Human-Centric Intelligent Systems (2022) 2:14–30
https://doi.org/10.1007/s44230-022-00001-3

RESEARCH ARTICLE

Online Spam Review Detection: A Survey of Literature

Li He[1] · Xianzhi Wang[1] · Hongxu Chen[1] · Guandong Xu[1]

Received: 22 November 2021 / Accepted: 22 February 2022 / Published online: 5 May 2022

Abstract
The increasingly developed online platform generates a large amount of online reviews every moment, e.g., Yelp and Amazon. Consumers gradually develop the habit of reading previous reviews before making a decision of buying or choosing various products. Online reviews play an vital part in determining consumers' purchase choices in e-commerce, yet many online reviews are intentionally created to confuse or mislead potential consumers. Moreover, driven by product reputations and merchants' profits, more and more spam reviews were inserted into online platform. This kind of reviews can be positive, negative or neutral, but they had common features: misleading consumers or damaging reputations. In the past decade, many people conducted research on detecting spam reviews using statistical or deep learning method with various datasets. In view of that, this article first introduces the task of spam online reviews detection and makes a common definition of spam reviews. Then, we comprehensively conclude the existing method and available datasets. Third, we summarize the existing network-based approaches in dealing with this task and propose some direction for future research.

Keywords Spam review detection · Machine learning · Graph convolution network · Deep learning

1 Introduction

Nowadays, the development of internet innovation makes a big difference in the way of human's life styles. Variant e-commerce websites, e.g., Yelp, Amazon and Taobao,[1] provide internet user with a convenient, efficient and relatively reliable online trading environment. More and more merchants prefer to build their virtual shops through different online platforms. Meanwhile, an increasing number of consumers gradually get used to this way of shopping, and automatically share their shop experiences and reviews by using an online review system which applied by the e-commerce website. Because most of these reviews come from the online consumers, they basically reflect the quality of product or the user experience. More and more people are accustomed to checking online reviews before placing an order to buy goods. Moreover, many merchants realize that the more positive online reviews they have, the more business transaction they have and they also can quickly expand and get high reputation.

In each online review system, the consumers' reviews have a significant influence on merchants' ranks. Some existing research works have proved that for each half-star upgrade, the restaurant's sales increased 19 percent more frequently [1], and one star increase in the online rating system will bring five to nine percent increase in revenue [2]. Unfortunately, driven by the substantial economic benefits, many malicious merchants beginning to run illegal operations, including post spam online reviews deliberately. They publish spam reviews or opinions to promote their brand reputation and attract more consumers, due to people tend to purchase such product or choose services that are frequently bought, have top ranking, and have more positive feedbacks [3]. From the BBC News, there were approximately a quarter of Yelp reviews could be spam.[2]

✉ Li He
li.he-1@student.uts.edu.au

✉ Guandong Xu
guandong.xu@uts.edu.au

Xianzhi Wang
xianzhi.wang@uts.edu.au

Hongxu Chen
hongxu.chen@uts.edu.au

[1] University of Technology Sydney, Sydney, Australia

[1] https://en.wikipedia.org/wiki/Taobao.

[2] http://www.bbc.com/news/technology-24299742.

Further, making spam online reviews has become an industrial chain, malicious merchants can easily find some professional spam review writing services online, like the Sponsored Reviews,[3] which is a site where advertisers and bloggers get in touch to write paid reviews. This deteriorating online review environment let us have to face the task of spam review detection.

Spam online reviews, which are similar to opinion spam in some certain situations, refer to violation activities, such as writing spam reviews, that try to confuse consumers with imitating real buyers. Furthermore, malicious merchants hire the real user to post spam reviews directly. Above all of these increased the difficulty of defining spam reviews. Jindal and Liu [4] make some contributions on spam review detection, and first generally proposed three kinds of definition of online reviews:

- Untruthful opinions: Those that intentionally deceive per users or conclusion mining frameworks via providing unworthy positive reviews to a lot of target objects in arrange to advance the objects and with giving unreasonable or pernicious harmful reviews to a few other objects in arrange to harm their notorieties.
- Brands Reviews only: Do not write any useful information on the items but as it were the brands, the producers or the venders of the items. In spite of the fact that they may be valuable, we consider them as spam since they are not focused on at the particular items and are regularly one-sided.
- Non-reviews: These kinds of reviews have two characteristics: (1) notices and (2) other unessential surveys containing no conclusions (e.g., questions, answers, and arbitrary writings).

Most online spam reviews can be covered by these three types up to now. The first type of reviews may cause the change of purchase decision or the negative effects on other begin merchants. Specially, this kind of spam reviews is difficult to identified. Based on these factors, many studies have applied some research work on detection untruthful reviews.

Besides, reviews on brands and advertisements also attract many researchers more interests, because these types of reviews have the potential of fraud. Some spammers may utilize the online review platform to broadcast their own illegal brand or stealing personal social contact information, or even induce consumers to conduct offline transaction, etc. These tricks bring some confusion to people, and make an enormous challenge for anti-spam framework [5]. In arrange to recognize different extortion designs, many scholars work on some adversarial tasks.

[3] http://www.sponsoredreviews.com/.

In this paper, we integrate the untruthful opinions and the fraud one comes from the other two types of reviews, and uniformly called as “spam reviews”. Spam audits are conflicting with the genuine assessment of items and attempt to deceive per users or intentionally overestimate or belittle one category things. The source of spam reviews might come from malicious merchants, individual spammers and fraud groups. Spam reviews take the form of various patterns designed by spammers [5, 6]. For occasion, the taking after has appeared two spam surveys was writing to Amazon review platform, which was identified with a model survey spam discovery framework [7]. After observing from “Review 1”, it is troublesome for human per users to decide whether the audit is spam or generous. Fortunately, on the off chance that a per user finds these two reviews at the same time, he/she will be able to capture the fundamental spam include to classify these two as spam reviews, due to both of them have settled semantic design almost diverse items. Clearly, the manual approaches of identifying spam audits are not attainable for this event.

- Review 1: I did broad investigate some time recently selecting the SD60D, and I am excited with my buy. This camera is modest (littler than my iPod) and lightweight, but still takes extraordinary picture. The screen is much bigger than my friends’cameras, and it has all the additional settings that the normal individual should take incredible photographs is all sorts of conditions, I have not had any terrible or hazy pictures with it however. I am excited with this camera and would suggest it to everybody.
- Review 2: I did broad investigate some time recently selecting the Kodak EasyShare C875, and I am excited with my buy. This camera takes extraordinary picture. The screen is much bigger than my friends’cameras, and it has all the additional settings that the normal individual has to take extraordinary photographs is all sorts of conditions, I have not had any terrible or hazy pictures with it however. I am excited with this camera and would suggest it to everybody.

The number of online spam reviews has increased year by year. From existing insights, spam reviews for around 2–6% at Proceline, Orbitz, Tripadivisor and Expedia [8, 9]. Especially, in online review platform, e.g., Yelp, the proportion of spam reviews is already up to 14–20% [10]. It is very urgent to build effective detection framework to identify spam reviews automatically. Until now, there are so many state-of-the-art methods proposed driven by the detection task. The main challenge of this kind of task has three folds as follows:

- How to recognize a set of linguistic features from spam reviews?
- How to deal with the lack of labelled reviews datasets?
- How to utilize the relationship between products, consumers and reviews?

This research direction has attracted a lot of research attentions [4, 11–13]. Among them, they are mainly focused on basic language models which do not consider deep and relational information. Deep learning models have broadly been connected into numerous NLP assignments. Compared with conventional measurable models, new methods (e.g., deep neural networks and graph based methods) create a large space for new researches [14–16]. Recently, a few survey works have been published, there are three works to summarize the existing method for the spam review detection [5, 17, 18]. However, these three works have several shortages. First, they do not systematic summarize the labelled datasets and verify the availability of their listed data source. Secondly, they lack of the conclusion of graph-based technique, especially the rapidly graph convolution network method developed in recent years. Third, they fail to give a complete task classification to cover existing methods. To address these issues, we focus on three aspects to systematically summarize previous research works: existing method and available datasets, and provide some suggestions for future research. Especially, we will disentangle the graph-based strategies that have been proposed to unravel the issue of spam review discovery.

Our work first defines the mainly task of spam review detection. Then we present the existing state-of-art approaches, including four types of directions, such as feature engineering, traditional statistical models, neural network models and graph networks frameworks. In addition, we summarize some existing data resources and their data structure. Finally, we provide some construction research direction for the future.

2 Task Definitions and Concepts

Most of the existing spam review detection researches are driven by different task. Some researches are on the spam reviews themselves, and others are dedicated to identifying spammers or groups. From our selected research works, we identified tasks for detection of four categories: review mining task, end-to-end classification, cold-start problem respectively and spammer detection task, as shown in Fig. 1.

2.1 Review Mining Task

Review mining is a process of extracting linguistic features and context from online review platform, utilizing statistical models and feature engineering evaluates customers'opinions, and learning reviewer behavior to identify the spam reviews. It is generally modeled as a rule-based system.

