unlabeled (PU) learning incorporating several reliable negative examples identified from the unlabeled dataset (Ren et al. 2014). The representative positive examples and negative examples were generated based on LDA. For the remaining unlabeled examples which could not be explicitly identified as positive and negative, two similarity weights were assigned, thereby displaying the probability of a spy example belonging to the positive class and the negative class. Finally, the spy examples and their similarity weights were incorporated into SVM to build an accurate classifier (Li et al. 2014).

In addition to that, Rout et al. (2017) demonstrated how four popular semi-supervised learning approaches—Co-training algorithm, expectation maximization algorithm, label propagation and spreading and PU learning—can be used to improve the classification by incorporating new dimensions in the feature vector such as parts-of-speech features, linguistic and word count features and sentimental content features. They achieved high-performance levels using PU learning-based classification. It was observed that spammers' ratings tend to deviate from the average review rating at a far higher rate than legitimate reviewers, thus identifying that user rating deviations may help in detection of dishonest reviewers (Crawford et al. 2015). Moreover, it was identified that the presence of similar reviews for different products by the same reviewer has been shown to be a strong indication of a spammer (Crawford et al. 2015). If the reviews are written for verified purchases they will most likely reflect a genuine review and a reviewer with a higher ratio of verified purchases can be considered as more trustworthy (Crawford et al. 2015).

Jindal et al. (2010) suggested using unexpected deviations from the expected behavior of reviewers since unexpected rules discovered the suspicious behavior of reviewers. Furthermore, Lim et al. (2010) proposed a behavioral approach to detect review spammers who tried to manipulate review ratings on some target products or product groups. They derived an aggregated behavior scoring method to rank reviewers as per the degree they demonstrated spamming related behaviors. In addition, Feng et al. (2012) identified different types of distributional footprints of deceptive reviews which necessarily distorted its distribution of review scores.

3.1.3 Un-supervised learning techniques

Due to the difficulty of human labeling which is required for supervised learning and evaluation, a novel unsupervised text mining model was suggested by confirming the semantic language modeling (SLM). Thus, the text mining-based computational model is effective for detection of untruthful reviews, even if spammers use different strategies (Lau et al. 2011).

A mechanism exists to delete dishonest reviews that would distort the popularity ranking significantly when compared with the removal of a similar set of reviews at random (Wu et al. 2010b). This distortion could be quantified by comparing popularity rankings before and after deletion, using a distortion rank correlation. Mukherjee et al. (2013) proposed a mechanism to observe reviewing behaviors to detect fraudster reviewers using an unsupervised Bayesian inference framework. This unsupervised model is called 'Author Spamicity Model (ASM)' and its basis lies in the theoretical foundation of probabilistic model-based clustering.

The Bayesian framework facilitates the characterization of many behavioral phenomena of opinion spammers, using the estimated latent population distributions. Fei et al. (2013) modeled reviewers and their co-occurrence in bursts as a Markov random field (MRF) and employed the loopy belief propagation (LBP) method to infer whether a reviewer is a spammer or not.

In addition, Kolhe et al. (2014) proposed a scoring algorithm comprising three models; group spam-product model (GPM), member spam-product model (MPM) and group spam-member spam model (GSMS), which were used to analyze the dataset and form candidate groups using the process of frequent itemset mining (FIM) method (Mukherjee et al. 2012). All three types of features—review content features, reviewer behavioral features, and product-related features—are used in unison to identify the deceptive reviews. Mukherjee et al. (2013) formulated opinion spam detection as a Bayesian clustering problem, in which their model considered the degree of spamming of authors and reviews as latent variables with other observed behavioral and linguistic features in latent spam model (LSM).

Nevertheless, Lin et al. (2014) applied different features like review-centric and reviewer-centric with supervised classifiers LR and SVM, while using the unsupervised method by calculating a score value using weight parameters which turn the contributions of feature set and identify deceptive reviews using a threshold value. Further, Singh (2015) utilized the unsupervised Kth Nearest Neibour (KNN) method to detect the outliers of reviews based on the review, reviewer and product-centric features. In addition to that, Wu et al. (2017) defined the set of features that expected to be predictive of suspicious reviews by giving a score value for each feature. Thereafter, they used two aggregation methods named Singular Value Decomposition (SVD) and unsupervised Hedge Algorithm (UH) to calculate a single score of ranking.

3.2 Network based approaches

The Review Graph-based approach is used to capture relationships among reviewers, reviews, stores, items and group reviewers where each node of the graph is attached with a set of features. Wang et al. (2011) showed that certain types of spammers posted many similar reviews about one target entity, using the behavior of reviewers, text similarity, linguistic features, and rating patterns. They proposed a novel concept of a

heterogeneous review graph to capture the relationships among reviewers, reviews and stores that the reviewers have reviewed. They further explored the interactions between nodes in that graph which could reveal the cause of spam and proposed an iterative model to identify suspicious reviewers (Wang et al. 2011). This was the first time such complicated relationships had been identified for review spam detection. Wang et al. (2011) also created an algorithm called Iterative Computation Framework (ICF) for calculating the trustworthiness of reviewers, the honesty of reviews, and the reliability of stores by exploring inter-dependencies between those three.

In a further attempt, a unified probabilistic graphical model named, Unified Review Spamming Model (URSM) was proposed to detect suspicious review spam, the review spammers, and suspicious items in an unsupervised manner (Xu et al. 2015). The goal of URSM was to rank the review texts, the reviewers and items based on their spam behavior, which is modeled as a latent variable by observing abnormal features of reviews, users, items, and the text generality detected from the review text.

An iterative algorithm was proposed by Noekhah et al. (2014) to detect pseudo-reviews, review spammers and groups of spammers considering the above mentioned three types of features and entity relationships between reviews, reviewers, products and group of reviewers.

Further, Shehnepoor et al. (2017) proposed NetSpam framework that was a novel network-based approach which models review networks as heterogeneous information networks. Its classification step used different meta-path types which were innovative approaches in spam detection domain. Moreover, Shehnepoor et al. (2017) proposed a novel weighting method for spam features to determine the relative importance of each feature and demonstrated the effectiveness of each feature when identifying spam from normal reviews.

Another unsupervised network-based framework presented by Akoglu et al. (2013) successfully captured the correlations of labels among users and products. This was an iterative, propagation-based algorithm that exploited the network structure and the long-range correlations to deduce the class labels of users, products, and reviews.

Furthermore, a neural network model to learn document-level representation for detecting deceptive opinion spam was later generated and it processed sentence representation with convolution neural network (CNN) (Ren and Ji 2017; Ren and Zhang 2016). Then, sentence representations were combined using a Gated Recurrent Neural Network (GRNN). Finally, the document representations were directly used as features to identify deceptive opinion spam. Similarly, Wu et al. (2017) proposed a Twitter spam detection method based on deep learning which addressed the challenges of existing feature-based classifiers. Jain et al. (2016) demonstrated the procedure of spam detection using CNN and Recursive Neural Network (RNN) by learning from high-level features on their own with the help of raw data.

