Machine Learning Approaches for Fake Reviews Detection: A Systematic Literature Review

Mohammed Ennaouri* and Ahmed Zellou

Software Project Management Team, Mohammed V University in Rabat, High National School for Computer Science and Systems Analysis. Rabat, Morocco
E-mail: mohammed_ennaouri@um5.ac.ma; ahmed.zellou@ensias.um5.ac.ma
**Corresponding Author*

Received 08 October 2022; Accepted 04 November 2023;
Publication 19 December 2023

Abstract

These days, most people refer to user reviews to purchase an online product. Unfortunately, spammers exploit this situation by posting deceptive reviews and misleading 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 makes the task more difficult, especially on e-commerce platforms. The purpose of this study is to conduct a systematic literature review to analyze solutions put out by researchers who have worked on setting up an automatic and efficient framework to identify fake reviews, unsolved problems in the domain, and the 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, in particular, machine learning (ML), which

Journal of Web Engineering, Vol. 22_5, 821–848.
doi: 10.13052/jwe1540-9589.2254

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 and many websites allow users to share their own 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 affects significantly consumers' buying decisions. Unfortunately, as anyone can write anything and get away with it, a rise in the number of opinion spams has been witnessed. In some cases, the product manufacturers hire a "water army" to post online reviews [1]. For instance, in the context of e-commerce websites, customers have become used to going through the reviews available before buying any product. Thus, product reviews have developed into a crucial source of knowledge for consumers making purchasing decisions. Because of this tendency of customers, online reviews have become a target for spammers. Consequently, fake reviews, sometimes referred to as deceptive opinions, spam opinions, or spam reviews, may harm a company's brand name and result in financial loss for retailers and service providers, but they can also increase profits for businesses by publishing falsely positive ratings. Unfortunately, there are no restrictions on reviewing products and sharing them on 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 [2]. Indeed, several studies have 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. On the other hand, spammers can adopt different approaches to posting fake reviews. There are those who work individually, called individual spammers, and those who work in groups, called group spammers, which describes a group of reviewers who have collaborated to publish defamatory reviews of a certain category of target products. Because of their size, group spammers pose a greater threat than individual spammers. 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, another 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 cover the different methods and some existing datasets described in the literature that can help to determine the future works in this domain. The document is organized as follows. In the next section, the SLR research method will be covered. The review's results are then reported along with responses to the request questions obtained from the chosen research. A discussion and a conclusion complete the document.

2 Research Questions and Search Process

We conducted this SLR with the purpose of identifying and classifying the most relevant research related to fake review detection. The adopted process is inspired from Kitchenham's guidelines [3] 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 SLR, we formulated four research questions that match the expected goal. The questions are described as follows::

- RQ1: What techniques and approaches are applied to detect fake reviews?
- RQ2: What techniques and methods are most effective for preprocessing reviews in natural language processing?
- RQ3: What are the different important areas where fake reviews have been overwhelming?
- RQ4: What are the gaps in detecting fake reviews?
- RQ5: Is there any experimentation in the studies? If so, which datasets were 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

matches 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

Clarified inclusion and exclusion selection criteria are necessary to reduce bias in a review. Sometimes the studies that are chosen are not related to the goals and topic of the research. Selection criteria eliminate these problems.

2.2.1.1 *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 any of the research question(s) as formulated in Section 2.1.
- Only articles and papers that are published in journals or conferences.

2.2.1.2 *Exclusion criteria*

- Literature that does not meet the already mentioned criteria has been excluded.
- All the duplicate papers.
- Studies unrelated to the topic under investigation.
- Studies not accessible in full text.
- Papers in the form of end of study or a memoir.

2.2.2 Analysis of findings

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 2. However, we had to reduce the number of articles by following the systematic literature review process.

We began by establishing a limitation of date time between 2018 until 2022 and setting the preceding section's inclusion and exclusion criteria.

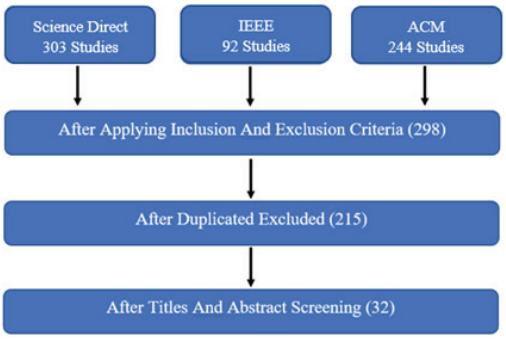


Figure 1 Diagram highlighting the stages of the screening process in quantitative terms.