Fig. 1 Distribution of Focused Research Works. Review Mining Task takes the percentage of 17%, end-to-end classification task is 22%, and cold-start problems have the same proportion as classification task. Spammer detection task takes the percentage of 39%

KL Divergence: Kullback–Leibler divergence is broadly utilized to assess the distance between two likelihood disseminations. Lai et al. [19] utilized KL divergence to degree the separate between the combine of language models. Reference [19], which denoted as M_{d_1} and M_{d_2}. They proposed untruthful review spam detection strategy which is supported by KL divergence and is characterized by:

$$\mathrm{KL}\left(M_{d_1} \| M_{d_2}\right) = \sum_{t \in \{d_1 \cup d_2\}} P\left(t \mid M_{d_1}\right) \times \log_2 \frac{P\left(t \mid M_{d_1}\right)}{P\left(t \mid M_{d_2}\right)} \quad (1)$$

Then, they applied Joachim's SVM package[4] for the classifier modules.

MI Measures: Mutual Data degree has been utilized in collocational investigation in existing inquire about works. Mutual Data is an information-theoretic strategy for computing the reliance between two substances:

$$MI\left(t_i, t_j\right) = \log_2 \frac{\Pr\left(t_i, t_j\right)}{\Pr\left(t_i \Pr\left(t_j\right)\right.} \quad (2)$$

Lau et al. [7] proposed a content mining show with a adjusted shared data degree for the location of untruthful reviews. They consider both term nearness and term nonappearance as prove for assessing the quality of the affiliation between a concept and its basic clear terms.

[4] http://svmlight.joachims.org/.

Table 1 Description of features in CATS

Feature name	Comments
Avg(Positive)	Average number of positive words
Avg(Pos/Neg)	Difference between positive and negative words counts
Ratio(Unique/All)	Ratio between unique and the overall words counts
Avg(Sentiment)	Average sentiment of reviews
Avg(Entropy)	Average entropy of reviews
Avg(Length)	Average length of reviews
Sum(Length)	Sum length of reviews
Sum(Punctuation)	Number of punctuations
Ratio(Punctuation)	Average ratio of punctuations
Avg(Ngram)	Average counts of positive Ngrams
Ratio(Ngram)	Average ratio of positive Ngrams

Multicriteria Decision: Viviani et al. [20] used a multicriteria choice making approach based both on the evaluation of numerous criteria and the utilize of accumulation administrators with the point of getting a veracity score related with each review. Based on this score, it is conceivable to identify spam reviews [20]. Specifically, they provide a definition of aggregation operator F:

$$F : I^n \to I \tag{3}$$

Then their aggregated value is calculated as follows:

$$F(a_1, a_2, \ldots, a_n) = \sum_{j=i}^{n} w_j b_j \tag{4}$$

Cross-Platform E-Commerce Fraud Detection: Weng et al. [13] distinguished a bunch of platform-independent include from the world level Table 1, the semantic level and the basic level to separate extortion and typical things on diverse e-commerce stages.

Advance, CATS did an execution comparison try to choose the leading one from the six commonly utilized models: AdaBoost, SVM, Xgboost, Neural Organize, Decision Tree and Naive Bayes. The Xgboost show appears way better execution than other benchmark and have the capability of recognizing more features that can segregate whether an thing is false or ordinary.

2.2 End-to-End Classification

Many research works aim to build an end-to-end classifier to detect spam reviews. They took advantage of word embedding as the input of their non-linear or linear classifier. This kind of task can be subdivided into two directions: text classification and graph classification. More specifically, text classification usually modeled as two-class classification task or binary classification task and graph-based problem contained node, edge and sub-graph classification task.

LR & RNNLM: Fontanarava [21] proposed two classifiers on linguistic features and another classifier on metadata and behavioral characters. They evaluated their models on the datasets from Yelp.com site. Particularly, they are based on Logistic Regression (LR) connected to literary features within the frame ofweighted bag-of-words, and on a generative model utilizing Repetitive Neural Network based Dialect Models (RNNLM) separately.

Graph-based Model: Recent years, there is a growing interest in constructing graph relationship among nodes and edges. Noekhah [22] built a graph-based model with three entities: review, reviewer and target. Then, they designed a algorithm to update corresponding spamicity degree of each entity iteratively to determine whether an entity is spam or not.

$$\text{Spamicity} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} f_i \times w_i}{n \times m} \tag{5}$$

Bipartite Graph: Yang [23] used a bipartite graph with popular ranking algorithm to detect spam reviews. In this work, they focused on measuring the coherence of a review based on two major flow smoothness information among sentences: Word transition probability and Word concurrence probability. Specially, they defined one step transition probability as follows:

$$H(r) = \exp\left\{ -\frac{\sum_{i=1}^{n-1} \log\left(P(s_i \to s_{i+1})\right)}{\sum_{i=1}^{n-1} |s_{i+1}|} \right\} \tag{6}$$

At that point, the word concurrence metric for word w_i and w_j as $\log(O_{i,j}/O_i O_j)$. Encourage, the esteem of coherent metric for spam reviews is frequently lower than that of honest reviews. So this coherent metric is additionally exceptionally supportive to distinguish spam reviews and it can be mutually utilized with the previous one.

$$\text{Con}(r) = \frac{1}{n-1} \sum_{i=1}^{n-1} \text{Con}(s_i, s_{i+1}) \tag{7}$$

GCN: Li [6] first applied graph convolution network to the spam review detection problem. The GCN-based strategies take after a layer-wise engendering way. In each proliferation layer, all the nodes upgrade at the same time. In common, a GCN based show can be composed as:

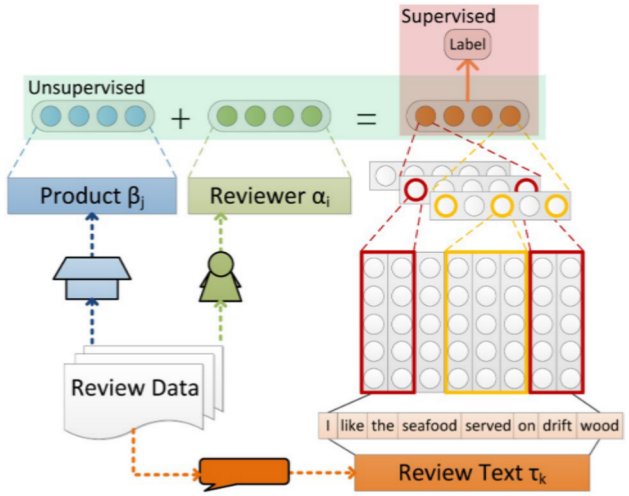

Fig. 2 Wang et al., model overview. They take the products items as the head part of the TransE network in their model, take the reviewers as the translation (relation) part and take the review as the tail part

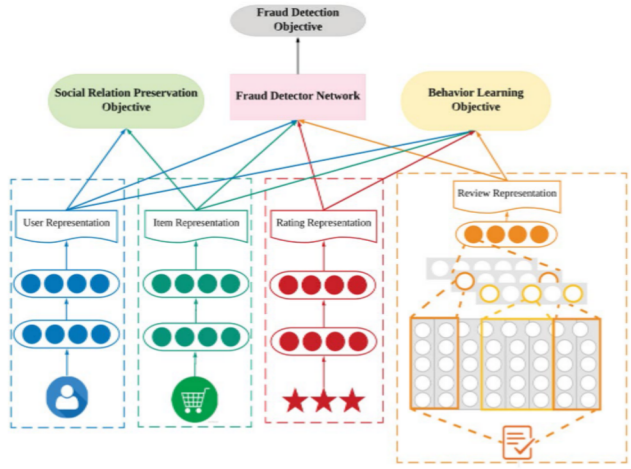

Fig. 3 Li et al., model overview. The embedding network consists four parts: user embedding layers, item embedding layers, review embedding networks, and rating embedding layers

$$h^l_{N(v)} = \sigma\left(W^l \cdot AGG\left(\left\{h^{l-1}_{v'}, \forall v' \in N(v)\right\}\right)\right)$$
$$h^l_v = COMBINE\left(h^{l-1}_v, h^l_{N(v)}\right) \tag{8}$$

Moverover, they proposed a heterogeneous graph to represent the interaction between products and users, and formulated to an edge classification problem with attributed nodes and edges.

2.3 Cold-Start Problem

Due to the lack of the labelled data and negative samples, some scholars start to deal with zero-shot learning issues. Moreover, fathoming the cold-start issue in survey spam discovery can offer assistance the online review websites to calm the harm of spammers in time. They mainly focus on positive labelled or unlabeled learning.

PU-learning: PU-learning learns from positive and unlabeled information, where P denotes a set of positive datasets and U a set of unlabeled datasets. The goal is to build a classifier using P and U to classify the data in U or a future test set T. Li and Liu [24] proposed the method of learning from positive and unlabeled illustrations (or PU-learning).