Along with the successful utilization of Neural Networks on various classification applications, Aghakhani et al. (2018) suggested to use Generative Adversarial Networks (GANs) for deceptive review detection. GANs are a class of artificial intelligence algorithms used in unsupervised machine learning, implemented by a system of two neural networks contesting with each other in a zero-sum game framework (Aghakhani et al. 2018). Thus, using a semi-supervised learning method such as GAN can eliminate the issue of scarcity of labeled dataset.

3.3 Pattern mining approaches

Temporal patterns have demonstrated several longitudinal studies along the time dimension. At times, review spammers were active in certain days and certain reviewers displayed identified rating patterns over a time period. Further, the deceptive reviewers could be focused on certain locations. These temporal patterns enabled some longitudinal studies which use features along the time dimension. Nevertheless, it has been brought into focus that a majority of reviewers generally write single reviews. 68% of the reviewers write only one review on Amazon and such an occurrence is called a 'singleton review' (Xie et al. 2012). Xie et al. (2012) introduced a novel approach that maps the singleton review spam detection problem to an abnormally correlated temporal pattern-detection problem. The algorithm they suggested was based on multi-scale multi-dimensional time series anomaly detection (Xie et al. 2012).

Furthermore, sentiment analysis techniques are incorporated for review spam detection by computing the sentiment score and then combining the discriminative rules from the abnormal time windows (Peng and Zhong 2014). Li et al. (2015) considered meta-data patterns, both temporal and spatial when detecting deceptive reviews. It was found that the registration pattern of the reviewers, the average travel speed of a reviewer from the written location of one review to another, writing patterns over the weekdays and weekends, IP addresses, cookies and pattern of the locations where reviews are written to be noteworthy features of this review writing process.

As per Li et al. (2016), Labeled Hidden Markov Model (LHMM) was proposed to detect spammers by modeling their temporal posting behaviors. While this approach required only the time-stamp of reviews, their analysis showed major differences of temporal patterns of spammers and non-spammers (Li et al. 2016). Further, Santosh and Mukherjee (2016) discovered various temporal patterns and their relationships with the rate at which fake reviews are posted. They had employed vector auto-regression (VAR) to predict the rate of deception across different spamming policies such as early spamming, mid spamming and late spamming. Additionally, they explored the effect of reviews on future rating and popularity prediction of entities as well.

4 Data collection

This paper is mainly focused on research findings on the area of deceptive review detection. Around 63 research papers were studied in-depth to gather findings on different deceptive review detection techniques. According to Fig. 1, around 30% of researchers used datasets from Amazon while 15% used AMT datasets by paying anonymous online workers to write fake reviews.

More than 30% of research has used the review content based features to detect deceptive reviews, while 24% has used both content features, reviewer behavioral features and product-related features (Fig. 2). Furthermore, less than 5% of research has used temporal and spatial approaches and big data analytic approaches (Fig. 2) and more than 40% of research detects deceptive reviews using the supervised method, while around 15% has used semi-supervised and unsupervised methods (Fig. 3). Overall, most of the research has incorporated machine-learning techniques whereas fewer have used self-supervised, LDA, pattern analysis, neural network and multitasking learning methods (Fig. 3).

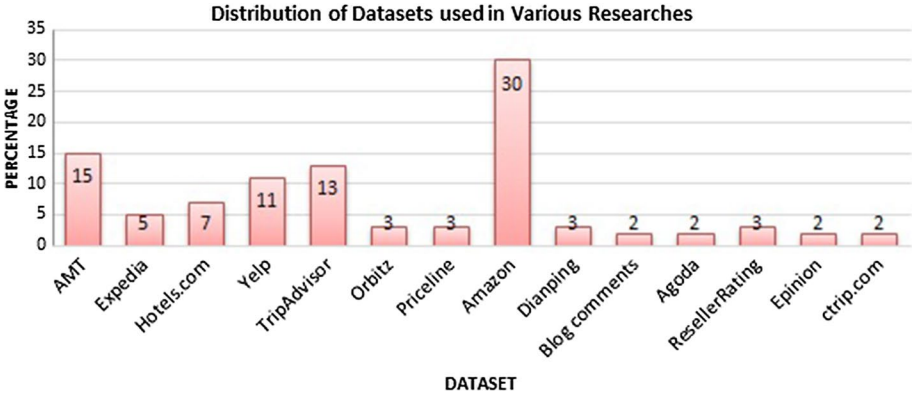

Fig. 1 Distribution of the dataset used in various researches

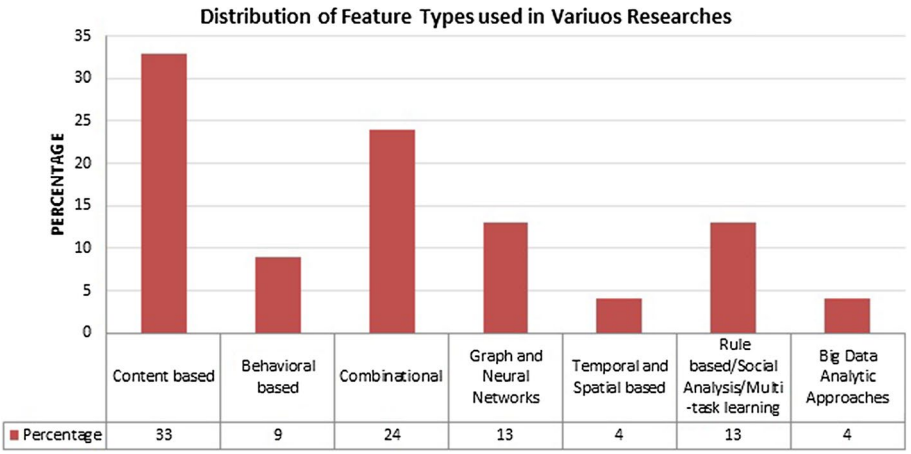

Fig. 2 Distribution of feature types used in various researches

5 Comparative evaluation

Tables 1 and 2 demonstrate that most of the research has used machine-learning techniques, Bayesian technique and topic-modeling technique for content analysis of the deceptive review detection. Further, the aforesaid approaches have used precision, accuracy, recall, F measure and Area Under the Curve (AUC) to evaluate the methods.

