Fake Reviews Detection through Machine learning Algorithms: A Systematic Literature Review

Mohammed Ennaouri (mohammed_ennaouri@um5.ac.ma)
École Nationale Supérieure d'Informatique et d'Analyse des Systèmes

Ahmed Zellou
École Nationale Supérieure d'Informatique et d'Analyse des Systèmes

Research Article

Keywords: Fake reviews, opinion spam, spam reviews, machine learning

Posted Date: September 9th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2039197/v1

Fake Reviews Detection Through Machine learning Algorithms: A Systematic Literature Review

Mohammed Ennaouri[1] and Ahmed Zellou[2]

[1]Software Project Management, National School of System Analysis and informatics, Mohamed Ben Abdellah, Rabat, 10112, Morocco.
[2]Software Project Management, National School of System Analysis and informatics, Mohamed Ben Abdellah, Rabat, 10112, Morocco.

Contributing authors: mohammed_ennaouri@um5.ac.ma; ahmed.zellou@um5.ac.ma;

Abstract

These days, most people refer to user reviews to purchase an online product. Unfortunately, spammers exploited this situation to post deceptive reviews and mislead consumers either to promote a product with poor quality or to demote a brand and damage its reputation. Among the solutions to this problem is human verification. Unfortunately, the real-time nature of fake reviews make the task more difficult especially on e-commerce platforms. The aim of this paper is to conduct a systematic literature review to analyze proposed solutions by researchers who were working on setting up an automatic and efficient framework to identify fake reviews, unsolved problems in the domain and future research direction. Our findings emphasize the importance of the use of certain features and provide researchers and practitioners with insights on proposed solutions and their limitations. Thus, the findings of the study reveals that most approaches focus on sentiment analysis, opinion mining and especially machine learning (ML) which contributes to the development of more powerful models that can significantly solve the problem and thus enhance further the accuracy and efficiency of detecting fake reviews.

Keywords: Fake reviews, opinion spam, spam reviews, machine learning

1 Introduction

In recent years, opinion-sharing platforms have been increasing exponentially, some websites allow people to present their personal experiences, emotions, attitudes, and feelings in order to help future customers who want to get a service or product already tested and approved. Consequently, posting reviews affect significantly the consumers' buying decision Unfortunately, as anyone can write anything and get away with it, an increase in the number of fake reviews has been witnessed. In some cases, the product manufacturers hire "water army" to post online reviews (Chen et al, 2013). For instance, in the context of e-commerce websites, customers got used to going through reviews available before buying any product. Thus, the reviews of products have become an important source of information for the buyers in their decision for purchase of goods. Because of this tendency of customers, online reviews have become a target for spammers. Consequently, fake reviews, also known as deceptive opinions, spam opinions, or spam reviews can on the one hand, cause financial loss for merchants and service providers because their brand reputation can be damaged by negative fake reviews but on the other hand make more profit for companies by posting fake positive reviews. Unfortunately, there is no constraint on writing reviews and posting them to social media. Everyone is allowed to post reviews of any company without any limits. In response to this issue, detecting fake reviews has become a primary concern for platform owners and a good challenge for researchers (Mukherjee et al, 2013). Indeed, several studies tried to harness the power of machine learning and deep learning techniques to classify the review as genuine or fake while most of them are based on supervised learning which is due to the binary aspect of the problem. From other side, spammers can adopt different approaches to post fake reviews. There are who can work individually called individual spammers and who can work in groups called group spammers which refer to a group of reviewers who worked together to write malicious reviews to promote or demote a set of target products. Group spammers are more dangerous than individual spammers because of their size. Consequently, some researchers have adopted three approaches for distancing judicious features, one based on the content called content-based or review features which can be extracted using generally the natural language techniques, other based on the reviewer called reviewer-based or user behavior and finally the features based on the product. In this document, we explore different studies of research about detecting fake reviews. This work was performed by means of a systematic literature review (SLR). We fly over the different methods and some existing datasets described in the literature which can help to determine the future works in this domain. The document is organized as follows. In the next session, the SLR research approach is discussed. Following this, the findings of the review are reported with answers to the request questions through the identified studies. The document ends with a discussion and a conclusion.

2 Research Questions and Search Process

We conduct this Systematic literature review with the aim of identifying and classifying the most relevant research related to the fake reviews detection. The adopted process inspired from Kitchenham's guidelines (Kitchenham and Charters, 2007) based on identifying the research questions, developing the search process, making the study selection and the data extraction.

2.1 Research Questions

To plan the review, we formulate four research questions that goes with the expected goal. The questions are described as follows::

- RQ1: What techniques and approaches are applied to detect fake reviews?
- RQ2: What are the different important areas where fake reviews have overwhelmed?
- RQ3: What are the gaps in detecting fake reviews?
- RQ4: Are there any experimentations in the finding studies? If so, which datasets are used and with what results?

2.2 Search Process

To increase the probability of having relevant articles, it's necessary to use an appropriate set of databases to make sure that the research scope is in adherence to the objectives. Consequently, three Bibliographic databases were used to search for primary studies:

- ScienceDirect (https://www.sciencedirect.com/)
- IEEE (https://www.ieee.org/)
- Acm (dl.acm.org)

The study employed the following search terms:

("fake reviews" OR "opinion spam" OR "spam reviews") AND ("detect" OR "detection") AND ("machine learning" OR "supervised" OR "unsupervised" OR "deep learning")

2.2.1 Inclusion and Exclusion Criteria

To minimize bias in a review, inclusion and exclusion selection criteria needs to be clearly defined. It happens in some cases that the selected studies are irrelevant to the research question and objectives. Having selection criteria eliminate such issues.

Inclusion Criteria

- Include studies in the range from January 2018 to April 2022
- Research studies written in English language only.
- Research Studies and articles with potential to answer either of the research question(s) as formulated in Section 2.1
- Only articles and papers that are published in journals or conferences.

Exclusion Criteria

- Literature that does not fulfill the above-mentioned criteria has been excluded.
- The researchers excluded all the duplicate papers.
- Studies that are not related to the research question.
- Studies not accessible in full text.
- Papers in the form of end of study or a memoir.