TransE: TransE may be a demonstration which can encode the arrange structure and speak to the nodes and edges with the triple frame connection and tail in low-level measurement vector space. TransE has been demonstrated that it is nice at depicting the worldwide data of the chart structure by the work approximately distributional representation for information base. Therefore, Wang et al. [16] utilized TransE to encode the literary and behavioral data into the review embeddings for the cold-start spam location assignment. As appeared in Fig. 2, they take the items β as the head portion of the TransE arrange in their model, take the reviewers α as the connection portion and review content embeddings which learnt by the CNN τ as the tail portion.

$$S' = \left\{(\boldsymbol{\beta}', \boldsymbol{\alpha}, \boldsymbol{\tau}) \mid \boldsymbol{\beta}' \in B\right\} \cup \left\{(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\tau}') \mid \boldsymbol{\tau}' \in T\right\} \tag{9}$$

Attribute Enhanced Domain Adaptive Model: Inspired by the previous work Wang et al. [16], You and Qian [25] presented a neural network encoding the attributes, entities, and their relations, and leveraged the abundant information to alleviate data scarcity problem in the cold-start scenario of spam review detection.

JESTER: Li [26] proposed a user-item-review-rating representation for fraud reviews detection, they embedded the user-item social relations into their models to tackle cold-start problem. JESTER simultaneously considers three tasks: client looking into behavior learning, social connection conservation, and extortion review discovery, comparing to three learning misfortune capacities: behavior learning misfortune, social connection conservation misfortune, and extortion discovery misfortune. By mutually optimizing these three misfortune capacities, Entertainer learns the user-item-review-rating inferable representations for extortion survey discovery. A toy example of the architecture of the above model is shown in Fig. 3

2.4 Spammer Detection Task

The semantic and context analysis of online review content, as well as the deep learning method, can well capture

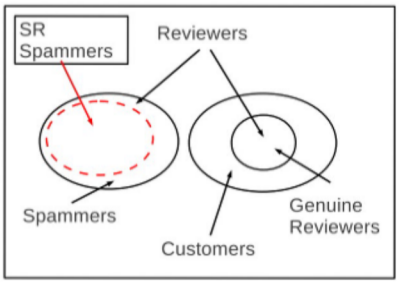

Fig. 4 Xie et al., model overview. Relationship of Spammers, Reviewers and Customers

the explicit and implicit features in language level from the review text only, but it will always have accuracy problems and affect the efficiency of identifying spam reviews. Therefore, increasing number of research works have devoted to the behavior analysis from the review posters, which can improve the detection performance. Online review spammer detection is a comprehensive analysis for the review content and the review posters.

SR Spammer: Xie et al. [27] focused on singleton review spammer (SR) detection and mapped the SR spammer detection problem to an abnoramlly correlated temporal pattern detection problem. They explored the relationship

Table 2 Mukherjee et al. indicators

Indicators	Models
Group time window	Demonstrating the degree of dynamic inclusion of a gather as its bunch time window $GTW_P(g,p) = \begin{cases} 0 & \text{if } L(g,p) - F(g,p) > \tau \\ 1 - \frac{L(g,p)-F(g,p)}{\tau} & \text{otherwise} \end{cases}$
Group deviation	Group deviation (GD) on a higher rating scale $D(g,p) = \frac{\lvert r_{p,g} - \bar{r}_{p,g} \rvert}{4}$
Group content similarity	Group member information similarity (GMCS): $CS_G(g,p) = \underset{m_i,m_j \in g, i<j}{\text{avg}} \left(\text{cosine}\left(c(m_i,p), c(m_j,p)\right)\right)$
Group early time frame	Group early time frame (GETF) models: $GTF(g,p) = \begin{cases} 0 & \text{if } L(g,p) - A(p) > \beta \\ 1 - \frac{L(g,p)-A(p)}{\beta} & \text{otherwise} \end{cases}$
Group size ratio	Group size to the total number of reviewers for each product $GSR(g) = \underset{p \in P_g}{\text{avg}} \left(GSR_P(g,p)\right) \quad GSR_P(g,p) = \frac{\lvert g \rvert}{\lvert M_p \rvert}$
Group size	Normalize it between 0 and 1. Maximum ($\lvert g_i \rvert$) among all discovered groups $GS(g) = \frac{\lvert g \rvert}{\max\left(\lvert g_i \rvert\right)}$
Group support count	Support count of a group is the overall number of products towards which the group has worked together $GSUP(g) = \frac{\lvert P_g \rvert}{\max\left(\lvert P_{g_i} \rvert\right)}$
Individual rating deviation	Like group deviation and model IRD as $IRD(m,p) = \frac{\lvert r_{p,m} - \bar{r}_{p,m} \rvert}{4}$
Individual content similarity	Modelling this behavior of a reviewer m across all its reviews according to each product p $\text{ICS}(m,p) = \text{avg}(\text{cosine}(c(m,p))$
Individual early time frame	Define a group member m as $IETF(m,p) = \begin{cases} 0 & \text{if } L(m,p) - A(p) > \beta \\ 1 - \frac{L(m,p)-A(p)}{\beta} & \text{otherwise} \end{cases}$
Individual member coupling in a group	Behavior measures how closely a member works with the other members of the group $\text{IMC}(g,m) = \text{avg}_{p \in P_g} \left(\frac{\lvert (T(m,p) - F(g,p)) - \text{avg}(g,m) \rvert}{L(g,p) - F(g,p)} \right)$

of reviewers, customers, spammers and SR spammers as shown in Fig. 4.

Group Spam Behavior Indicators: Mukherjee et al. [11] to begin with utilized a visit thing set mining strategy to discover a set of spam analyst groups. They utilized a few behavioral models inferred from the collaboration wonder among spam commentators and connection models based on the connections among bunches, people, and things they looked into to identify spammer bunches as appeared in Table 2.

Reviewer Trustiness: For further learning the relationship among reviews, item and reviewers, Wang et al. [28, 29] provided a review graph to recognize suspicious reviewers. They captured the connections by presenting three essential concepts, the trustiness of analysts, the trustworthiness of surveys, and the unwavering quality of stores. Uncommonly, the trustiness of a reviewer r, which denoted as $T(r)$ is a score of how much we can believe r. The author, in arrange to illuminate these relations gives the common frame of trustiness work as take after:

$$T(r) = \frac{K}{1 + e^{-KH_r}} \tag{10}$$

SRD-BM & SRD-LM: As of late, Hussain et al. [30] worked on this errand and utilizes thirteen distinctive spammer's behavioral features to calculate the review spam score which is at that point utilized to distinguish spammers and spam reviews. In the mean time, they utilized Linguistic Strategy (SRD-LM) works on the substance of the reviews and utilized change, include determination and classification to distinguish the spam reviews.

The system of SRD-BM begins with the recognizable proof and calculation of spammer behavioral features in unlabelled Amazon survey dataset. This demonstrate executes in four steps: (1) Calculating the normalized esteem of each spammer behavioral feature. (2) Computing the cruel score for each review and the in general precision of the total datasets. (3) Surveying the affect of each behavioral include and relegates a weight agreeing to the significance of each behavioral include. (4) Calculating spam score and distinguishes spam and not-spam surveys utilizing distinctive limit values.

In addition, generative adversarial models become more and more popular recently, because of they can generate negative labelled training data automatically and gradually enhance its discriminator. Aghakhani [31] first used generative adversarial network (GAN) to generate better double discriminator model for detecting deceptive fraudulent reviews. Zheng et al. [32] created one-class antagonistic nets (OCAN) for extortion discovery with as it were generous clients as preparing information.

3 Techniques of Spam Review Detection

Over the past decade, a developing number of analysts have worked to discover superior strategies to identify spam reviews. Some scholars have tried using traditional statistical methods to learn different aspects text features from large-scale datasets, and continually optimize evaluate results with various feature extractors. In addition, other people begin to utilize machine learning approaches to boost their detection frameworks. Specially, as more and more CNN/RNN methods are applied in the field of Natural Language Processing (NLP), researchers also propose some representative neural models to solve spam review detection problems. Due to the dependence among reviews, products and users, graph-based method is used to capture a set of features. This part will be introduced in Sect. 3.4.

3.1 Traditional Statistical Methods

From the perspective of existing investigate works, conventional methods need to extract various features from reviews and usually presented as a language model. Due to the spam reviews have some identified attributes, feature engineering is very necessary for statistical models. Further, the spam review writers could change the form of comments, and the detection models also need to adjust constantly. Previous researchers using various feature engineering are summarized as follows.

Jindal and Liu [4] first identified three types of spam reviews, and they then analyzed real-world datasets from Amazon. Through the statistic of datasets, they found a expansive number of copy and near-duplicate reviews. Then, they utilized 2-gram to calculate the closeness score of two reviews and review sets with closeness score of at slightest 90% were chosen as copies [33]. However, duplicates can only detect spam type reviews, Jindal et al. characterized a huge set of features to characterize reviews, totally up to thirty-five features, such as length of the review title, price of the product and so on. They isolated these features into three categories: review centric features, reviewer centric features and product centric features. For the model building, they used logistical regression with statistical package R.

Further, Mukherjee [11] proposed method first using several behavioral models derived from collusion phenomenon among spam reviewers to detect spam reviews as shown in Table 1.