When considering the review content-based deceptive review detection approach, the SVM technique outperformed NB, KNN, LR, and Decision Tree techniques while NB is more appropriate for small datasets, but SVM for large sets of data (Badresiya et al. 2014). PU learning on SVM detects a large number of potential pseudo-reviews hidden in the unlabeled set (Li et al. 2014; Ren et al. 2014). Nonetheless, LR has the potential of achieving high F Score value with PU learning (Rout et al. 2017). Further, the SLM technique proves to be quite effective for untruthful detection and a combination of LR, KNN, and SVM is better for non-review detection (Lau et al. 2011). Li et al. (2014, 2017) used a

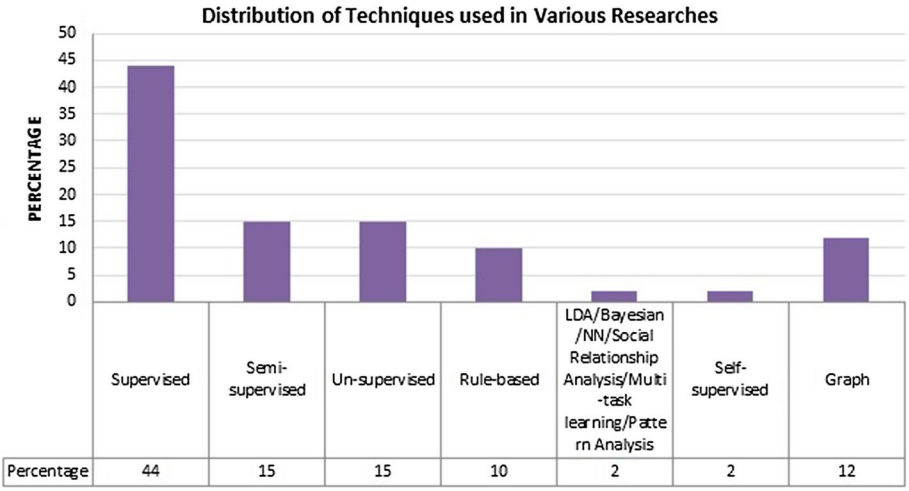

Fig. 3 Distribution of techniques used in various researches

generative Bayesian approach which achieved better results in comparison to SVM and generative LDA-based topic modeling approach with high accuracy. In addition to that, Hai et al. (2016) concluded that high-quality unlabeled review data may help SMTL-LLR to learn better.

When considering the features of review content, a high F-measure value could be achieved by applying SVM with SMO over a combination of both lexical and syntactic features (Shojaee et al. 2013) Also worth noting is that the sentence structure manipulates the genre of a text and detects deceptive opinions (Xu and Zhao 2012). It should also be highlighted that detecting pseudo-reviews requires both linguistic features and behavioral features (Shivagangadhar et al. 2015) and the combinatory approach using psycho-linguistically motivated features and n-gram features can perform slightly better (Shivagangadhar et al. 2015). Furthermore, Banerjee et al. (2015) exposed that the understandability, level of details, writing style, and cognition indicators offer useful linguistic clues to distinguish between authentic and fake reviews. Also, the titles of reviews are as useful as their descriptions can use to distinguish between authentic and fake reviews by incorporating meta-classification algorithms such as voting, which emerged as the best performing algorithm. In addition, Lai et al. (2010a) found that Self-supervised Decision Tree method is more flexible than most supervised spam detection systems which depend primarily on pre-classified data. The KL divergence and the probabilistic language modeling-based computational model are also considered to be effective for the detection of untruthful reviews. Moreover, the SVM-based method is yet another effective mechanism to detect non-reviews (Lai et al. 2010a).

All aforesaid methods suggested to combine different features and to evaluate the approaches with large datasets in diverse domains. Lau et al. (2011) suggested to further improve the accuracy of SLM method, evaluate the effectiveness and efficiency of design artifacts based on a larger dataset and examine more sophisticated language modeling approaches, such as n-gram language models, to improve the effectiveness of the untruthful review detection method. Also, the semantic granularity of text will be examined as a candidate feature for untruthful review detection and the impact of fake

Table 1 Summary of identifying deceptive reviews using content analysis (Part 1)

Paper	Dataset used	Features used	Approach	Method	Evaluation criteria	Result
Badresiya et al. (2014)	Amazon mechanical turk (AMT) dataset used in Ott et al. (2011)	Content analysis	Supervised	NB	Accuracy, precision and recall using Rapidminer data mining tool	Accuracy = 66.44%
				SVM		Accuracy = 83.19%
				KNN		Accuracy = 68.56%
				Logistic regression		Accuracy = 82.39%
				Decision tree		Accuracy = 51%
Hai et al. (2016)	10,000 Unlabeled data from doctor, hotel and restaurant domains	Text unigram and bigram term-frequency features	Semi-supervised	SMTL-LLR	Accuracy, precision and recall	Accuracy = 87.2%
				MTL-LR		Accuracy = 85.2%
				MTRL		Accuracy = 84%
				TSVM		Accuracy = 82.9%
				LR		Accuracy = 82.1%
				SVM		Accuracy = 81.8%
				PU		Accuracy = 72.6%
Shojaee et al. (2013)	800 Truthful reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp and 800 fake reviews from AMT	Lexical and syntactic features	Supervised	Support vector machine (SVM) with sequential minimal optimization (SMO)	F measure	84%
Xu and Zhao (2012)	English gold standard dataset from Ott et al. (2011) and Chineese dataset	Deep linguistic features	Supervised	SVM	F measure	89.80%
				Maximum entropy model using openNLP MAXENT		91.70%

Table 1 (continued)

Paper	Dataset used	Features used	Approach	Method	Evaluation criteria	Result
Lau et al. (2011)	Amazon dataset	Syntactical, lexical, and stylistic features for SVM	Unsupervised/supervised	Untruthful reviews: text mining model integrated with semantic language model (SLM)	1-AUC	0.13%
				Non-reviews: SVM	1-AUC	2.51%

Table 2 Summary of identifying deceptive reviews using content analysis (Part 2)

Paper	Dataset used	Features used	Approach	Method	Evaluation criteria	Result
Rout et al. (2017)	Amazon mechanical turk (AMT) dataset used in Ott et al. (2011)	Bigram frequency counts, parts-of-speech features, linguistic and word count features and sentimental content features	Semi-supervised	Co-training algorithm	F measure	78%
				Expectation maximization algorithm		83%
				Label propagation and spreading		83%
				PU learning for the classifiers KNN, LR, random forest, stochastic gradient descent		84%
Li et al. (2014)	Chinese reviews hosting site Dianping	Unigram and bigram features	Semi-supervised	PU learning on SVM	F measure	67%
Ren et al. (2014)	Gold standard dataset from Ott et al. (2011)	Similarity weights of spy examples	Semi-supervised	PU learning on LDA and SVM	Accuracy	83.91%
Shivagangadhar et al. (2015)	Yelp dataset of hotels and restaurant	Unigram, bigram features, review length	Supervised	LR, SVM, NB	Accuracy	Different accuracies for different features on different classifiers
Ott et al. (2011)	Gold-standard opinion spam dataset from Amazon	LIWC + Bigram	Supervised	SVM	Accuracy	89.60%
Li et al. (2017)	Dataset from Ott et al. (2011) contains hotel reviews from TripAdvisor	Word frequency	A generative LDA-based topic modeling approach	TopicSpam	Accuracy	94.80%
Bhattarai and Dasgupta (2012)	Blog comments	Content based features	Self-supervised	Self supervised decision tree (J48)	Accuracy	71.58%