2.2.2 Analysis of finding

The initial search resulted in 639 papers from the three bibliographic databases specified above, which confirm the great interest in fake reviews by the researchers and our selection of this subject as a trending area of research described in section two. However, we had to reduce this number of articles by following the systematic literature review process. We began by establishing a limitation of date time between 2018 until 2022 and the establishing of inclusion and exclusion criteria listed in the previous section. Resulting in a total of 215 studies after elimination of duplicate ones. These studies are then subjected for two-stage screening: analysis by title then by abstracts and keywords: in this step we eliminate all the papers that certainly talk about fake reviews but will not respond to the research questions specified before which result in a total of 32.

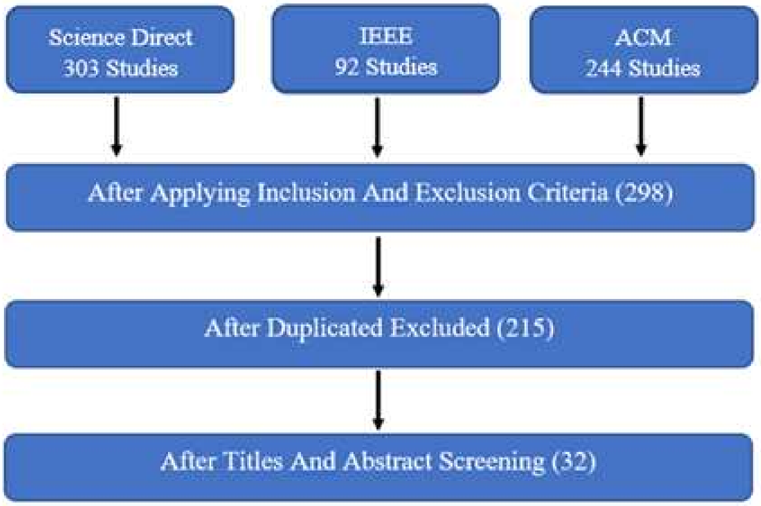
Science Direct
303 Studies
IEEE
92 Studies
ACM
244 Studies
After Applying Inclusion And Exclusion Criteria (298)
After Duplicated Excluded (215)
After Titles And Abstract Screening (32)

Fig. 1 Diagram highlights the stages of the screening process in quantitative terms

3 Result and Answers

Given the valuable studies retrieved from the systematic literature review process, we performed a summary of these studies by giving answers to the research questions defined in the previous section.

RQ1: What techniques and approaches are applied to detect fake reviews?

Research on fake reviews detection is a relatively new area of research, despite that, researchers have designed many methods, most recent ones was Machine Learning algorithms. Machine learning is defined by (Arthur Samuel, 1959) as the "field of study that gives computers the ability to learn without being explicitly programmed". Methods that use machine learning extract and gain knowledge from experience and analytical observations. Because of these advantages of machine learning, the applications of its techniques are expansive. Thus, in this section we list the categories of ML techniques that are used to detect fake reviews.

- Supervised approaches

Most of the literature found throughout our methodology employed supervised methods to detect fake reviews due to their polarity and the high accuracy provided. Therefore, to collect relevant inputs for our classifier, each researcher considers diverse types of features. Mainly, there are three types of features which are being used for fake review detection: review-centric features, reviewer-centric features and product-centric features. First, review-centric features, in which users' textual content is analyzed according to methods, such as bag-of-words, word frequency, n-grams, skip-grams and length of the text. Second, reviewer-centric features, which describe users' information, their interactions, actions, timestamp and may include text counting without deeply analyzing textual content. Finally, there is a product centric features which depends directly on product information. In Table 1, some examples of the features extracted considering the three types of the feature engineering.

Furthermore, recent studies as (Rout et al, 2018) point out the need to address the detection of fake reviews using these feature engineering which consider all the three types. The main idea of their study was to exploit all extracted data to apply Supervised, Semi-Supervised and Unsupervised learning methods and compare theme to deploy the one with the best accuracy. On the other hand, (Martinez-Torres and Toral, 2019) focused just on the content of the text by taking a set of unique attributes based on sentiment polarity while (Siagian and Aritsugi, 2020) found a way to combine word and character n-grams as a feature to detect fake reviews. However, the huge number of attributes which comes with this combination present a problem to apply machine learning algorithms. Fortunately, they applied the Principal Component Analysis (PCA) to classify dominant and non-essential feature attributes which decrease the size of feature attributes. Finally, provide the data to be

Table 1 Examples of features used in the Fake reviews Detection

	Features	Description
Review-Centric-Features	Basic text information	Total (Letters, words, stop words, sentences) in the review Total negative terms Total elongated words (e.g., ‘fiiiine’, ‘Yeees’)
	Linguistic characteristics	The ratio of adjectives and adverbs Average of number of words per sentence Average of number of letters per word
	Sentiment analysis	Total of sentiment terms Total of (positive, neural, negative) of sentiment terms
Reviewer-Centric-Features	Basic user behavior	Total reviews given by the user Total product reviewed by the user Total star given by the user
	Behaviors based on time difference	Minimum, maximum, mean, median and coefficient of variation of the time difference between two consecutive reviews
	Behaviors based on rating/star given	Minimum, maximum, mean, mode, variance, entropy of ratings given by the reviewer
Product-Centric-Features	Basic user behavior	Total reviews for the product, total users who reviewed the product and total rank give, for the product
	Behaviors based on time difference	Minimum, maximum, mean, median and coefficient of variation of the time difference between two consecutive reviews of the same product
	Behaviors based on rating/star given	Minimum, maximum, mean, mode, variance, entropy of ratings given to the product

learned with the use of Machine Learning classifiers for labeling the testing data.

Moreover, to efficiently explore the side of supervised method an Ensemble model has been proposed for classifying data into fake or genuine (Taneja and Kaur, 2021). The approach followed consists of incorporating labels in the existing Cloud Armor dataset by imposing restriction on the number of review counts, service count and probability of collusion feedback factor so that supervised machine learning can be applied using classification models on this labeled dataset. From other side, the problem has been treated differently by (Wang et al, 2020) performed a technique which include two phases to design an alarm system that can monitor the review data stream. First, they generate the most abnormal review subsequences (MARS) by monitoring online reviews from a data stream of a product, during the computation of abnormal subsequences a large number of candidates are produced. Then, depending to the large of the output, they apply the conditional random field (CRF) to label each review in a MARS as fake or genuine by training the MARSs and predict the RFCs with high precision and recall based on two kind of features taking advantage of the relationships between random variables: Node and Edge Features Functions. The process data is an incremental manner with fast response time and less memory consuming. Finally, the authors compared the results with the supervised benchmark classifiers which are SVM, NB and RF.