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 talk about fake reviews but do not respond to the research questions specified, which resulted in a total of 32. Figure 1 shows the established steps.

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 review detection is a recently developed field of study. Despite that, researchers have designed many methods, the most recent one being ML algorithms. ML is defined by Arthur Samuel (1959) as the "field of study that gives computers the ability to learn without being explicitly programmed". Machine learning techniques draw knowledge from analytical observations and experience. These benefits of ML lead to a wide range of uses for its methods. As a result, we list the many ML approach categories used to identify fake reviews in this section.

- 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, reviewer centric and product centric features. First, review centric features analyze the textual content of users 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 user information, their connections, actions, and timestamp, and may incorporate text counts. Finally, there is product centric features which depend directly on product information. In Table 1 some examples of the features extracted considering the three types of the feature engineering are given.

Furthermore, recent studies such as Rout et al. [4] draw attention to the need to address the issue of detecting fake reviews by this feature engineering which considers 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 them to deploy the one with the best accuracy. On the other hand, Martinez-Torres et al. [5] focused just on the content of the text by taking a set of unique attributes based on sentiment polarity. While Siagian et al. [6] developed a feature that merged word and character n-grams to detect fake reviews. However, the huge number of attributes that comes with this combination presents a problem in applying machine learning algorithms. Fortunately, they used principal component analysis (PCA) to divide feature characteristics into dominant and non-essential categories, resulting in a reduction in the number of feature attributes. Finally, to provide the data to be learned with the use of ML classifiers for labeling the testing data.

Moreover, to efficiently explore the side of supervised method an ensemble model was proposed for classifying data into fake or genuine [31]. 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. On the other hand, the problem was treated differently by Wang et al. [7] who performed a technique which included two phases to design an alarm system that can monitor the review data stream. First, they generated 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 on the size of the output, they applied the conditional random field (CRF) to label each review in a MARS as fake or genuine by training the MARSs

Table 1 Examples of features used in the fake review detection.

	Features	Description
Review Centric Features	Elementary text information	Total (letters, words, stop words, sentences) in the review
		Total negative terms
		Total elongated words (e.g., "fiiiine", "Yeees")
	Linguistic features	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 number of sentiment phrases (positive, neural, and negative)
Reviewer Centric Features	Basic user behavior	Total reviews left by the user
		Total product reviewed by the user
		Total star given by the user
	Differentiated behaviors dependent on time	Minimum, maximum, mean, median and coefficient of variation of the time difference between two consecutive reviews
	Behaviors determined by the rating or star granted	Minimum, maximum, mean, mode, variance, entropy of ratings given by the reviewer
Product centric features	Basic product reviews	Total number of product reviews, total number of reviewers, and total number of ranking given for the item
	Differentiated reviews	Minimum, maximum, mean, median and coefficient of variation of the time difference between two consecutive reviews of the same product
	Reviews determined by the rating or star granted	Minimum, maximum, mean, mode, variance, entropy of ratings given to the product

and predicting the RFCs with high precision and recall based on two kinds of features, taking advantage of the relationships between random variables: node and edge feature functions. The process data is an incremental manner with fast response time and is less memory consuming. Finally, the authors compared the results with the supervised benchmark classifiers which are support vector machine (SVMs), naïve Bayes (NB) and random forest (RF).

Wang et al. [7] proposed a method based on multi-feature fusion including sentiment analysis, text features of reviews and behaviors features of

reviewers extracted with a related algorithm (Doc2vec for text representation as the pre-processing step), then they used seven classifiers in a sample labeled dataset and the most accurate classifier was selected to classify new reviews, and, finally, the output of this step was added into the initial samples and so on.

Otherwise, there is still possibility for improved accuracy with new approaches because supervised learning classification algorithms have shown to be relatively useful. Aiyar et al. [8] employed custom heuristics like character n-grams, which have been successfully used to identify and subsequently combat spam reviews, in addition to more traditional machine learning techniques like random forest, support vector machine, and naïve Bayes, to try and detect fake reviews. In the same context, Jamshidi Nejad et al. [9] proved that decision tree and adaBoost can be effective in detection fake reviews by creating new collection of data features using text normalization and part of speech tagging.

Wang et al. [10] applied supervised machine learning techniques and two different types of features to the data to classify it. They took readability characteristics and theme features into consideration when choosing the features to be employed. They suggested that fake reviews and reviewers participate in the detection of fake reviews. They developed a new set of readability elements for reviews, such as the Coleman–Liau index (CLI) and the automated readability index (ARI), which primarily assess each review's readability. From a different reviewer's vantage point, they provided a list of behavioral traits, like the restaurant number (RN) and the date range (DI). In addition to the above two types of characteristics, natural language processing (NLP)-based n-gram features (such unigrams and bigrams) are also employed to categorize reviews as fraudulent or real.