Table 2 (continued)

Paper	Dataset used	Features used	Approach	Method	Evaluation criteria	Result
Lai et al. (2010a, b)	Amazon dataset	syntactical, lexical, and stylistic features	Supervised/un-supervised probabilistic language model	Untruthful reviews: Kullback–Leibler (KL) divergence method	AUC	0.858
				Non-reviews: SVM	AUC	0.806
Li et al. (2014)	Dataset created by Turkers and experts, Ott et al. (2011) dataset from hotels, restaurants and doctors	Unigram, LIWC, POS	Generative Bayesian Approach	SAGE model	Accuracy	0.65
Banerjee et al. (2015)	Reviews from Agoda.com, Expedia.com and Hotels.com	Linguistic features	Supervised	LogReg, C4.5, BPN, JRip, NB, RF, SVML, SVMP, SVMRBF and voting	AUC	0.815

reviews on product sales will be examined based on an econometric analysis. Table 3 summarizes the research which identified deceptive reviews based on reviewer behavioral analysis using machine learning approaches while Jindal et al. (2010) used a rule-based approach with unexpected rules.

When considering the reviewer behavior, there exists a connection between the distributional anomaly and deceptive reviews (Feng et al. 2012). Further, the aggregated behavior scoring methods depicted outperformed results (Lim et al. 2010) and unexpected rules and groups represented abnormal or unusual behaviors of reviewers, thereby indicating spam activities (Jindal et al. 2010). Moreover, Mukherjee et al. (2013) found that the Author Spamicity Model is effective and it outperforms strong competitors. In view of this, it could be suggested to incorporate review spammer detection for review detection and vice versa to explore new ways to learn spamming related behavior patterns (Lim et al. 2010).

Tables 4 and 5 summarise the research which used a combination of all features such as content specific, reviewer specific, product specific and semantic features, without being limited to one set of features. Most of the researches, in this context, focused on machine-learning techniques. Even though Jindal and Liu (2007, 2008) found LR to be highly effective, SVM presented a high accuracy and precision by including a new set of features (Christopher and Rahulnath 2016; Daiyan et al. 2014; Li et al. 2014). Nonetheless, Rout et al. (2016) asserted that Decision Trees have the potential of achieving high performance in comparison to other supervised approaches, while Mushtaq et al. (2016) articulated that Bayes probabilistic classifier is more sensitive to detect spam.

Lin et al. (2014) elaborated in their study that the unsupervised method which uses a spam score with a threshold value can achieve a relatively positive effect without training samples. Also worth noting is that unsupervised methods such as LSM-UP and LSM-HE significantly outperform baseline clustering methods on entropy, purity, and F1 (Mukherjee et al. 2013). Further, singular value decomposition (SVD) is also known to outperform the unsupervised hedge algorithm (UH) (Wu et al. 2010a). The unsupervised clustering explored distortion in popularity ranking and that distortion contributes to demarcate true positives from false positives (Wu et al. 2010b). The semi-supervised two-view co-training algorithms can achieve better results than the single-view algorithms (Li et al. 2011). Jindal and Liu (2007, 2008) further suggested using a dataset in a different domain and aligning with that, Christopher and Rahulnath (2016), analysed on social media data which revealed more behavioral traits of reviewers. Li et al. (2011), exploiting the probabilistic two-view algorithm such as Co-EM, suggested to model the uncertainty in review spam identification task and to test the co-training algorithm in other opinion resources such as blogs, or Twitter. Daiyan et al. (2014) also suggested improving the detection method with a different corpus.

Table 6 summarises the research which used graph-based and network-related approaches. Best results from graph-based approach were obtained when all behavioral and linguistic features are combined together (Noekhah et al. 2014) and it identified subtle spamming activities with fine precision and human evaluation agreement (Wang et al. 2011). Also worth noting is that experimental results on three popular review datasets demonstrated the effectiveness of the probabilistic graphical model (Xu et al. 2015). Shehnepoor et al. (2017) calculated weights using the meta-path concept which could be quite effective in identifying spam reviews and may lead to a better performance. Even without a train set, NetSpam could calculate the importance of each feature and it yielded better performance in the feature addition process, and performed better than previous work, with a limited number of features. Furthermore, reviews behavioral category performed better

Table 3 Summary of identifying deceptive reviews using behavioral analysis

Paper	Dataset used	Features used	Approach	Method	Evaluation criteria	Result
Feng et al. (2012)	Pseduo-gold strandard labeled dataset based on different types of distributional footprints from TripAdvisor and Amazon	Rating distribution or the historic rating distribution of a reviewer	Supervised	LIBSVM	Accuracy	High accuracy than the human judged dataset
Lim et al. (2010)	Amazon dataset	Rating behavior features	Unsupervised	Aggregated behavior scoring methods	Inter-evaluator consistency	Cohen's Kappa value = between 0.48 and 0.64
					Spammer ranking performance	High
Jindal et al. (2010)	Amazon dataset	Behavioral features	Rule based approach	Unexpected rules	Statistical test	Significant values
Mukherjee et al. (2013)	Reviews of manufactured products from Amazon.com	Author features and review features	Unsupervised	ASM	Accuracy and human evaluation	High accuracy than the human evaluated results

Table 4 Summary of identifying deceptive reviews using combination feature approach (Part 1)

Paper	Dataset used	Features used	Approach	Method	Evaluation criteria	Result
Jindal and Liu (2007, 2008)	5.8 Million reviews, 2.14 million reviewers and 6.7 million products from Amazon	Review, reviewer and product centric features	Supervised	LR	AUC	78%
Mukherjee and Venkataraman (2014)	AMT dataset from Ott et al. (2011), Amazon dataset from Mukherjee et al. (2012) and Yelp dataset	Review features and reviewer features	Unsupervised	LSM-UP	Purity and entropy, precision, recall, and F score	74.6%, 69.2%, 58.4%
				LSM-HE	Purity and entropy, precision, recall, and F score	75.9%, 74.3%, 61.6%
Christopher and Rahulnath (2016)	Reviews for musical instruments from Amazon	Review and reviewer centric features including burst review detection, sentiment scores, content similarity, review deviation	Supervised	SVM	Precision, recall, accuracy and F measure	F = 73.7%
Li et al. (2011)	Manually build corpus from product reviews in Epinions	Content, product, sentiment, meta data, behavior features	Semi-supervised	Co-training algorithm	F measure	63%
Daiyan et al. (2014)	Amazon reviews of mobile and camera—manually labeled	Review, reviwer and product features	Supervised	NB SVM	Accuracy, F measure	SVM Accuracy and F Measure is high for all product categories
Lin et al. (2014)	Jindal and Liu (2007) review dataset	Review content and reviewer behaviors	Unsupervised/supervised	Supervised: LR	F Score	85.60%
				SVM		92.30%
				Unsupervised: spam score with a threshold value		85.1% with 0.55 threshold value