(Wang et al, 2020) propose a method based on multi-feature fusion includes sentiment analysis, text features of reviews and behaviors features of reviewers extracted with a related algorithm (Doc2vec for text representation as preprocessing step), then they use 7 classifiers in sample labelled dataset and the most accurate classifier is selected to classify new reviews and finally, the output of this step is added into the initial samples and so on.

Otherwise, since supervised learning classification algorithms have proven to be somewhat effective, there is still room for better accuracy with new approaches. (Aiyar and Shetty, 2018) attempt to detect deceptive reviews by applying conventional machine learning algorithms such as Random Forest, Support Vector Machine, Naive Bayes along with certain custom heuristics such as character n-grams which have proven to be very effective in detecting and subsequently combating spam reviews. In the same context, (Jamshidi Nejad et al, 2020) proved that Decision Tree and adaBoost can be effective in detection fake reviews by creating new collection of data feature using text normalization and part of speech tagging.

(Wang et al, 2018) Propose two types of features and apply supervised machine learning algorithms to classify the data. In terms of the features used, they considered two new sets of semantic features: readability features and theme features. They suggest that fake reviews and reviewers participate in the detection of fake reviews. From the perspective of reviews, they proposed a new set of readability features (for example, Automated Readability Index (ARI) and Coleman-Liau Index (CLI)), which mainly evaluate the readability of each review. From another perspective of reviewers, they presented a set of behavioral characteristics, such as restaurant number (RN) and date interval (DI). In addition to the above two types of features, n-gram features based on NLP technology (for example, unigrams and bigrams) are also used to classify reviews as fake or genuine.

On the other hand, some authors propose a graph-based method such as (Noekhah et al, 2020). The main idea of this model is to illustrate the intra and inter relationships among entities and analyze importance of features throw calculating weight based on feature fusion techniques and consequently determine the most effective combination of weighted features. After that, the authors applied a multi-iterative algo is designed to update spamicity. Indeed, there was a preprocessing step based on noise removal and text normalization in both the structure and the data-content. After that, the feature Selection is applied using IG (information Gain) and TF-IDF and the most effective features where selected by applying well-known classifier (SVM, NB and DT). Then, they apply feature fusion techniques to determine the most effective and helpful combination of features and finally calculate spamicity after performing a limited number of iterations as multi-iterative algorithm.

The recent achievements of deep learning techniques in complex natural language processing tasks, make it as a promising solution for fake reviews detection. Therefore, there are successful applications of Convolutional Neural Networks (CNN) for natural language processing problems which have achieved improved performance. Following this approach, (Liu et al, 2022) propose a hierarchical attention network in which distinct attentions are purposely used at the two layers to capture important, comprehensive, and multi-granularity semantic information. At the first layer, they especially use an n-gram CNN to extract the multi-granularity semantics of the sentences. Then, they use a combination of convolution structure and Bi-LSTM to extract important and comprehensive semantics in a document at the second layer. Therefore, the most common way to deal with fake reviews detecting. Also, (Archchitha and Charles, 2019) present a CNN model which is developed to detect opinion spam using the features extracted from the pre-trained Global Vectors for Word Representation model. They made 3 parallels convolution layers with different filter sizes combining data extracted from traditional text and behavioral features. Moreover, some word and character level features used in existing research work, are extracted from the text and concatenated with a feature set extracted by the convolutional layers of the model to improve the performance. In the same context, (Shahariar et al, 2019) proposed deep learning methods for spam review detection which includes Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN) and a variant of Recurrent Neural Network (RNN) that is Long Short-Term Memory (LSTM). They had also applied some traditional machine learning classifiers such as Nave Bayes (NB), K Nearest Neighbor (KNN) and Support Vector Machine (SVM) to detect spam reviews and finally, they have shown the performance comparison for both traditional and deep learning classifiers. Finally, by conducting this SLR,

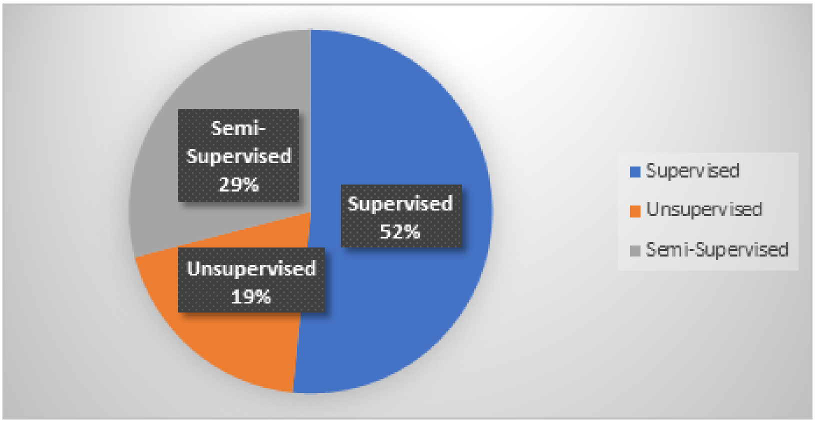

Fig. 2 Percentage of techniques and approaches found on the selected studies

we found out that 52% of the reviewed studies used the supervised learning techniques.

- Unsupervised approaches

In search of an alternative to supervised methods, the authors consider detecting fake with unsupervised manner. In unsupervised machine learning, the algorithm finds its way to cluster the input data. Thus, there is no need for labeled data to detect bots using this approach, and rather than depending on distinct features' values to classify each account, the unsupervised approaches focus more on what is common between groups of accounts and cluster accounts based on similarity between accounts (Mewada and Dewang, 2021)in a single cluster. This approach was illustrated by (Huang and Liang, 2021)) who present two study. First, after a preprocessing step, they evaluate the effect of textual features on the trustworthiness of the review text by conducting a k-means clustering on semantic similarities between the text of reviews to select diverse collection of reviews. In the second study, they tested the impact of attribute salient, review valence and content concreteness on review trustworthiness using multiple regression models.