Then, Mukherjee and Venkataraman [34] at that point considered an extra set of behavioral features around commentators and their review for learning, which significantly make strides the classification result on real world reviews

datasets. For the behavioral study, they crawled profiles of all reviewers in their hotel and restaurant datasets and proposed eight behavioral features:

- Activity window (AW): The difference of timestamps of the last and first reviews for that reviewer.
- Maximum number of reviews (MNR): Due to 35.1% of spammers posted reviews in one day, the MNR per day is a suitable feature.
- Review count (RC): The number of audits that a commentator has. This feature appears a clear division of spammers from non-spammers based on their looking into exercises.
- Percentage of positive reviews (PR): The percentage of positive reviews, which have got 4 or 5 stars.
- Review length (RL): The average number of words per review for each reviewer.
- Reviewer deviation (RD): The sum that spammers veer off from the common rating agreement. They compute the supreme rating deviation of a review from other surveys on the same commerce.
- Maximum content similarity (MCS): The creator computes the most extreme substance closeness based on cosine closeness between any two reviews of a analyst.
- Tip count (TC): "Tip" function is unique to Yelp website which is a short (less than 140 charactors) descriptions and insights about a business.

In addition, Alom et al. [35] conducted spam reviews detection research work on Twitter website. They made use of several new features, which were more effective and robust than existing used features. Specially, some graph-based features was put forward: (a) triangle count of user's network: Triangle_Count(u) (b) the ratio of triangle count to number of followers of user:

$$\text{RateTNF}(u) = \frac{\text{Triangle\_Count}(u)}{N_{fer}(u)}$$

(c) the ratio of bidirectional links from the users' social network:

$$\text{Rate}_{\text{bilink}}(u) = \frac{N_{\text{bilink}}(u)}{N_{fer}(u) + N_{\text{fing}}(u)}$$

Meanwhile, Myo Myo Swe [36] proposed a modern and vigorous boycott creation for recognizing spam accounts on Twitter was proposed. The boycott was made utilizing LDA and TF-IDF strategies. In this work, there are fourteen content-based features proposed that can distinguish fake accounts from legitimate accounts, such as spam words ratio: $ratio = \frac{CountsOfSpamWords}{TotalNumberOfWords}$, hashtag ratio: $ratio =$ and so on.

Table 3 Comparison of supervised methods in previous work

Methods	Metrics	Performance	Datasets	Reference
SVM	F-measure	78.1%	Yelp etc	[38]
LIBSVM	Accuracy	89.6%	Yelp etc	[8]
NB-SVM	Accuracy	91.87%	IMDB	[39]
WMUSVM	Recall	82.5%	TripAdvisor	[40]
BERT	Accuracy	90.5%	Yelp	[41, 42]
RF	Accuracy	97%	Yelp	[43]
LR	F-score	92.6%	Yelp	[44]
XGBoost	Precision	99%	Yelp	[45]

Jia [37] used linguistic feature, which respectively aims to term frequency, Latent Dirichlet Allocation (LDA) and word2vec, then merged into one model to conduct experiment. In details, they utilized topic modeling technique in natural language processing and extracted hidden topics from Yelp datasets. This work has extreacted five topics and each topic contains eight words in terms of fake review as follows:

- Topic1: promise, quality, pushy, peeled, rationalize, podium, decorated and gulped.
- Topic2: care, shots, spray, cliff, ramps, edge, comments and park.
- Topic3: swirling, settle, breadth, strict, eavesdrop, split, discarded and stones.
- Topic4: writing, reserve, injure, damn, autographed, hate, olfactory and zealand.
- Topic5: cube, parings, shined, pomp, bamboo, heroin, absurd and unsalted.

Recently, Weng [13] from Alibaba Group, summarized eleven stage free features from the word level, the semantic level and the basic level to separate extortion and typical things. They had developed a cross-platform e-commerce fraud detection system, called CATS. Additionally, they selected Xgboost model as a binary classifier, and their evaluation results indicated that CATS achieves both high precision and recall [13]. According to their feature engineering, this research work have identified several features as shown in Table 1. These above features have achieved good results in cross-platform spam detection.

Another statistical based method presented by Lai et al. [19] successfully carried out review spam detection for untruthful review detection with an unsupervised probabilistic language model, for non-review detection with a supervised classifier. They outlined the probabilistic language modeling and Kullback–Leibler (KL) divergence based strategy for the discovery of untruthful reviews, and the Support Vector Machine (SVM) based approaches for the detection of non-reviews. LAU et al. [7] also proposed a content

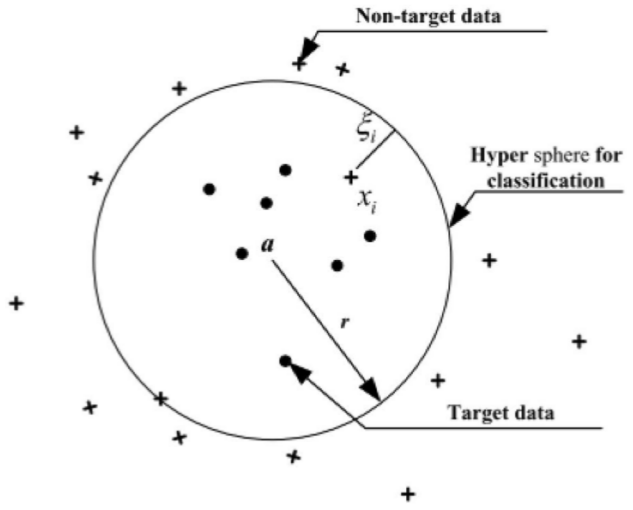

Fig. 5 Yang et al., WMUSVM Algorithm Overview. This demonstrate builds up hypersphere with greatest volume, containing all beguiling reviews information, and all genuine reviews information are exterior of this hypersphere

mining model is created and coordinates into a semantic language model for the detection of untruthful reviews.

3.2 Machine Learning Approaches

Machine learning approaches have been broadly utilized in handling spam reviews discovery. Most of existing investigate works can be classified three directions: supervised learning, semi-supervised learning and unsupervised learning (Table 3).

3.2.1 Supervised Learning Model

Supervised learning model relies on a large amount of labelled datasets. Ott et al. [8, 46] collected a balanced datasets of TripAdvisor reviews and trained a supervised deceptive classifier based on the labelled datasets. They proposed a common system for assessing the predominance of double dealing in online review communities and discovered a Bayesian based model as a classifier that distinguished truthful from deceptive reviews. They further compared with Naïve model and Bayesian model, and found that Bayesian Predominance Demonstrate (bayes) addresses Naïve strategy confinements by modeling the generative prepare through the joint likelihood dissemination of the watched and inactive information. Their proposed prevalence models is denoted as:

$$\pi^* = \frac{1}{N^{\text{test}}} \sum_{i=1}^{N^{\text{test}}} y_i \tag{11}$$

Inspired by previous work that Support Vector Machines (SVM) trained on n-gram features perform well in spam detection tasks, this work trained linear SVM classifiers using the LIBSVM software package, and represented reviews using unigram and bi-gram bag-of-words (BoW) features.

Mesnil [39] used a supervised reweighing of the counts as in the Naïve Bayes Support Vector Machines method and achieved strong results on a dataset of IMDB movies reviews.

Siagian [38] abused work words as a imminent include and combined with character n-grams as an input include for recognizing beguiling and honest review.

Yang [40] proposed an unbalanced support vector machine to deal with the lack of manual labeled deceptive reviews. WMUSVM model was first proposed in this paper based on hypersphere with maximum volume, containing all deceptive reviews data, and all true reviews data are outside of this hypersphere, as shown in Fig. 5.

Kennedy et al. [41, 42] utilized the BERT model for pre-training their word embeddings. So far, most well-known machine learning methods were used as benchmark classifiers, such as Naïve Bayes, neural network and support vector machine [47–49]. Barushka et al. [47] demonstrated the central importance of text preprocessing strategies in detecting spam reviews among these three methods. The experiment result indicated that number and length of the extracted word segments had major effect on the performance of the classifiers.

A random forests method was chosen by Nilizadeh [43] as a classifier. This was since of its value in a wide assortment of applications, its resistance to over-fitting, and its utility in understanding feature significance [50, 51].

Tingxuan et al. [44] used under-sampling and over-sampling techniques to expend training datasets for imbalance learning. An implementation of gradient boosted decision trees designed by Sihombing [45], which aimed to build a deep learning model that can detect spam and non-spam reviews on datases come fomr Yelp.com.

3.2.2 Semi-Supervised Learning Model

Semi-supervised learning models combine many labeled information and a huge number of unlabelled information to prepare a classifier for the discovery of spam reviews.

Most of the previous research works focus on proposing a novel angle to the problem by modeling positive unlabelled (PU) learning. PU learning generally has two classes of framework:

- Constructing a classifier by using positive sample dataset and some samples of the unlabeled dataset.
- Learning a classifier by using positive sample dataset and the full unlabelled dataset.

Further, PU learning aims to build the major classifiers using positive and unlabelled samples with four steps:

- Extracting the reliable negative samples.
- Calculating the representative positive and negative samples.
- Generating the similarity weights for spam samples
- Building the major SVM-based classifier.