Table 5 Summary of identifying deceptive reviews using combination feature approach (Part 2)

Paper	Dataset used	Features used	Approach	Method	Evaluation criteria	Result
Jindal and Liu (2007)	Reviews on manufactured products	Review and reviewer centric features	Supervised	LR	AUC	78%
Rout et al. (2016)	Supervised approach: TripAdvisor.com, Mechanical Turk, Expedia, Yelp, Orbitz, Hotels.com	Linguistic features, POS features and the sentiment score, n-gram	Supervised/unsupervised	Decision trees	Accuracy	92.11%
	and Priceline			NB		91.90%
				SVM		88.71%
	Unsupervised: Amazon unlabeled dataset	Review centric, textual features, rating related features, reviewer centric features, product centric features		Outlier detection from KNN		
Mushtaq et al. (2016)	Amazon dataset	Page rank and content based approach along with prestige of reviewer	Supervised	Bayes probabilistic classifier	Kendals rank correlation	Helpfulness = 0.594, review goodness = 0.589
					Spearmans rank order correlation	Helpfulness = 0.62, review goodness = 0.60
					Osim correlation methods	Helpfulness = 0.613, review goodness = 0.587

Table 5 (continued)

Paper	Dataset used	Features used	Approach	Method	Evaluation criteria	Result
Wu et al. (2010a, b)	Reviews from TripAdviso	Proportion of positive singletons, concentration of positive singletons, reactive positive singletons, review weighted rating, contribution weighted rating, truncated rating, sentiment shift, positive review length difference	Unsupervised	Singular value decomposition (SVD), and the unsupervised hedge algorithm (UH)	Correlation analysis and grapical method	SVD outperforms UH
Wu et al. (2010a, b)	Irish TripAdvisor data	Proportion of positive singletons, concentration of positive singletons	Unsupervised Clustering	Explore distortion in popularity ranking	Time plots and scatter plots	

Table 6 Summary of identifying deceptive reviews using graph based and neural network approaches

Paper	Dataset used	Features used	Approach	Method	Evaluation criteria	Result
Noekhah et al. (2014)	Labeled data from Amazon	Behavioral and linguistic features of reviews, reviewers, group of reviewers, and their targets	Graph based	Graph based model and iterative multi-level algorithm	Accuracy	93%
Wang et al. (2011)	Store review data from www.resellerratings.com	Review, reviewer and store based features	Graph based	Graph based approach	Precision of human evaluators	60.30%
Ren and Zhang (2016)	Li et al. (2014) dataset of domains hotels, restaurant and doctor	Neural and descrete features	Neural network	CNN	Accuracy	75.90%
				RNN		63.20%
				GRNN		80.10%
				Bi-directional GRNN		83.60%
				Bi-directional GRNN (attention)		84.10%
Mukherjee et al. (2012)	Product reviews from Amazon	Group spam behavior features	Relational Analysis	Relation-based model of GSRank	AUC	Over 90%
Xu et al. (2015)	Amazon audioCD data, the TripAdvisor hotel data, and the Yelp restaurant data	Abnormal features about reviews, users and items, text generality features	Probabilistic graphical model	URSM	F measure and accuracy	Higher F value and accuracy for all the datasets for the proposed model
					Human evaluation	Larger value for top part (TP), smaller values for middle part (MP) and bottom part (BP)
Shehnepoor et al. (2017)	Real-world review datasets from Yelp and Amazon Web sites	Review-behavioral, Userbehavioral, review-linguistic, and user-linguistic	Graph based	NetSpam	AUC	High accuracy

Table 6 (continued)

Paper	Dataset used	Features used	Approach	Method	Evaluation criteria	Result
Aghakhani et al. (2018)	Hotel reviews from TripAdvisor	Content features	Semi-supervised neural network-based learning method	Generative Adversarial Network	Accuracy	0.891

than other categories. In terms of the neural network approach, bi-directional gated recurrent neural network model with attention mechanism presented high accuracy.

In addition to that, Ren and Zhang (2016) suggested to improve the accuracy by integrating discrete and neural features for the feature set and Shehnepoor et al. (2017) suggested to use a similar meta-path concept which could be used to find spammer communities, utilizing the product features. It certainly is a promising direction for future studies since they used features mostly related to spotting spammers and spam reviews, while single networks have received considerable attention from various disciplines for over a decade. It should also be noted that information diffusion and content sharing in multi-layer networks is still a young branch of research.

In recent years, deep learning techniques are successfully used for text classification tasks. Though recurrent neural network (RecurrentNN) and convolutional neural network (CNN) showed satisfactory performance in classification, the complexity of such neural network-based methods requires much larger datasets to reach a reasonable performance (Aghakhani et al. 2018). Further, Aghakhani et al. (2018) proposed a GAN based approach to detect deceptive reviews which required few credible datasets.

Table 7 summarises the research which used temporal and spatial features for the detection of deceptive reviews. As per Li et al. (2015), a combination of uni-gram and bi-gram features, behavioral features, temporal and spatial features demonstrated high accuracy. The multi-scale anomaly detection algorithm on multi-dimensional time series based on curve fitting is effective in detecting singleton review spam (Xie et al. 2012). Xie et al. (2012) further performed an extensive analysis on the temporal patterns of opinion spamming and analysed different temporal patterns of spammers and non-spammers.

5.1 Alternative approaches

Table 8 summarises the research which followed association rule-based mining approaches, social relationship analysis approaches and multi-task learning approaches used to detect deceptive reviews. Wahyuni and Djunaidy (2016) stated that the precision value of rule-based ICF++ is higher than ICF, while Lai et al. (2010b) proposed an inferential language model which outperformed all the other methods in TREC-like experiment, with a large dataset downloaded from Amazon. The trust-based rating predictions—RWR algorithm achieved a higher accuracy than standard Computation Framework (CF) method and a strong correlation between social relationships and the computed trustworthiness scores (Xue et al. 2015).