(Gao et al, 2021) focuses primarily on a neglected emerging domain "movie review" and develops a novel unsupervised spam detection model with an attention mechanism. It is revealed, by extracting the statistical features of reviews, that users will express their sentiments on different aspects of movies in reviews. An attention mechanism is introduced in the review embedding, and the conditional generative adversarial network is exploited to learn users' review style for different kind of movies.

Another approach was done by (Wang et al, 2021). Throughout their work, the authors propose an unsupervised network embedding-based approach to jointly combine direct and indirect neighborhood exploration for learning the user embeddings for more accurately identifying spam reviewers. Indeed, direct relevance is used to measure the degree of the spammer agreements based on

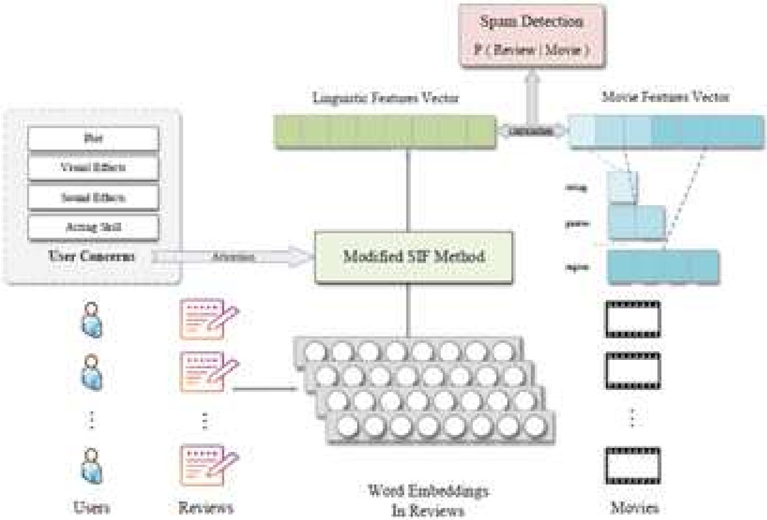

Fig. 3 The fake review detection model illustrates by (Gao et al, 2021)

their direct co-rating associations and then define a user-based signed network based on direct relevance of user for embedding while indirect relevance employs a truncated random walk to measure the indirect relevance for positive users. The authors consider for different type of pairwise features: Product Rating Proximity, Product Time Proximity, Category Rating Proximity and Category Time Proximity.

From other side, in the context of detecting spammer groups using unsupervised learning methods, (Xu et al, 2019) use clues from behavioral data and relational data to conduct a suspicious reviewer graph in three real world datasets from Yelp (Mukherjee et al, 2013) and then they use CPM (Clique Percolation Method) to generate candidate group Spammers (k-clique cluster) and finally ranks opinion spammer groups by group-based and individual-based spam indicators and the top ranked groups are most likely to be opinion spammer groups.

- Semi-supervised approaches

One problem faced on fake reviews detection is a lack of labeled data, there is a limited source of opensource dataset that can be considered in this purpose, the details will be described in next section. To deal with this problem, some authors propose a semi-supervised machine learning technique. Indeed, this approach falls between supervised and unsupervised machine learning because it uses partially labeled data. In other words, to reduce the cost of collecting labeled instances and increase the accuracy of classification, approaches of this type use a large amount of unlabeled data and a proportion of labeled data to build classifiers(Yilmaz and Durahim, 2018)(Navastara et al, 2019).

Even though the number of publications that implement this method is relatively small, semi-supervised machine learning is an interesting research area. The following shows a brief summary of different applications of semi-supervised learning in detecting fake reviews. (Tian et al, 2020) tried to deal with the lack of labeled data by addressing the one-class SVM algorithm. however, to perform their method they tried to introduce a Ramp loss function to minimize the effect of noise and outlier data where does the name "Ramp one-class SVM" come from. The experiment follows these steps: preprocessing (removing stop-words and stemming . . .), after that there was a feature extraction with TF-IDF, then validate the result by applying their algorithm in splitting datasets using 10-fold cross validation to prevent overfitting and finally specifying the Ramp loss function parameters by the grid search techniques. Moreover, some study tried to compare the effectiveness of the baseline semi-supervised methods as (Ligthart et al, 2021). The authors of this article compared six of the main Semi-Supervised learning algorithms (Graph-based learning, Self-training, Co-training, Multi-View learning and Low-diversity separation (LDS) Generative methods) to the following supervised classification methods (SVM, NB and RF). First, after the traditional preprocessing task, they select 1000 top features considered as strongest predictors with chi-square test (with unigram and bi-grams) by taking the high F-value. Then the classification is applied and evaluated based on their accuracy, precision, recall and f1-score. And to find the optimal hyper parameters they applied a Grid Search for each base classifier or semi-supervised method which increase in performance. In the same context, (Hassan and Islam, 2019) introduced some semi-supervised (Expectation Maximization 1) and supervised (NB and SVM algorithms) text mining models to detect fake online reviews as well as compares the efficiency of both techniques. The Expectation Maximization is described as follow: First, a classifier is derived from the labeled dataset to label the unlabeled dataset. The predicted set is named PU. Then, another

Algorithm 1 EM Algorithm

```
    INPUT: Labeled Instance set L, and Unlabeled instance set U.
    OUTPUT: Deployable classifier, C
$C \Leftarrow train(L)$;
$PU = \emptyset$
while true do
    $PU = predict(C, U)$;
    if $PU$ same as in previous iteration then
        $returnC$;
    end if
    $C \Leftarrow train(L \cup PU)$;
end while
```

classifier is taken from the retrieved sets of the unlabeled and labeled datasets and is used to classify the unlabeled dataset again. This process is repeated until the set PU stabilizes. After that, the classification algorithm is trained with the combined training set and deployed for predicting test dataset (Rout et al, 2018).

Further, detecting fake reviews through considering spammer groups has also been treated with unsupervised manner. Indeed, (Zhang et al, 2018) followed this lead and proposed a method which depend on context as many baseline spammer group detectors. In their study, the first thing to do is extraction spammer group candidates using the Frequent items mining (FIM) and then, proceed to manually label positive spammer group after that they apply PU-Learning to extract reliable negative set (RN) from these steps they resulted into some labeled data and unlabeled data. After that a naïve bayes classifier is trained on labeled dataset and then incorporate unlabeled dataset with an expectation maximization (EM) algorithm.