Li and Liu (2014) first reported a supervised learning study of two classes, spam and unknown with a Chinese review dataset from Dianping.[5] They focused on using text content, since it can detect spam reviews right after posting and spam reviews thus has less damage. Further, they used Support Vector Machines (SVM) and Positive and Unlabeled learning (PU) to detect spam reviews [52]. Li and Chen [52] moreover utilized Dianping's online review datasets to examine the basic instrument of supposition spamming and perform a supervised learning on the double classification errand. They utilized the perplexing conditions among reviews, clients and IP address to propose collective classification calculation called Multi-typed Heterogeneous Collective Classification (MHCC) and after that amplified it to Collective Positive and Unlabelled learning (CPU). Ren et al. [53] created a blending populace and person property PU learning (MPIPUL) show. For this work, the PU learning was proposed based on Latent Dirichlet Assignment (LDA) and SVM.

Hai et al. [54] proposed a semi-supervised different errand learning strategy through Laplacian regularized calculated relapse to boost the review spam discovery capability. Wu et al. [55] made use of both labeled and unlabelled data to conduct a semi-supervised learning model based on Bayesian inference. A semi-supervised learning system, named SPR2EP (SPam Survey REPresentation), built by utilizing report (review) and hub (reviewer and item) embeddings independently [56]. They assessment comes about appeared that demonstrate that was built by utilizing the combined include vectors accomplish superior execution.

3.2.3 Unsupervised Learning Model

Unsupervised learning model only utilizes a set of unlabeled data to discover the potential relationship among reviews. The existing research works refer to utilizing Generative Adversarial Nets (GAN) to generate spam samples from the original labelled samples and enhance further training process. In general, GAN is formalized as a minimax game with the value function:

$$V(G,D) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{Z}\sim p_z}[\log(1 - D(G(\mathbf{z})))] \tag{12}$$

Due to the difficulty of manually labelling, Zheng et al. [32] applied a one-class classification method to solve the lack of labelled spam reviews datasets. For this work, they only learned the representations of benign reviewers with LSTM-Autoencoder method, named OCAN [32]. Interestingly, OCAN generated many complementary samples of benign reviewers and boosted the capability of discriminator, and then the generator kept trying to make the discriminator fail to identify. After this self-enhancement processing, the detection model can adaptively update a text representation once the reviewer commits a new comment and predict whether the review was a spam or non-spam.

3.3 Neural Networks Models

Neural networks methods, also known as deep neural networks, have been broadly utilized within the field of computer vision, such as Convolutional Neural network (CNN) and represented the sequential information, such as Long Short-Term Memory (LSTM) or Recursive Neural Network (RNN).

Wang et al. [12] endeavored to utilize Long Short-Term Memory Repetitive Neural Organize system to distinguish spamming reviews. They established three types of layers to predict spam reviews, the input layer for receiving data, hidden layer of LSTM and output layer respectively. Liu and Jing [57] explored with bidirectional long short-term Memory (BiLSTMWF) to study document embeddings for detecting deceptive reviews. They formulated the spam reviews detection task as a two-class classification problem and then added the feature embeddings to BiLSTMWF model by aggregating the feature representation. Generally, the BiLSTM neural networks consist of input gate, forget gate and output gate, which these following equations as:

$$\begin{aligned} i_t &= \sigma\left(W_{mi}m_t + P_{mi}h_{t-1} + b_i\right) \\ f_t &= \sigma\left(W_{mf}m_t + P_{mf}h_{t-1} + b_f\right) \\ o_t &= \sigma\left(W_{mo}m_t + P_{mo}h_{t-1} + b_o\right) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tanh\left(W_{mc}m_t + P_{mc}h_{t-1} + b_c\right) \\ h_t &= o_t \odot \tanh\left(c_t\right) \end{aligned} \tag{13}$$

Barushka [58] used deep feed-forward neural network (DNN) to handle the high-dimensional feature representation and classified online reviews into spam and legitimate categories. A CNN model was developed by Archchitha

[5] http://www.dianping.com/.

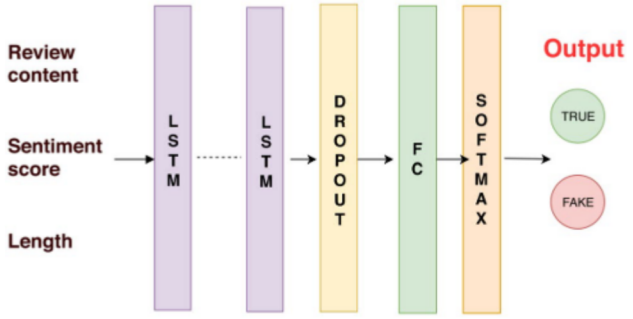

Fig. 6 Nayak et al., review indicator framework. They demonstrate the review spamicity expectation as a two-class classification issue. The review spamicity pointer was built utilizing stacked LSTM models and show yields the individual probabilities of each review being genuine or fakes

et al. [59] to detect opinion spam. They mapped highlight tokens to a particular inserting utilizing Worldwide Vectors for Word Representation demonstrate (GloVe) pre-trained word implanting demonstrate and developed their CNN with three parallel convolution layers with diverse channel sizes.

Yuan et al. [60] designed a hierarchical fusion attention network to facilitate learning semantic representations from reviewers and items level. They considered a reviewer may post several reviews and used TransH to encode the relationship among reviewer, reviews and products [61]. Additionally, they evaluated their models, named HFAN, on four public spam reviews datasets, such as Mobile01 Reviews, YepCHI, YelpNYC and YelpZip, and the neural networks based models outperformed feature based methods. This work designed some major components for HFAN model: (1) To capture the user-related semantic features of the review at word level, they design the multi-attention unit (MAU). The MAU was a attention mechanism to summarize the local context matrix to extract user-related words, which was denoted as follows:

$$\begin{aligned} \boldsymbol{v}_j^{(.)} &= \sum_{t=1}^{2r+1} \alpha_t^{(.)} \boldsymbol{X} \\ \alpha_t^{(.)} &= \frac{\exp\left(\boldsymbol{u}_t^{(.)}\right)}{\sum_{k=1}^{2r+1} \exp\left(\boldsymbol{u}_k^{(.)}\right)} \\ \boldsymbol{u}^{(.)} &= \tanh\left(\boldsymbol{X}\boldsymbol{W}_x^{(.)} + \boldsymbol{U}_j \boldsymbol{W}_u^{(.)}\right) \end{aligned} \tag{14}$$

(2) To obtain the sentence representation, this paper used linear layer and max pooling on sentence matrix:

$$\begin{aligned} \boldsymbol{S}_i &= \tanh\left(\boldsymbol{V}_i \boldsymbol{W}_v + \boldsymbol{b}\right) \\ \boldsymbol{s}_i^u &= \max_{dim=1}\left(\boldsymbol{S}_i\right) \end{aligned} \tag{15}$$

(3) They utilized dot-product attention to calculate the fusion matrix to build the relationship between the two reviews matrices. A TransH based model applied to model user-review-product relationship and defined as:

$$l(\boldsymbol{u},\boldsymbol{d},\boldsymbol{p}) = \left\| \left(\boldsymbol{u} - \boldsymbol{w}_d^T \boldsymbol{u} \boldsymbol{w}_d\right) + \boldsymbol{d} - \left(\boldsymbol{p} - \boldsymbol{w}_d^T \boldsymbol{p} \boldsymbol{w}_d\right) \right\|_2^2 \tag{16}$$

(4) The fully connected layers are proposed and *softmax*(.) function is used to convert the output values:

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{W}_c\left(\text{relu}\left(\boldsymbol{d}_j \boldsymbol{W}_d + \boldsymbol{b}_d\right)\right) \\ p_i\left(c \mid \boldsymbol{u}_k, \boldsymbol{d}_j, \boldsymbol{p}_i; \theta\right) &= \frac{\exp\left(\boldsymbol{y}_i\right)}{\sum_{k=1}^{c} \exp\left(\boldsymbol{y}_k\right)} \end{aligned} \tag{17}$$

A DeepSpot proposed by Nayak [62] to recognize spam and non-spam reviews based on the real-world reviews and the generated reviews. The DeepSpot applied three well-known supervised learning algorithms for text classification, such as support vector machines, naive bayes, and random forest. Further, they prepared an encoder-decoder system as the reviews generator utilizing Bidirectional LSTM with word embeddings. The change of exhibitions assist demonstrated the thought that the neural arrange can capture more complex setting data that was troublesome to extricate utilizing conventional discrete manual features [63]. Specifically, the spam review indicator in DeepSpot was built by stacked LSTM models and outputs the prediction of each review being spam or non-spam. The proposed architecture of the spam review indicator as shown in Fig. 6.

3.4 Graph-based Methods

The graph-based algorithm has been widely used in representation learning on networks, such as the social network and knowledge graph. Recent years, the researchers gradually realize that above feature-based methods ignore the relationship among reviews, reviewers and products. However, under some circumstances, the connection between different objects plays an important role in spam review detection. For this reason, some researchers began to apply the graph-based method to capture text features among different entities. Due to the motivation of graph embedding, existing works are focused mainly on graph neural networks and graph convolutional networks.