Furthermore, a multi-task learning method via logistic regression (MTL-LR) has been developed in order to allow boosting of the learning for one task by means of sharing the knowledge contained in training signals of other related tasks (Hai et al. 2016). To leverage the unlabeled data and regularize each base model, Hai et al. (2016) introduced a Laplacian graph and proposed a semi-supervised multi-task learning model via Laplacian regularized logistic regression (SMTL-LR), thus generating better performance standards. In addition to that, Yang et al. (2015) suggested a general framework to detect spam reviews based on coherent examination and iterative computational framework. This model attempted to analyze review coherence in the granularity of sentence using several metrics to investigate the coherence between sentiment words and other related words based on the smooth flow between sentences: word transition probability and word concurrence probability.

It has also been suggested to optimize the proposed process (Wahyuni and Djunaidy 2016) and to improve computation of the sentiment score considering the modifier word

Table 7 Summary of identifying deceptive reviews using temporal and spatial features

Paper	Dataset used	Features used	Approach	Method	Evaluation criteria	Result
Li et al. (2015)	Restaurant reviews from Dianping.com	Spatial and temporal features	Supervised	SVM	Accuracy, precision and recall	Accuracy = 85%
Xie et al. (2012)	Review data from (resellerratings.com)	Temporal features	Pattern analysis	Multi-scale anomaly detection algorithm on multi-dimensional time series based on curve fitting	Human evaluation	High accuracy
Li et al. (2016)	Restaurant reviews from dianping.com	Temporal behavior of reviewers	Pattern analysis	Labeled Hidden Markov Model to detect spammers	Accuracy, precision, recall and F1-score	F1 Score = 84% Precision = 78% Recall = 88% Accuracy = 83%
			Co-bursting network to detect spammer groups	Louvain	Purity and entropy	Purity = 83% entropy = 67%
				Kmeans		Purity = 86% entropy = 73%
				Hierarchical		Purity = 88% entropy = 73%
Santosh and Mukherjee (2016)	Reviews from yelp restaurant	Time series features (TSF), behavioral features (BF) and n-gram features (NF)	Combinational feature approach	Linear Kernal SVM	Accuracy, precision, recall and F1-score	Combining TSF with NG and BF feature sets, resulted the highest F-scores

Table 8 Summary of identifying deceptive reviews using other approaches

Paper	Dataset used	Features used	Approach	Method	Evaluation criteria	Result
Wahyuni and Djunaidy (2016)	Product reviews from Amazon	Review content and sentiment polarity	Rule based approach	ICF++	Precision of human evaluators	63%
Peng and Zhong (2014)	Product reviews from Resellerrating.com	Sentiment score	Rule based approach	Discriminative rules and time series analysis	Accuracy	Over 82%
					Human evaluation	65%
Lai et al. (2010a, b)	Amazon dataset	Not rely on features	Association mining	Un-supervised inferential language model	Measure of failures	lam = 1.44%
Xue et al. (2015)	Yelp dataset	Rating, rating deviation	Social relationship analysis	Trust-based rating predictions—RWR algorithm	MAE and MAUE	Complements only, friendships only and two-faceted approaches have similar MAE and MAUE values about 80% of the users have high trustworthiness scores larger than 0.9
				Overall trustworthiness score	CDF	
Hai et al. (2016)	Unlabeled doctor, hotel, restaurant reviews data from ratemdscom, TripAdvisor and Yelp	Text unigram and bigram term-frequency features	Multi-task learning method	MTL-LR	Accuracy	87.20%
				SMTL-LR		85.20%
Yang et al. (2015)	Human labeled dataset with reviews from ctrip.com	Word transition probability, word concurrence probability	Rule based approach	Coherence matrics based on iterative computation	Human evaluation	46%

and the expansion of the discriminative rules (Peng and Zhong 2014). Lai et al. (2010b) too proposed an evaluation of both the effectiveness and the efficiency of the proposed method, based on a larger dataset and more sophisticated language modeling approaches such as n-gram language models.

5.2 Big data analytic approach

Apart from the discussed methods of deceptive review detection, big data analytics methods and tools can also be integrated with the detection process. Big data analytics can be classified as distributed or single host-based approaches. For single host approaches, analytical frameworks such as the BIG Data Suite support batch mode processing of big data, whereas the MOA framework can be used to analyze evolving big data streams (Zhang et al. 2014). It is believed that combining both batch and streaming modes of analytics is essential to develop an operational big data analytics framework to support real-world applications. Since spammers tend to modify their deceptive styles, it is crucial to design an adaptive detection method that could continuously improve its feature set and system parameters based on users' feedback (Zhang et al. 2014). Accordingly, Zhang et al. (2014) designed a novel big data analytic framework which influenced the distributed computing and streaming to efficiently process big social media data streams. Subsequently, they applied the proposed framework that was supported by a novel parallel co-evolution genetic algorithm to adaptively detect deceptive reviews with respect to different social media contexts (Zhang et al. 2014).

Further, Chavan et al. (2017) investigated opinion spam in reviews and proposed a technique to identify spam review using sentimental analysis and statistics using the MapReduce technique developed by Apache Hadoop. Ultimately, the deceptive reviews were decided by analysing the outliers.

5.3 Characteristics and drawbacks of the approaches

Table 9 describes the characteristics and drawbacks of the most prominent approaches discussed above. Approaches which mainly focus on the content of the review considered only the linguistic features. The behavior-based approach selects reviewer profile and behavioral features for the deceptive review detection. These two approaches require a large amount of labeled data for supervised learning.

Even though the graph-based approach analyses the relationship among reviews, reviewers, and products using graphs to identify the interdependent nature among all approaches, it may ignore textual features of reviews. Moreover, the interdependent nature of reviews, reviewers, and products need to be carefully analysed to filter spammers. Although this approach is effective, it is less efficient and scalable.

Neural network approaches are ideal for deciphering semantics of lengthy texts. The continuous learning ability of Neural Networks has the potential of producing better deceptive detection rates by acquiring few input features. However, this approach requires high computational power and large amounts of data. Even though the approaches which consider the temporal and spatial features discovered unusual patterns to detect spammers, this approach tends to ignore the textual features of reviews.