The graph-based approach was interpreted by (Liu et al, 2019). The authors propose a novel approach combining a multimodal neural network-based representation learning and a probabilistic graph model. Indeed, they train a neural network with attention mechanism to learn the multimodal embedded representation of nodes (reviews, authors, and products) by leveraging both textual and rich features after that they incorporate the embedded representation learned from multimodal neural network into a probabilistic review graph for effective spamicity computation. At last, they compare some baselines classifiers as SVM, Linear regression, CNN, Bi-LSTM with mPGM (the proposed model) by using n-gram text and rich features to finalize the prediction based on real life datasets of restaurant and hotel reviews.

Finally, not far from previous approach, (Budhi et al, 2021) propose 133 unique features from the combination of content and behavior-based features (80 for content features, 29 behaviors and 24 product features). They did a sampling process (over-sampling or under-sampling) to increase the accuracy of the minority class and deal with unbalancing data. The experiments were conducted by using Machine learning and Deep learning classifier (MLP, LR, DT, CNN and SVM) with 10-fold cross validation method and by investigating the improvement in processing speed with parallel processing (several CPU works together).

RQ2: What are the different important areas where fake reviews have overwhelmed?

The experiment done in fake reviews detection require mostly a large dataset. However, each dataset uses a specific domain and the extracted features are based on this context which influence the results and make the nature of dataset as a factor of the evaluation metrics. Furthermore, some model obtained with high accuracy and precision perform less when changing the area of its application. The Figure 5 illustrate the areas interpreted by researchers in the 32 selected studies.

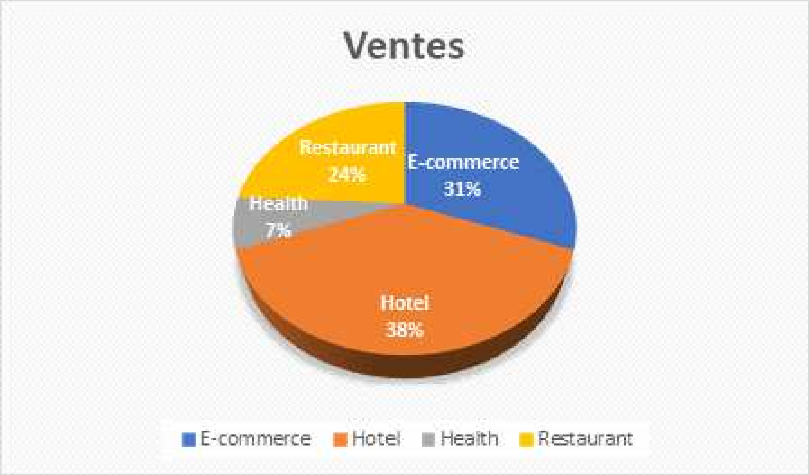

Fig. 4 Areas where Fake Reviews have overwhelmed in last five years

In this figure we considered the most common domain used which are: Hotel, Restaurant, E-commerce, and Health. Indeed, most researchers from the selected studies used hotel area which explained by the application of the Ott and Yelp datasets (Luca and Zervas, 2013). (Ott et al, 2011) generate an opensource dataset that contains hotel reviews from 4 sectors in USA. Thus, the application of NLP, POS and N-grams for instance keep specific to that purpose. From other side, Yelp focus not only in hotel area but also in restaurant and health domain which explain the high percentage of hotel's field application followed by restaurant area where some interesting studies were performed by (Luca and Zervas, 2016). Also, other researchers, as described in previous section, focus more on e-commerce domain by managing their own dataset collected manually even if require more human interference, more effort and lot of time to build a labeled dataset. In fact, E-commerce is the most filed affected by fake reviews in last few years. People rush to buy product or services from online store and most of those consumers prefer to support their decision by consulting reviews of other consumers who sales the same product or service before proceeding to buy them. This reactivity force stores and companies to improve their service or their product. Unfortunately, other stores or other company recruits people to give false positive reviews so that they sell more or give more service.

RQ3: What are the gaps in detecting fake reviews?

Many challenges faced researchers in fake reviews detection. In this section, some gaps in current literature will be identified and discussed.

First, most studies focus on detecting fake reviews using Ott (Ott et al, 2011) or Yelp datasets (Mukherjee et al, 2013) which concern most of time Hotels field. Even though Hotel reviews is one of a great interest for discussion and opinion sharing, there are many areas that need to be investigated too.

For instance, there is no standard labeled dataset specialized in e-commerce reviews even if it's considered as trending area in last decades. Therefore, experiments must be focused in one area at time which help using the Natural Language Processing, N-grams, and POS fluently and enhance the accuracy of the proposed models. Indeed, while some detection models are platform independent, many are not, which is an obstacle against detecting fake reviews in other popular and important platforms.

(Taneja and Kaur, 2021) built their experiences on Cloud datasets in supervised manner by using the Ensemble Voting (EV) which outperform all other studies. However, the proposed model is not efficient when applying in other datasets. Thus, researchers are recommended to release their datasets for research community. This will help in training new models, testing them or evaluating existing models. Furthermore, even if it requires a great effort in term of human resources and time, new public datasets are needed. For the reasons mentioned earlier and because some of the existing widely used datasets use ambiguous terminology. From other side, the features selected in each study has a major impact on the efficiency of the results. Indeed, the chosen features may be different based on the area where they are applied. This upraises the need for platforms dependent models that employ all possible features of a platform to optimize the recall and precision of fake reviews detection. Thus, there are vague areas that require more investigation. A new direction that started to attract attention of researchers is detecting each type of fake reviews independently rather than detecting all types using one general model using the same feature values.

Second, detecting spammers from early phases was found in few studies. If it succeeds, this approach can be very powerful as it can avoid future spammers from posting fake reviews, but it needs more development and require real experiments.

RQ4: Are there any experimentations in the finding studies? If so, which datasets are used and with what results?