3.4.1 Graph Neural Networks

Graph neural networks (GNNs) are deep learning based methods that operate on homogeneous or heterogeneous graphs [64]. Machine learning assignments in GNN can be classified into hub classification for foreseeing a sort of a given hub, connect expectation for foreseeing whether two hubs are connected, community location to recognize thickly connected clusters of hubs and organize closeness for assessing the degree of two systems.

Review Graph: The primary GNN-based spam review location strategy was proposed by Wang et al. [28]. They built

a heterogeneous "review graph" to speak to the relationship among analysts, surveys and stores [28]. This was the primary time a chart demonstrate with three sorts of nodes that have been utilized to capture spam survey clues. Each node was joined with a set of highlights. For occasion, a store node had highlights approximately its number in case reviews, its rank rating, etc. They encourage proposed three crucial concepts to recognize diverse substances, i.e. the trustiness of commentators, the trustworthiness of reviews, and the unwavering quality of stores. Wang too created an iterative computation system (ICF) to compute unwavering quality, trustiness and genuineness, by investigating the inter-relationship among them.

Spamicity Degree: Noekhah [22] first proposed "Spamicity" concept to define what extent the entity is spam. First, they extracted proper and efficient features from Amazon datasets, and then designed an effective learning algorithm to update corresponding "Spamicity" degree of each entity iteratively [22]. Further, Noekhah updated entity "Spamicity" degree iteratively based on their features and the values from last iteration and utilized final value to distinguish whether an entity is spam or not.

SpEagle: Rayana and Akoglu [14] utilized clues from all metadata, such as content, timestamp and rating from Yelp.com datasets [14], as well as social information, and combined them employing a bound together system to identify spam reviews. Encourage, they built a user-review-product demonstrate with a pairwise Markov Irregular Field (MRF) [65], to handle a network-based classification assignment. They too planned a light adaptation of SpEagle called SpLite which employments a really little set of review features to boost the computation speed.

Coherence Metrics Computation: Yang [23] found that human composing will illustrate certain word move designs actually between two continuous sentences. When a word was given in one sentence, certain words can be watched in its taking after sentence with a few probability [66]. Be that as it may, such move designs in spam reviews can be impeded due to their beguiling nature. At that point, they to begin with characterized some reviews' coherent measurements to analyze review coherence within the unit of sentence [23]. They proposed a bipartite chart to show all store-sentiment word sets, set of reviews and the association between reviews and store-sentiment word sets. Assist, Yang given a few measurements to degree the coherence of a review based on two sorts data: word move likelihood, word concurrence likelihood.

NetSpam: Shehnepoor [15] utilized spam features for modeling review datasets as heterogeneous data systems to outline spam review location strategy into a classification issue in such networks [15]. The most commitment in this work was that they proposed distinctive metapath sorts which were the inventive within the spam review discovery assignment. The metapath expanded the concept of edge sorts to way sorts and depicted the diverse relations among hubs through roundabout joins, i.e. ways, additionally inferred differing semantics. Encourage, they created the classification portion for recognizing errand with two steps, such as metapath weight calculation and last likelihood calculation.

ATF: Weng et al. [67] created an productive and adaptable AnTi-Fraud framework (ATF) to identify e-commerce fakes for large-scale e-commerce platforms. For the engineering of ATF, they found three components: preprocessor, Graph-Based Discovery module (GBD), and Time Arrangement based Location module (TSD). The GBD was planned as a user-item bipartite chart as portion of the by and large framework for performing spam discovery leveraging the basic and behavioral characteristics of e-commerce spam activities.

Trust Propagation: Xue [68] proposed a three-layer believe engendering demonstrate based on the inter-dependencies between three sorts of hubs: clients, surveys, and statements. Distinctive from the bipartite graph-based two-layer models, the three-layer demonstrate given an extra halfway layer to speak to the impact on one review due to the other review approximately the same protest. Assist, they created an iterative content-based computational demonstrate to compute genuineness scores for diverse substances.

Ianus: Yuan [69] used a Sybil detection method that leverages account registration information. They modeled spam detection as a graph inference problem, which integrated heterogeneous features. Further, they constructed a registration graph that integrated the heterogeneous synchronization and anomaly patterns.

SemiGNN: The research work from Wang et al. [70] mainly focused on tackling three challenges: the bridge between labeled data and unlabeled data, the data heterogeneity and the study of an interpretable model. To address these challenges, they proposed semi-supervised graph neural model with attention mechanism.

GEM: Liu et al. [71] displayed a neural network-based chart demonstrate, named Chart Embeddings for Pernicious accounts (Pearl), which both considered "Device aggregation" and "Activity aggregation" in heterogeneous charts. They centered on managing with the situation of different sorts of nodes and proposed an consideration instrument to memorize the significance of each sort of nodes. Further, they partitioned the arrange into subgraphs concurring to

node sorts and calculated the consideration coefficients in recognizing the spam accounts [71].

3.4.2 Graph Convolutional Networks

Graph convolutional networks can be considered as a simplification of the traditional graph spectral methods, and its common strategy is to model a node's neighborhood as the receptive field and then apply the convolution operation to the deep-learning processing. The graph convolution operator is denoted as feature aggregation of one-hop neighbors. Utilizing the multi-layer convolution operation, information can be transferred among multiple hops. GCN-based method achieves significant improvements compared to previous graph neutral network methods such as DeepWalk [72]. After Perozzi et al. [72] first proposed DeepWalk, which applies SkipGram model [73] on the generated random walks, a large number of scholars have engaged in this area.

GCNN: Alhosseini [74] developed a model based graph convolutional neutral networks (GCNN) for spam bot detection. For this work, the key idea was that an inductive representation learning approach for spam review detection based on the reviewer profile information and the social network graph on twitter datasets was proposed. Further, the inductive representation learning method was similar to GraphSAGE [75] that had a propagation layer with two sub-layers: aggregation and combination. Finally, they compared with multilayer perceptron (MLP) and belief propagation (BP) [76] and gained better performance in detection task.

FdGars: Wang [77] to begin with connected chart convolution arrange approach for spam review location in online app review system. Particularly, they extricated substance highlights and behavior highlights for each analyst based on their review logs. At that point, the review logs were changed into a rule-based chart structure. They moreover planned a naming strategy to name tall suspicious spammers and start clients. Encourage, they prepared a semi-supervised GCN show to memorize node highlight and chart structure, and assessed FdGars by leveraging the real-world review datasets from Tencent App Store.

MNCN: Ghadery [78] proposed a deep convolutional network architecture with three different objective functions at the same time to address spam review detection. Specially, they considered three parallel convolution layers to capture text features from the input reviews, such as convolution filer, n-gram feature maps and max-pooling layer [78].

GAS: Li and Qin [6] first applied GCN-based method to the spam advertisements detection problem and extended GCN algorithm for heterogeneous graph. The heterogeneous graph presented the local context among reviews, users and products, while the homogeneous graph utilized global context which only extracted from reviews. Particularly, the keypoint of heterogeneous chart was to customize conglomeration sub-layer and combination sub-layer with consideration instrument and time-related inspecting procedure. Other than, they utilized surmised KNN chart algorithm [79] to develop the comment chart based on K closest neighbor of nodes. Assist, they utilized the TextCNN [80] show to urge comment implanting and coordinated it to their chart neural organize show as an end-to-end classification system.

Table 4 Op_spam_v1.4 Datasets statistics

Dataset	Positive corpus	Negative corpus
#Product	20	20
#spam Reviews	400	400
#Non-spam Reviews	400	400

Table 5 Various features of different categories of products

Category	Reviews	Items	Reviewers	Total items
All	5,838,032	1,195,133	2,146,048	6,272,502
Books	2,493,087	637,120	1,076,746	1,185,467
Music	1,327,456	221,432	503,884	888,327
DVD/VHS	633,678	60,292	250,693	157,245
mProducts[a]	228,422	36,692	165,608	901,913

[a]Industry manufactured products like electronics, computers, etc.

Table 6 Review datasets in Yelp.com

Dataset	Review counts	User counts	Products
YelpNYC	359,052	160,225	923
YelpCHI	67,395	38,063	201
YelpZIP	608,598	260,277	5044

4 Data Resource

As mentioned above, most of the spam review detection tasks are highly dependent on labeled data. However, there are less well-labeled datasets for supervised learning task or semi-supervised learning task in real-world. Moreover, the available data resources from existing research work are also hard to collect. In this section, we mainly focus on the open source datasets.

Table 7 Statistics of the 500 restaurants in Shanghai

	Spam reviews	Unlabelled reviews	Sum
#(Review)	3523	6242	9765
#(Users	3310	5894	9067
#(Ips)	1314	4564	5535
AVG(User)	1.064	1.059	1.077
AVG(IP)	2.681	1.368	1.764
AVG(words)	53.17	63.21	59.59

Tripadvisor: This corpus comprises of honest and beguiling lodging reviews of 20 Chicago lodgings. The information is depicted in two papers concurring to the assumption of the review [46, 81]. Modeled as a graph, it only has two entities {hotel, reviews} and two classes {truthful, deceptive}.