Table 9 Characteristics and Bottlenecks of approaches

Approach	Characteristics	Bottlenecks
Content-based approach	1. Depends on the content of the review	1. Consider the linguistic features only
	2. Review based features are used for feature extraction	2. Lack of large amount of labeled data for supervised learning
	3. Machine learning techniques are used for classification	
Behavioral-based approach	1. Depends on the behavior of reviewers	1. Consider the reviewer centric features only
	2. Reviewer profile and behavioral features are used for feature extraction	2. Lack of large amount of labeled data for supervised learning
Graph based approach	1. Analyse the relations among reviews, reviewers, and products through a heterogeneous graph (review graph) to identify review spammers	1. Ignore the textual information of the reviews
	2. Inter-dependent nature of data	2. Inter dependence of reviews, reviewers, and products need to be carefully accounted for spams
	3. Relational nature of problem domain	3. Definitions of spams are much more diverse
		4. Graph-based spam detection algorithms need to be designed not only for effectiveness but also for efficiency and scalability
Neural network approach	1. Better captures the contextual information and is ideal for realizing semantics of long texts	1. Required high computational power
	2. Continuous learning feature produces favorable detection rates	2. Required more data
	3. Require fewer input features	
Temporal and spatial approach	1. Discover unusual temporal patterns, because they allocate a large portion of reviews and can have effect on the product rating trend	1. Ignore the textual features of the reviews

6 Conclusion

Due to the advancements of the Internet, people are more inclined to buy and sell online products and services. Since the products and services through e-commerce sites are intangible for consumers, their trust and credibility are solely dependent on reviews; thus manipulating the purchasing decisions to a great extent. Manufacturers also rely heavily on reviews in order to improve the quality and features of the items according to customer preferences. Due to these reasons, it is integral to maintain and ensure the reliability of reviews. Alarming statistics on untruthful reviews on the internet have led many researchers to explore techniques to effectively and efficiently identify deceptive reviews from genuine reviews, yet their commentary is limited to opinions only on truthful reviews.

Thus, the main concern of the current research can be summarized as follows. Supervised or semi-supervised methods require real-world data for model building. Even though certain researchers presume duplicate and near duplicate reviews as deceptive, such assumptions are often unreliable. Further, the usage of manually labeled data or AMT crowd-sourced data cannot be treated as reliable. Hence, future researchers should primarily focus on less number of labeled datasets or neither. Also worth nothing is that it does not suffice to extract a permanent set of features concerning the content of the review or reviewer behavior. Therefore, most of the researchers have used a combination of review-centric, reviewer-centric and product-related features when detecting deceptive reviews. Another major concern is that the consumer reviews are generated in high volume with velocity and variety, thereby being treated as big data which requires the application of big data analytic approaches to analyze reviews and identify deceptive reviews in real-time or near real-time. After having compared different approaches in a wider spectrum, it can be concluded that each approach has its own positive and negative impact. Hence, it is wise to consider a hybrid approach, taking all positive aspects identified in the above discussion.

Thus, the ultimate goal of this research was to facilitate a way forward to ensure a satisfied customer who would increase the revenue of the business by increasing purchases led by trustworthy and reliable reviews.

Acknowledgements We are immensely grateful to all the researchers with whom we had the pleasure of working during the completion of this research. Further, we would like to thank each member of our staff who extended us their extensive personal and professional guidance throughout this research.

References

Aghakhani H, Machiry A, Nilizadeh S, Kruegel C, Vigna G (2018) Detecting deceptive reviews using generative adversarial networks. In: 1st deep learning and security workshop co-located with the 39th IEEE symposium on security and privacy. arXiv:1805.10364 [cs.CR]

Akoglu L, Chandy R, Faloutsos C (2013) Opinion fraud detection in online reviews by network effects. In: International AAAI conference on web and social media, North America. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/5981. Accessed 03 June 2017

Badresiya A, Vohra S, Teraiya J (2014) Performance analysis of supervised techniques for review spam detection. Int J Adv Netw Appl Special Issue:21–24

Banerjee S, Chua A, Kim, J (2015) Using supervised learning to classify authentic and fake online reviews. In: International conference on ubiquitous information management and communication. ACM, New York, pp 1–6

Bhattarai A, Dasgupta D (2012) A self-supervised approach to comment spam detection based on content analysis. In: Nemati HR (ed) Privacy solutions and security frameworks in information protection, 1st edn. IGI Global, pp 1–253

Campbell K (2014) The importance of reviews in eCommerce.2x Consulting Limited. https://2xecommerce.com/importance-reviews-ecommerce/. Accessed 15 May 2017

Chavan A, Darekar O, Kulkarni O, Jain Y (2017) Spam reviews detection using Hadoop. Int J Eng Comput Sci 6(2):20320–20323

Christopher S, Rahulnath H (2016) Review authenticity verification using supervised learning and reviewer personality traits. In: 2016 International conference on emerging technological trends

Crawford M, Khoshgoftaar TM, Prusa JD, Richter AN, Najada HA (2015) Survey of review spam detection using machine learning techniques. J Big Data 2(1):2–23

Daiyan M, Tiwari S, Alam M (2014) Mining product reviews for spam detection using supervised technique. Int J Emerg Technol Adv Eng 4(8):619–623

Farooq S, Khanday H (2016) Opinion spam detection: a review. Int J Eng Res Dev 12(4):1–8

Fei G, Liu B, Hsu M, Castellanos M, Ghosh R, Mukherjee A (2013) Exploiting burstiness in reviews for review spammer detection. In: Proceedings of the 7th international AAAI conference on weblogs and social media. Association for the Advancement of Artificial Intelligence. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewFile/6069/6356. Accessed 03 June 2017

Feng S, Xing L, Gogar A, Choi Y (2012) Distributional footprints of deceptive product reviews. In: Proceedings of the 6th international AAAI conference on weblogs and social media. Association for the Advancement of Artificial Intelligence, pp 98–105

Hai Z, Zhao P, Cheng P, Yang P, Li X, Li G (2016) Deceptive review spam detection via exploiting task relatedness and unlabeled data. In: Proceedings of the 2016 conference on empirical methods in natural language processing. Association for Computational Linguistics, Texas, pp 1817–1826

Jain G, Manisha, Agarwal B (2016) An overview of RNN and CNN techniques for spam detection in social media. Int J Adv Res Comput Sci Softw Eng 6(10):126–132

Jindal N, Liu B (2007) Analyzing and detecting review spam. In: 7th IEEE international conference on data mining (ICDM 2007), pp 547–552

Jindal N, Liu B (2008) Opinion spam and analysis. In: WSDM. ACM, New York, pp 219–230

Jindal N, Liu B, Lim E (2010) Finding unusual review patterns using unexpected rules. In: Proceedings of the 19th ACM international conference on information and knowledge management—CIKM'10, pp 1549–1552

Kolhe N, Joshi M, Jadhav A, Abhang P (2014) Fake reviewer groups detection system. IOSR J Comput Eng 16(1):06–09

Lai C, Xu K, Lau R, Li Y, Song D (2010a) High-order concept associations mining and inferential language modeling for online review spam detection. In: 2010 IEEE international conference on data mining workshops. IEEE, Sydney, pp 1120–1127

Lai C, Xu K, Lau R, Li Y, Jing L (2010b) Toward a language modeling approach for consumer review spam detection. In: IEEE international conference on e-business engineering. IEEE, New York, pp 1–8