The previous Research Question' answer presented a diverse solution using machine learning methods for detecting fake reviews. However, the performance validation of these models is a crucial task. Thus, several metrics are available, but the most popular used between researchers in fake reviews detection are: F1-Score, Recall and Precision. Accuracy is also a common metric for model evaluation as it provides an accurate result if the numbers of instances of both classes (genuine and deceptive) are equal. The application of such measures is performed using the following formulas:

$$Accuracy = \frac{TP + TN}{\text{TP+TN+FP+FN}} \tag{1}$$

$$Precision = \frac{TP}{\text{TP+FP}} \tag{2}$$

$$Recall = \frac{TP}{\text{TP+FN}} \tag{3}$$

$$F1 = \frac{2xRecallxPrecision}{\text{Recall+Precision}} \tag{4}$$

where,
TP is the number of truly classified positive fake reviews
TN is the number of truly classified negative fake reviews
FP is the number of wrongly classified fake reviews
FN is the number of missed fake reviews

To evaluate their approach, researchers found a real problem in collecting real life data, there is a critical issue of availability of labeled datasets. Labeled datasets are needed to train supervised classifiers or to measure the performance of current detection methods. Moreover, the fact that spammers expand rapidly doubles the need for an up-to-date sufficient dataset. Consequently, most of them use real life Yelp datasets (Luca and Zervas, 2013) or Ott datasets called also gold standard datasets. The gold standard dataset provided by (Ott et al, 2011) was used by many researchers in the state-of-the-art studies.It contained, first, spam reviews generated by using Amazon Mechanical Turk (AMT) which designate a crowdsource anonymous online workers called Turkers that construct the first text-based spam review dataset and many supervised classification-based works. Indeed, (Ott et al, 2011) recruited a group of people from (AMT) to write fake reviews for the same hotels. Second, truthful reviews collected from TripAdvisor.com for 20 popular hotels in the Chicago area of the United States. This dataset consists of 1600 truthful and deceptive labeled reviews in text format: 800 fake reviews and 800 true reviews, from genuine reviews 400 are written with a negative sentiment polarity and 400 includes positive sentiment polarity. Similarly, for fake reviews, 400 include positive and 400 reviews contain negative sentiment. The Table 2 gives more details about the number of reviews in this dataset.

Table 2 OTT DATASET DETAILS

Numbers of Reviews 1	Type	Source
400	Truthful positive	TripAdvisor
400	Deceptive positive	Amazon Mechanical Turk
400	Truthful negative	Expedia, Hotels.com, Orbitz,Priceline,TripAdvisor
400	Deceptive negative	Amazon Mechanical Turk

Therefore, the authors who use the supervised or the semi-supervised manner are those who manipulating the gold standard dataset since they need a labeled data as (Rout et al, 2018), (Martinez-Torres and Toral, 2019) and

(Hassan and Islam, 2020) who shared this path to evaluate their approach and which gave a better performance in terms of Accuracy. Also, some previously cited Deep Learning techniques used this famous open-source dataset as (Liu et al, 2022), (Archchitha and Charles, 2019) and (Neisari et al, 2021) which showed their effectiveness on single and multi-domain contexts with accuracy between 88% and 90%. (Noekhah et al, 2020) tried in turn to use two datasets, crowdsourced one from Amazon.com and the second from Ott dataset (Ott et al, 2011) (Ott et al, 2013). The result of this experience showed that the nature of the data affects the evaluation of the proposed method. The following Table 3 resume the results obtained for the supervised methods proposed by the selected articles: On the other side, Yelp dataset (Luca and Zervas, 2013) from Yelp.com is largely used in fake reviews detection to verify the performance of the proposed methods. Yelp does not publish details of their spam filtering algorithm, but the data list is available on Yelp website. It contains some subcategories datasets which are YelpChi, YelpNYC and YelpZip. YelpChi is the smallest dataset, which contains reviews for a set of restaurants and hotels in the Chicago area. YelpNYC and YelpZip are collected in New York City. YelpNYC contains reviews for restaurants located in NYC, and YelpZip contains restaurant reviews from several areas with continuous zip codes starting from NYC.

Table 3 The accuracy of the supervised selected methods by the selected dataset

Article	Dataset	Accuracy
(Rout et al, 2018)	Ott dataset	88.67%
(Martinez-Torres and Toral, 2019)	Ott dataset	85%
(Hassan and Islam, 2020)	Ott dataset	88,75%
(Noekhah et al, 2020)	Ott dataset Amazon.com	95% 93%

The figure 5, shows the databases used by the selected studies to perform their experiment to test the reliability of their algorithm. We can clearly deduce that the Yelp dataset is the most used followed by the gold standard dataset. One can remark that some researchers use both Yelp and Ott dataset, especially if it's about semi-supervised methods. Indeed, (Tian et al, 2020) validate their result by applying their semi-supervised algorithm in splitting dataset (Ott and Yelp datasets) using 10-fold cross validation. Their proposed method called the "Ramp one-class SVM" method (detailed in RQ2) outperforms others method by realizing 92.13% of accuracy in ott dataset with positive reviews, 90.25% in Ott dataset with negative reviews and 74.37% in yelp. Similarly, the experiment of (Shahariar et al, 2019) with their methods based on CNN and LSTM gave a best result for CNN and LSTM over Ott and Yelp with 94.56% accuracy. Moreover, in their testing semi-supervised algorithms and with the use of both datasets, (Ligthart et al, 2021) result that Self-training model with

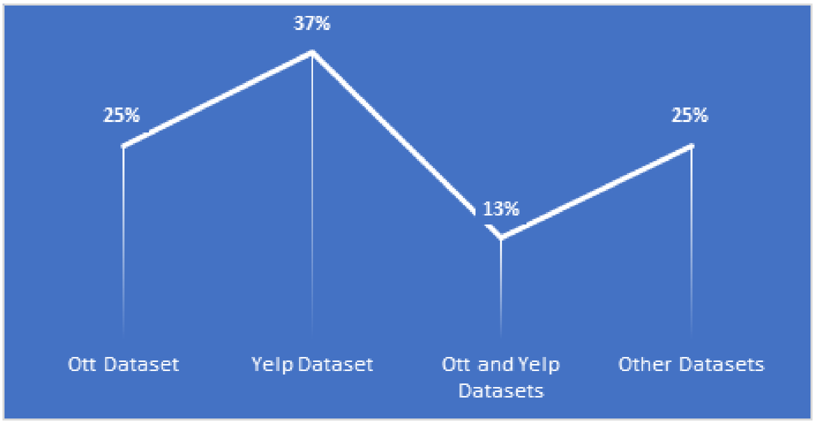

Fig. 5 Databases used on the articles studied

multinomial Naïve Bays as a base classifier and bigrams as input feature yield the best performance of 93% accuracy.