This dataset contains 400 honest positive reviews from TripAdvisor, 400 misleading positive reviews from Mechanical Turk, 400 honest negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Cry and 400 beguiling negative reviews from Mechanical Turk. Each of the over cluster comprises of 20 reviews for each of the 20 most prevalent Chicago inns in Table 4.

Amazon: This dataset contains item surveys and metadata from Amazon, counting 142.8 million surveys crossing May 1996–July 2014 [82, 83]. This dataset incorporates reviews (evaluations, content, supportiveness votes), item metadata (depictions, category data, cost, brand, and picture features), and joins (too viewed/also bought charts). Table 5 has appeared different features of diverse categories of products [84].

Each amazon.com review data contains the following features: user/item interactions, star ratings, helpful score, timestamps, product reviews, price, brand, category information and other metadata. Specially, the "helpful score" describes the helpfulness rating of the review. e.g., 2/3. Some scholars used this indicator to carry out the threshold of supervised training and apply it to the classification task [85].

Yelp: We are able utilize the comes about of the Yelp.com commercial spam review channel as the ground-truth for execution assessment by slithering the "not-recommended" information at the foot of review page. In the mean time, there are three datasets from Yelp.com accessible to utilize, its rundown insights is given in Table 6.

YelpCHI dataset has been first utilized by Liu [86] and contains user comments from restaurant and hotel domains in the city of Chicago from Yelp website. NYC and ZIP was created by Rayana et al. [14]. NYC contains online reviews for restaurant located in NYC. ZIP was collected reviews for restaurants according to zipcode. The zipcodes are organized by topography, as such this prepare gives us reviews for eateries in a ceaseless locale of the U.S. outline, counting NJ, VT, CT, and PA [14].

Dianping: This Chinese dataset consists of filtered (spam) reviews and unfiltered (unlabeled) reviews from 500 restaurants in Shanghai. It can be able to make three types of nodes: User, Review, IP address. This dataset was used in Liu et al. [52] and the statistics of this dataset has been shown in Table 7.

5 Conclusion and Future Work

In above sections, we introduce the basic motivation of detecting spam review in e-commence platforms. Then, we present the category of detection task, including review mining, end-to-end classification and cold-start problem. Further, we summarize the existing techniques of spam review detection and divide into three categories: machine learning, neural networks and graph-based methods. Specifically, we discuss machine leaning methods in details from three aspects: supervised learning, semi-supervised learning and unsupervised learning. Then, we first collect some state-of-art graph-based techniques for spam review detection and summarize the core idea of each approach from two subcategories: GNN and GCN, respectively. Finally, we show four available datasets from public websites and describe the data structure of each open source dataset.

Previous researches have done substantial work in spam reviews detection. Most scholars have used supervised learning, pattern discovery, graph-based methods, and relational modeling to solve the problem. However, there is a lack of state-of-art GCN based techniques for real-world spam review detection. So, designing an effective graph convolution network algorithm will be a promising research direction for spam review detection task.

Declarations

Conflict of Interest The authors declare they have no conflicts of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Anderson M, Magruder J. Learning from the crowd: regression discontinuity estimates of the effects of an online review database. Econ J. 2012;122(563):957–89.
2. Luca M. Reviews, reputation, and revenue: the case of yelp. com. In: Com (March 15, 2016). Harvard Business School NOM Unit Working Paper,2016; no. 12-016.
3. Park C-H, Kim Y-G. Identifying key factors affecting consumer purchase behavior in an online shopping context. Int J Retail Distrib Manage. 2003.
4. Jindal N, Liu B. Opinion spam and analysis. In: Proceedings of the 2008 international conference on web search and data mining, 2008; pp. 219–30.
5. Wu Y, Ngai EW, Wu P, Wu C. Fake online reviews: literature review, synthesis, and directions for future research. Decis Support Syst. 2020;132: 113280.
6. Li A, Qin Z, Liu R, Yang, Y, Li D. Spam review detection with graph convolutional networks. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019; pp. 2703–11.
7. Lau RY, Liao S, Kwok RC-W, Xu K, Xia Y, Li Y. Text mining and probabilistic language modeling for online review spam detection. ACM Trans Manage Inf Syst (TMIS). 2012;2(4):1–30.
8. Ott M, Cardie C, Hancock J. Estimating the prevalence of deception in online review communities. In: Proceedings of the 21st international conference on World Wide Web, 2012; pp. 201–10.
9. López V, Del Río S, Benítez JM, Herrera F. Cost-sensitive linguistic fuzzy rule based classification systems under the mapreduce framework for imbalanced big data. Fuzzy Sets Syst. 2015;258:5–38.
10. Fei G, Mukherjee A, Liu B, Hsu M, Castellanos M, Ghosh R. Exploiting burstiness in reviews for review spammer detection. In: Proceedings of the International AAAI Conference on Web and Social Media, 2013; vol. 7, no. 1.
11. Mukherjee A, Liu B, Glance N. Spotting fake reviewer groups in consumer reviews. In: Proceedings of the 21st international conference on World Wide Web, 2012; pp. 191–200.
12. Wang C-C, Day M-Y, Chen C-C, Liou J-W. Detecting spamming reviews using long short-term memory recurrent neural network framework. In: Proceedings of the 2nd International Conference on E-commerce, E-Business and E-Government, 2018; pp. 16–20.
13. Weng H, Ji S, Duan F, Li Z, Chen J, He Q, Wang T. Cats: cross-platform e-commerce fraud detection. In: 2019 IEEE 35th International Conference on Data Engineering (ICDE). IEEE, 2019; pp. 1874–85.
14. Rayana S, Akoglu L. Collective opinion spam detection: bridging review networks and metadata. In: Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining, 2015; pp. 985–94.
15. Shehnepoor S, Salehi M, Farahbakhsh R, Crespi N. Netspam: a network-based spam detection framework for reviews in online social media. IEEE Trans Inf Forensics Secur. 2017;12(7):1585–95.
16. Wang X, Liu K, Zhao J. Handling cold-start problem in review spam detection by jointly embedding texts and behaviors. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017; pp. 366–76.
17. Ren Y, Ji D. Learning to detect deceptive opinion spam: a survey. IEEE Access. 2019;7:42934–45.
18. Vidanagama DU, Silva TP, Karunananda AS. Deceptive consumer review detection: a survey. Artif Intell Rev. 2020;53(2):1323–52.
19. Lai C, Xu K, Lau RY, Li Y, Jing L. Toward a language modeling approach for consumer review spam detection. In: 2010 IEEE 7th International Conference on E-Business Engineering. IEEE, 2010; pp. 1–8.
20. Viviani M, Pasi G. Quantifier guided aggregation for the veracity assessment of online reviews. Int J Intell Syst. 2017;32(5):481–501.
21. Fontanarava J, Pasi G, Viviani M. An ensemble method for the credibility assessment of user-generated content. In: Proceedings of the International Conference on Web Intelligence, 2017; pp. 863–8.
22. Noekhah S, Fouladfar E, Salim N, Ghorashi SH, Hozhabri AA. A novel approach for opinion spam detection in e-commerce. In: Proceedings of the 8th IEEE international conference on E-commerce with focus on E-trust, 2014.
23. Yang X. 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, 2015; pp. 99–102.
24. Li H, Liu B, Mukherjee A, Shao J. Spotting fake reviews using positive-unlabeled learning. Computación y Sistemas. 2014;18(3):467–75.
25. You Z, Qian T, Liu B. An attribute enhanced domain adaptive model for cold-start spam review detection. In: Proceedings of the 27th International Conference on Computational Linguistics, 2018; pp. 1884–95.
26. Li Q, Wu Q, Zhu C, Zhang J, Zhao W. An inferable representation learning for fraud review detection with cold-start problem. In: 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, 2019; pp. 1–8.
27. Xie S, Wang G, Lin S, and Yu PS. Review spam detection via temporal pattern discovery. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012, pp. 823–31.
28. Wang G, Xie S, Liu B, Philip SY. Review graph based online store review spammer detection. In: IEEE 11th international conference on data mining. IEEE. 2011;2011:1242–7.
29. Wang G, Xie S, Liu B, Yu PS. Identify online store review spammers via social review graph. ACM Trans Intell Syst Technol (TIST). 2012;3(4):1–21.
30. Hussain N, Mirza HT, Hussain I, Iqbal F, Memon I. Spam review detection using the linguistic and spammer behavioral methods. IEEE Access. 2020;8:53801–16.
31. Aghakhani H, Machiry A, Nilizadeh S, Kruegel C, Vigna G. Detecting deceptive reviews using generative adversarial networks. In: IEEE Security and Privacy Workshops (SPW). IEEE. 2018;2018:89–95.
32. Zheng P, Yuan S, Wu X, Li J, and Lu A. One-class adversarial nets for fraud detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, 2019, pp. 1286–93.
33. Chen C. Mining the web: discovering knowledge from hypertext data. J Am Soc Inf Sci. 2004;55(3):275.
34. Mukherjee A, Venkataraman V, Liu B, Glance N, et al. Fake review detection: classification and analysis of real and pseudo reviews. UIC-CS-03-2013. Technical Report, 2013.