Lau RY, Liao SY, Kwok RCW, Xu K, Xia Y, Li Y (2011) Text mining and probabilistic language modeling for online review spam detecting. ACM Trans Manag Inf Syst 2(4):1–30

Li F, Huang M, Yang Y, Zhu X (2011) Learning to identify review spam. In: IJCAI proceedings-international joint conference on artificial intelligence, vol 22(3), pp 2488–2493

Li H, Liu B, Mukherjee A, Shao J (2014) Spotting fake reviews using positive-unlabeled learning. Comput Syst 18(3):467–475

Li H, Chen Z, Mukherjee A, Liu B, Shao J (2015) Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns. In: Proceedings of the 9th international AAAI conference on web and social media

Li H, Fei G, Wang S, Liu B, Shao W, Mukherjee A, Shao J (2016) Modeling review spam using temporal patterns and co-bursting Behaviors. Preprint. arXiv:1611.06625

Li J, Cardie C, Li S (2017) TopicSpam: a topic-model-based approach for spam detection. In: Proceedings of the 51st annual meeting of the association for computational linguistics. Association for Computational Linguistics, Bulgeria, pp 217–221

Lim EP, Nguyen VA, Jindal N, Liu B, Lauw HW (2010) Detecting product review spammers using rating behaviors. In: CIKM'10 proceedings of the 19th ACM international conference on information and knowledge management, pp 939–948

Lin Y, Zhu T, Wang X, Zhang J, Zhou A (2014) Towards online review spam detection. In: Proceedings of the 23rd international conference on world wide web—WWW'14 companion

Mukherjee A, Venkataraman V (2014) Opinion spam detection: an unsupervised approach using generative models. Technical report, UH

Mukherjee A, Liu B, Glance N (2012) Spotting fake reviewer groups in consumer reviews. In: Proceedings of the 21st international conference on world wide web—WWW'12

Mukherjee A, Venkataraman V, Glance N, Liu B (2013) What yelp fake review filter might be doing? In: 7th international AAAI conference on weblogs and social media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6006. Accessed 08 June 2017

Mushtaq S, Faisal C, Iqbal K, Shah P, Mehmood I, Akram ABU (2016) Review spam detection using sentiments and novel features. Int J Comput Sci Inf Secur (IJCSIS) 14(10):324–328

Noekhah S, Fouladfar E, Salim N, Ghorashi S, Hozhabri A (2014) A novel approach for opinion spam detection in e-commerce. In: 8th international conference on e-commerce with focus on e-trust. IEEE, Iran

Ott M, Choi Y, Cardie C, Hancock J (2011) Finding deceptive opinion spam by any stretch of the imagination. In: HLT'11 proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. Association for Computational Linguistics, Oregon, pp 309–319

Peng Q, Zhong M (2014) Detecting spam review through sentiment analysis. Int J Softw Inf 8:2065–2072

Ren Y, Ji D (2017) Neural networks for deceptive opinion spam detection: an empirical study. J Inf Sci 385:213–224

Ren Y, Zhang Y (2016) Deceptive opinion spam detection using neural network. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers, Japan, pp 140–150

Ren Y, Ji D, Hang H (2014) Positive unlabeled learning for deceptive reviews detection. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 488–498

Rout J, Singh S, Jena S, Bakshi S (2016) Deceptive review detection using labeled and unlabeled data. Multimed Tools Appl 76(3):3187–3211

Rout J, Dalmia A, Choo K, Bakshi S, Jena S (2017) Revisiting semi-supervised learning for online deceptive review detection. IEEE Access 5:1319–1327

Santosh KC, Mukherjee A (2016) On the temporal dynamics of opinion spamming: case studies on yelp. In: WWW 2016, pp 369–379

Shehnepoor S, Salehi M, Farahbakhsh R, Crespi N (2017) NetSpam: a network-based spam detection framework for reviews in online social media. IEEE Trans Inf Forensics Secur 12(7):1585–1595

Shivagangadhar K, Sagar H, Vanipriya C, Sathyan S (2015) Fraud detection in online reviews using machine learning techniques. Int J Comput Eng Res (IJCER) 5(5):52–56

Shojaee S, Murad M, Azman A, Sharef N, Nadali S (2013) Detecting deceptive reviews using lexical and syntactic features. In: 2013 13th international conference on intelligent systems design and applications, pp 53–58

Singh S (2015) Improved techniques for online review spam detection. MTech, National Institute of Technology, Odisha

Wahyuni E, Djunaidy A (2016) Fake review detection from a product review using modified method of iterative computation framework. In: The 3rd Bali international seminar on science, technology (BISSTECH 2015). EDP Sciences, pp 1–7 https://doi.org/10.1051/matecconf/20165803003. Accessed 08 June 2017

Wu G, Greene D, Cunningham P (2010a) Merging multiple criteria to identify suspicious reviews. In: Proceedings of the 4th ACM conference on recommender systems—RecSys'10

Wu G, Greene D, Smyth B, Cunningham P (2010b) Distortion as a validation criterion in the identification of suspicious reviews. In: Proceedings of the 1st workshop on social media analytics—SOMA'10

Wang W, Xie S, Liu B, Yu PS (2011) Review graph based online store review spammer detection. In: 11th IEEE international conference on data mining

Wu T, Liu S, Zhang J, Xiang Y (2017) Twitter spam detection based on deep learning. In: Proceedings of the Australasian computer science week multi-conference on—ACSW'17

Xie S, Wang G, Lin S, Yu P (2012) Review spam detection via temporal pattern discovery. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining—KDD'12

Xu Q, Zhao H (2012) Using deep linguistic features for finding deceptive opinion spam. In: Proceedings of COLING 2012: posters, pp 1341–1350

Xu Y, Shi B, Tian W, Lam W (2015) A unified model for unsupervised opinion spamming detection incorporating text generality. In: IJCAI'15 proceedings of the 24th international conference on artificial intelligence. AAAI Press, Argentina, pp 725–731

Xue H, Li F, Seo H, Pluretti R (2015) Trust-aware review spam detection. In: Trustcom/BigDataSE/ISPA, 2015 IEEE. IEEE, Helsinki, pp 726–733

Yang X (2015) One methodology for spam review detection based on review coherence metrics. In: Proceedings of 2015 international conference on intelligent computing and internet of things. IEEE, Harbin, pp 99–102. http://ieeexplore.ieee.org/document/7111547/. Accessed 01 June 2017

Zhang W, Lau R, Li C (2014) Adaptive big data analytics for deceptive review detection in online social media. In: 35th international conference on information systems. https://pdfs.semanticscholar.org/d310/e53480bc1257e7c9119fceb5e72f5ddf7229.pdf. Accessed 03 June 2017