The Yelp dataset was used alone, especially by researchers who adopt unsupervised methods. (Xu et al, 2019) conduct the CPM-Based Group Spamming Detection (GSCPM) by using three real world datasets from Yelp. Their experiment outperforms the four compared methods (GSBP, Wang, Fraud Eagle and SPEagle) in terms of prediction and precision in the condition that the proposed method be applied to a larger dataset.

Finally, other datasets were used in the selected studies. (Taneja and Kaur, 2021) used a labeled dataset called CloudArmor which contains reviews about cloud. The proposed model outperforms all other models with 97.5% of accuracy. Also, (Aiyar and Shetty, 2018) applied their n-gram assisted YouTube spam detection by extracting 13000 comments using public YouTube API and, as a result, they performed 84.37% of accuracy by applying Naïve Bias method and 88.75% of accuracy by Support Vector Machine algorithm. the rest of the studies are divided between those who use their own dataset (Jnoub and Klas, 2020) and those who use Chinese platforms (Gao et al, 2021).

4 Discussion

The detection of fake reviews remains a complex task despite the great effort of researchers made in this direction. Indeed, the nature of the reviews promotes the use of the Naturel language Processing (NLP), Sentiment Analysis, N-grams and Part of Speech Tagging in the features extraction. Thus, each study adopts different kind of features either the content-based features, the behavioral features, or the product-based features.

From other side, the selected studies analyze different machine learning techniques and tools based. It is analyzed that a large number of studies support the use of machine learning algorithms to deal with opinion spam detection. Investigation shows that the most common techniques used is supervised learning especially SVM and RF algorithms to classify deceptive reviews from a selected dataset. Indeed, SVM is an immensely powerful classifier and it is more suited for two class problem. We compared experimentally SVM, Naïve Bayes and K-NN in performance from our selected studies and concluded that SVM has very good predictive power with the higher accuracy. Similarly, Recurrent Neural Network (RNN) can be more effective in detecting fake reviews using Long Short-Term Memory (LSTM) version which opens up another search path based on deep learning. However, unsupervised approaches are in general less effective and have been incorporated so far for detecting fake reviews which are based on graphical methods and which are not very reliable but have the advantage that they do not need labeled datasets for training.

Furthermore, most experiments in the selected studies are based on some specifies open-source datasets from Yelp (Luca and Zervas, 2013), Amazon and Ott dataset (Ott et al, 2011) because of the hard task that can be provided to build a dataset oneself.

5 Conclusion

The widespread use of fake reviews may affect the validity of a reputation system and mislead consumers' purchase decision making. Nevertheless, it is difficult to identify them due to the characteristics of fake reviews. This document particularly focuses on giving an overview of the various approaches that have been used in the state-of-the-art to detect fake reviews. For this purpose, we performed a systematic literature review to identify the different methods that were used in the state-of-the-art studies to detect fake reviews within the last five years. Indeed, the findings of this document are highly beneficial. Moreover, existing studies mostly consider machine learning techniques which have been analyzed in our study. Consequently, we concluded that detection of fake reviews is a complex process.

However, more research is needed to directly explore how to precisely detect deceptive reviews and propose some tools for that purpose. Moreover, spammers always tend to overcome the week point of the researchers' approach by developing new methods based on adopted features. Thus, one of the main future directions is to find robust features that can't be easily manipulated by spammers as well as elaborating an efficient evaluation prototype of fake reviews.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

References

Aiyar S, Shetty NP (2018) N-gram assisted youtube spam comment detection. In: , vol 132. Procedia Computer Science, p 174–182, https://doi.org/https://doi.org/10.1016/j.procs.2018.05.181, URL https://www.sciencedirect.com/science/article/pii/S1877050918309153

Archchitha K, Charles E (2019) Opinion spam detection in online reviews using neural networks. In: 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer), pp 1–6, https://doi.org/10.1109/ICTer48817.2019.9023695

Budhi GS, Chiong R, Wang Z, et al (2021) Using a hybrid content-based and behaviour-based featuring approach in a parallel environment to detect fake reviews. In: , vol 47. Electronic Commerce Research and Applications, p 101048, https://doi.org/https://doi.org/10.1016/j.elerap.2021.101048, URL https://www.sciencedirect.com/science/article/pii/S156742232100020X

Chen C, Wu K, Srinivasan V, et al (2013) Battling the internet water army: Detection of hidden paid posters. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-84893309191&doi=10.1145%2f2492517.2492637&partnerID=40&md5=512d4e84e58327531493fbc330a5cce0

Gao Y, Gong M, Xie Y, et al (2021) An attention-based unsupervised adversarial model for movie review spam detection. In: , vol 23. IEEE Transactions on Multimedia, p 784–796, https://doi.org/10.1109/TMM.2020.2990085

Hassan R, Islam MR (2019) Detection of fake online reviews using semi-supervised and supervised learning. In: 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp 1–5, https://doi.org/10.1109/ECACE.2019.8679186

Hassan R, Islam MR (2020) A supervised machine learning approach to detect fake online reviews. In: 2020 23rd International Conference on Computer and Information Technology (ICCIT), pp 1–6, https://doi.org/10.1109/ICCIT51783.2020.9392727

Huang G, Liang H (2021) Uncovering the effects of textual features on trustworthiness of online consumer reviews: A computational-experimental approach. In: , vol 126. Journal of Business Research, p 1–11, https://doi.org/https://doi.org/10.1016/j.jbusres.2020.12.052, URL https://www.sciencedirect.com/science/article/pii/S014829632030881X

Jamshidi Nejad S, Ahmadi-Abkenari F, Bayat P (2020) Opinion spam detection based on supervised sentiment analysis approach. In: 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE), pp 209–214, https://doi.org/10.1109/ICCKE50421.2020.9303677

Jnoub N, Klas W (2020) Declarative programming approach for fake review detection. In: 2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA, pp 1–7, https://doi.org/10.1109/SMAP49528.2020.9248468

Kitchenham B, Charters S (2007) Guidelines for performing systematic literature reviews in software engineering 2

Ligthart A, Catal C, Tekinerdogan B (2021) Analyzing the effectiveness of semi-supervised learning approaches for opinion spam classification. In: , vol 101. Applied Soft Computing, p 107023, https://doi.org/https://doi.org/10.1016/j.asoc.2020.107023, URL https://www.sciencedirect.com/science/article/pii/S1568494620309625