35. Alom Z, Carminati B, and Ferrari E. Detecting spam accounts on twitter. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2018, pp. 1191–8.
36. Swe MM and Myo NN. Fake accounts detection on twitter using blacklist. In: 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS). IEEE, 2018, pp. 562–6.
37. Jia S, Zhang X, Wang X, and Liu Y. Fake reviews detection based on lda. In: 2018 4th International Conference on Information Management (ICIM). IEEE, 2018, pp. 280–3.
38. Aritsugi M, et al. Exploiting function words feature in classifying deceptive and truthful reviews. In: 2018 Thirteenth International Conference on Digital Information Management (ICDIM). IEEE, 2018, pp. 51–6.
39. Mesnil G, Mikolov T, Ranzato M, and Bengio Y. Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews. arXiv preprint; 2014. arXiv:1412.5335.
40. Yang X and Yu X. Recognizing deceptive reviews based on weighted multi-instance unbalanced support vector machine. In: Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science, 2019, pp. 705–8.
41. Kennedy S, Walsh N, Sloka K, Mccarren A, and Foster J. Fact or factitious? Contextualized opinion spam detection. In: Proceedings of the 57th Annual Meeting of the association for computational linguistics: student research workshop, 2019.
42. Devlin J, Chang MW, Lee K, and Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. 2018.
43. Nilizadeh S, Aghakhani H, Gustafson E, Kruegel C, and Vigna G. Think outside the dataset: Finding fraudulent reviews using cross-dataset analysis. In: The World Wide Web Conference, 2019, pp. 3108–15.
44. Tingxuan S and Lau RYK. Collective classification for social opinion spam detection. In: Proceedings of the 2019 2nd international conference on data science and information technology, 2019, pp. 181–6.
45. Sihombing A and Fong ACM. Fake review detection on yelp dataset using classification techniques in machine learning. In: 2019 International conference on contemporary computing and informatics (IC3I). IEEE, 2019, pp. 64–8.
46. Ott M, Choi Y, Cardie C, and Hancock JT. Finding deceptive opinion spam by any stretch of the imagination. arXiv preprint; 2011. arXiv:1107.4557.
47. Barushka A and Hajek P. The effect of text preprocessing strategies on detecting fake consumer reviews. In: Proceedings of the 2019 3rd international conference on e-business and internet, 2019, pp. 13–7.
48. Hassan R and Islam MR. Detection of fake online reviews using semi-supervised and supervised learning. In: 2019 International conference on electrical, computer and communication engineering (ECCE). IEEE, 2019, pp. 1–5.
49. Prakash P, Shashank N, Arjun M, Yadav PS, Shreyamsa S, and Prazwal N. Fake review prevention using classification and authentication techniques. In: ICT Systems and Sustainability. Springer, 2020, pp. 397–406.
50. Caruana R and Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international conference on machine learning, 2006, pp. 161–8.
51. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2001. J Roy Stat Soc. 2004;167(1):192–192.
52. Li H, Chen Z, Liu B, Wei X, Shao J. Spotting fake reviews via collective positive-unlabeled learning. IEEE Int Conf Data Min. 2014;2014:899–904.
53. Ren Y, Ji D, and Zhang H. Positive unlabeled learning for deceptive reviews detection. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 488–98.
54. Hai Z, Zhao P, Cheng P, Yang P, Li X-L, and Li G. Deceptive review spam detection via exploiting task relatedness and unlabeled data. In: Proceedings of the 2016 conference on empirical methods in natural language processing, 2016, pp. 1817–26.
55. Wu Z, Cao J, Wang Y, Wang Y, Zhang L, Wu J. hpsd: a hybrid pu-learning-based spammer detection model for product reviews. IEEE Trans Cybernet. 2018;50(4):1595–606.
56. Yilmaz CM and Durahim AO. Spr2ep: a semi-supervised spam review detection framework. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2018, pp. 306–13.
57. Liu W, Jing W, Li Y. Incorporating feature representation into bilstm for deceptive review detection. Computing. 2020;102(3):701–15.
58. Barushka A and Hajek P. Review spam detection using word embeddings and deep neural networks. In: IFIP International conference on artificial intelligence applications and innovations. Springer, 2019, pp. 340–50.
59. Archchitha K and Charles E. Opinion spam detection in online reviews using neural networks. In: 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer), vol. 250. IEEE, 2019, pp. 1–6.
60. Yuan C, Zhou W, Ma Q, Lv S, Han J, and Hu S. Learning review representations from user and product level information for spam detection. In: 2019 IEEE international conference on data mining (ICDM). IEEE, 2019; pp. 1444–9.
61. Wang Z, Zhang J, Feng J, Chen Z. Knowledge graph embedding by translating on hyperplanes. In: Proceedings of the AAAI conference on artificial intelligence, 2014; vol. 28, no. 1
62. Nayak A, Chen H, Ruan X, and Ouyang J. Deepspot: understanding online opinion spam by text augmentation using sentiment encoder-decoder networks. In: Proceedings of the 3rd ACM SIGSPATIAL international workshop on analytics for local events and news, 2019, pp. 1–10.
63. Ren Y, Zhang Y. Deceptive opinion spam detection using neural network. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers, 2016; pp. 140–50.
64. Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, Wang L, Li C, Sun M. Graph neural networks: a review of methods and applications. AI Open. 2020;1:57–81.
65. Kindermann R. Markov random fields and their applications. Am Math Soc. 1980.
66. Sun H, Morales A, Yan X. Synthetic review spamming and defense. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013; pp. 1088–96.
67. Weng H, Li Z, Ji S, Chu C, Lu H, Du T, He Q. Online e-commerce fraud: a large-scale detection and analysis. In: 2018 IEEE 34th international conference on data engineering (ICDE). IEEE, 2018; pp. 1435–40.
68. Xue H, Wang Q, Luo B, Seo H, Li F. Content-aware trust propagation toward online review spam detection. J Data Inf Quality (JDIQ). 2019;11(3):1–31.
69. Yuan D, Miao Y, Gong NZ, Yang Z, Li Q, Song D, Wang Q, and Liang X. Detecting fake accounts in online social networks at the time of registrations. In: Proceedings of the 2019 ACM SIGSAC conference on computer and communications security, 2019, pp. 1423–38.
70. Wang D, Lin J, Cui P, Jia Q, Wang Z, Fang Y, Yu Q, Zhou J, Yang S, and Qi Y. A semi-supervised graph attentive network for financial fraud detection. In: 2019 IEEE international conference on data mining (ICDM). IEEE, 2019, pp. 598–607.

71. Liu Z, Chen C, Yang X, Zhou J, Li X, and Song L. Heterogeneous graph neural networks for malicious account detection. In: Proceedings of the 27th ACM international conference on information and knowledge management, 2018, pp. 2077–85.
72. Perozzi B, Al-Rfou R, and Skiena S. Deepwalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, pp. 701–10.
73. Mikolov T, Chen K, Corrado G, and Dean J. Efficient estimation of word representations in vector space. arXiv preprint; 2013. arXiv:1301.3781.
74. Ali Alhosseini S, Bin Tareaf R, Najafi P, and Meinel C. Detect me if you can: Spam bot detection using inductive representation learning. In: Companion proceedings of The 2019 World Wide Web conference, 2019, pp. 148–53.
75. Hamilton WL, Ying R, and Leskovec J. Inductive representation learning on large graphs. arXiv preprint; 2017. arXiv:1706.02216.
76. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Elsevier, 2014.
77. Wang J, Wen R, Wu C, Huang Y, Xion J. Fdgars: fraudster detection via graph convolutional networks in online app review system. In: Companion proceedings of The 2019 World Wide Web conference, 2019; pp. 310–6.
78. Ghadery E, Movahedi S, Faili H, Shakery A. Mncn: a multilingual ngram-based convolutional network for aspect category detection in online reviews. In: Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, 2019; pp. 6441–8.
79. Dong W, Moses C, Li K. Efficient k-nearest neighbor graph construction for generic similarity measures. In: Proceedings of the 20th international conference on World wide web, 2011; pp. 577–86.
80. Rakhlin A. "Convolutional neural networks for sentence classification," *GitHub*, 2016.
81. Ott M, Cardie C, Hancock JT. Negative deceptive opinion spam. In: Proceedings of the 2013 conference of the north American chapter of the association for computational linguistics: human language technologies, 2013; pp. 497–501.
82. He R, McAuley J. Ups and downs: modeling the visual evolution of fashion trends with one-class collaborative filtering. In: Proceedings of the 25th international conference on world wide web, 2016; pp. 507–17.
83. McAuley J, Targett C, Shi Q, Van Den Hengel A. Image-based recommendations on styles and substitutes. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, 2015; pp. 43–52.
84. Jindal N, Liu B. Opinion spam and analysis. In: WSDM'08 - Proceedings of the 2008 international conference on web search and data mining, no. November,2008; pp. 219–29.
85. Learning to identify review spam. IJCAI international joint conference on artificial intelligence, no. January 2011,2011; pp. 2488–93
86. Mukherjee A, Venkataraman V, Liu B, Glance N. What yelp fake review filter might be doing?. In: Proceedings of the international AAAI conference on web and social media, 2013; vol. 7, no. 1.