Liu Y, Pang B, Wang X (2019) Opinion spam detection by incorporating multimodal embedded representation into a probabilistic review graph. In: , vol 366. Neurocomputing, p 276–283, https://doi.org/https://doi.org/10.1016/j.neucom.2019.08.013, URL https://www.sciencedirect.com/science/article/pii/S0925231219311324

Liu Y, Wang L, Shi T, et al (2022) Detection of spam reviews through a hierarchical attention architecture with n-gram CNN and bi-LSTM. In: , vol 103. Information Systems, p 101865, https://doi.org/https://doi.org/10.1016/j.is.2021.101865, URL https://www.sciencedirect.com/science/article/pii/S0306437921000934

Luca M, Zervas G (2013) Fake it till you make it: Reputation, competition, and yelp review fraud. In: , vol 62. SSRN Electronic Journal, https://doi.org/10.2139/ssrn.2293164

Luca M, Zervas G (2016) Fake it till you make it: Reputation, competition, and yelp review fraud. In: , vol 62. INFORMS, Linthicum, MD, USA, p 3412–3427, https://doi.org/10.1287/mnsc.2015.2304, URL https://doi.org/10.1287/mnsc.2015.2304

Martinez-Torres MR, Toral SL (2019) A machine learning approach for the identification of the deceptive reviews in the hospitality sector using unique attributes and sentiment orientation. In: , vol 75. Tourism Management, p 393–403, https://doi.org/https://doi.org/10.1016/j.tourman.2019.06.003, URL https://www.sciencedirect.com/science/article/pii/S0261517719301189

Mewada A, Dewang RK (2021) Deceptive reviewer detection by analyzing web data using HMM and similarity measures. In: . Materials Today: Proceedings, https://doi.org/https://doi.org/10.1016/j.matpr.2020.12.1126, URL https://www.sciencedirect.com/science/article/pii/S2214785320408442

Mukherjee A, Venkataraman V, Liu B, et al (2013) What yelp fake review filter might be doing? In: Kiciman E, Ellison NB, Hogan B, et al (eds) ICWSM. The AAAI Press, URL http://dblp.uni-trier.de/db/conf/icwsm/icwsm2013.html#MukherjeeV0G13

Navastara DA, Zaqiyah AA, Fatichah C (2019) Opinion spam detection in product reviews using self-training semi-supervised learning approach. In: 2019 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA), pp 169–173, https://doi.org/10.1109/ICAMIMIA47173.2019.9223407

Neisari A, Rueda L, Saad S (2021) Spam review detection using self-organizing maps and convolutional neural networks. In: , vol 106. Computers & Security, p 102274, https://doi.org/https://doi.org/10.1016/j.cose.2021.102274, URL https://www.sciencedirect.com/science/article/pii/S0167404821000985

Noekhah S, Salim Nb, Zakaria NH (2020) Opinion spam detection: Using multi-iterative graph-based model. In: , vol 57. Information Processing & Management, p 102140, https://doi.org/https://doi.org/10.1016/j.ipm.2019.102140, URL https://www.sciencedirect.com/science/article/pii/S0306457318306460

Ott M, Choi Y, Cardie C, et al (2011) Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Portland, Oregon, USA, pp 309–319, URL https://aclanthology.org/P11-1032

Ott M, Cardie C, Hancock JT (2013) Negative deceptive opinion spam. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Atlanta, Georgia, pp 497–501, URL https://aclanthology.org/N13-1053

Rout JK, Dash AK, Ray NK (2018) A framework for fake review detection: Issues and challenges. In: 2018 International Conference on Information Technology (ICIT), pp 7–10, https://doi.org/10.1109/ICIT.2018.00014

Shahariar GM, Biswas S, Omar F, et al (2019) Spam review detection using deep learning. In: 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp 0027–0033, https://doi.org/10.1109/IEMCON.2019.8936148

Siagian AHAM, Aritsugi M (2020) Robustness of word and character n-gram combinations in detecting deceptive and truthful opinions. In: , vol 12. J. Data and Information Quality, https://doi.org/10.1145/3349536, URL https://doi.org/10.1145/3349536, place: New York, NY, USA Publisher: Association for Computing Machinery

Taneja H, Kaur S (2021) An ensemble classification model for fake feedback detection using proposed labeled CloudArmor dataset. In: , vol 93. Computers & Electrical Engineering, p 107217, https://doi.org/https://doi.org/10.1016/j.compeleceng.2021.107217, URL https://www.sciencedirect.com/science/article/pii/S004579062100210X

Tian Y, Mirzabagheri M, Tirandazi P, et al (2020) A non-convex semi-supervised approach to opinion spam detection by ramp-one class SVM. In: , vol 57. Information Processing & Management, p 102381, https://doi.org/https://doi.org/10.1016/j.ipm.2020.102381, URL https://www.sciencedirect.com/science/article/pii/S0306457320308761

Wang J, Kan H, Meng F, et al (2020) Fake review detection based on multiple feature fusion and rolling collaborative training. In: , vol 8. IEEE Access, p 182,625–182,639, https://doi.org/10.1109/ACCESS.2020.3028588

Wang X, Zhang X, Jiang C, et al (2018) Identification of fake reviews using semantic and behavioral features. In: 2018 4th International Conference on Information Management (ICIM), pp 92–97, https://doi.org/10.1109/INFOMAN.2018.8392816

Wang Z, Wei W, Mao XL, et al (2021) User-based network embedding for opinion spammer detection. In: . Pattern Recognition, p 108512, https://doi.org/https://doi.org/10.1016/j.patcog.2021.108512, URL https://www.sciencedirect.com/science/article/pii/S0031320321006889

Xu G, Hu M, Ma C, et al (2019) GSCPM: CPM-based group spamming detection in online product reviews. In: ICC 2019 - 2019 IEEE International Conference on Communications (ICC), pp 1–6, https://doi.org/10.1109/ICC.2019.8761650

Yilmaz CM, Durahim AO (2018) SPR2ep: A semi-supervised spam review detection framework. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp 306–313, https://doi.org/10.1109/ASONAM.2018.8508314

Zhang L, Wu Z, Cao J (2018) Detecting spammer groups from product reviews: A partially supervised learning model. In: , vol 6. IEEE Access, p 2559–2568, https://doi.org/10.1109/ACCESS.2017.2784370