On the other hand, other authors like Noekhah et al. [11] proposed a graph-based approach. The major goal of this model is to show the internal and external relationships between entities, to measure the value of features using feature fusion techniques, and to ultimately identify the best weighted feature combination. After that, the authors applied a multi-iterative algo 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. The authors then used a multi-iterative algorithm that updates spamicity. In fact, both the data's structure and its content underwent a preparation stage based on noise removal and text normalization. The most useful features were chosen by applying well-known classifiers after the feature selection was applied using IG (information gain) and TF-IDF (term

frequency-inverse document frequency) and (SVM, NB and decision tree (DT)). Then, using a multi-iterative algorithm with a restricted number of iterations, they calculated spamicity by applying feature fusion approaches to find the most advantageous and effective combination of features.

Deep learning techniques have recently made progress in difficult natural language processing (NLP) tasks, making it a promising method for spotting bogus reviews. Thus, convolutional neural networks (CNNs) were successfully used to address problems in natural language processing and have demonstrated improved performance. Following this approach, and to capture significant and multi-granularity semantic information, Liu et al. [12] suggested a hierarchical attention network with two levels that strategically used various attentions. They employed an n-gram CNN in particular at the phrase's multi-granularity semantics from the top layer. Then, at the second layer, they extracted significant and comprehensive semantics from a document using a combination of convolution structure and Bi-LSTM. Also, Archchitha et al. [13] presented a CNN model designed to detect opinion spam using features collected from the pretrained global vectors for a word representation model. They merged data from behavioral and traditional text elements into three parallel convolution layers with different filter widths. Additionally, to enhance performance, some word- and character-level characteristics from previous research studies were taken from the text and combined with a feature set taken from the model's convolutional layers. In the same context, Shahariar et al. [14] presented deep learning techniques for detecting spam reviews, including multi-layer perceptrons (MLP), convolutional neural networks (CNNs), and a long short-term memory (LSTM) which is a variation of recurrent neural networks (RNNs). In addition, they used typical machine learning classifiers to identify spam reviews, including naïve Bayes (NB), K nearest neighbor (KNN), and support vector machine (SVM). Finally, they presented a performance comparison between conventional and deep learning classifiers.

Finally, by conducting this SLR, we found out that 52% of the reviewed studies used the supervised learning techniques as shown in Figure 2.

- Unsupervised approaches

The authors explore detecting fake reviews utilizing unsupervised approaches as an alternative to supervised methods. Unsupervised machine learning involves program clustering the input data. As a result, the unsupervised approaches concentrate more on what unites groups of accounts and sorts accounts according to how similar they are [15] in a single cluster. These

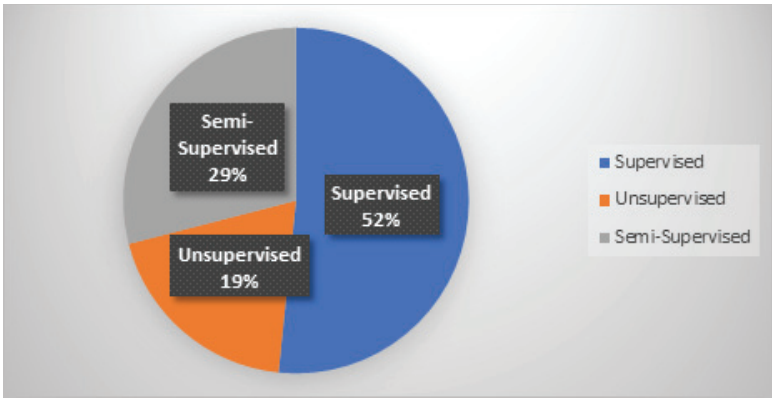


Figure 2 Percentage of techniques and approaches found on the selected studies.

approaches do not require labeled data and instead classify each account based on the values of separate attributes. This approach was illustrated by Huang et al. [16] who presented two studies. They initially evaluated the effect of textual variables on the reliability of the review text by applying a k-means clustering on semantic similarity between the text of reviews to select a diverse set of reviews. In the second study, they utilized multiple regression models to look at how review valence, content concreteness, and attribute salience affected review trustworthiness.

Gao et al. [17] create a novel unsupervised spam detection model with an attention mechanism, focusing especially on the under-utilized developing sector of "movie reviews". By analyzing the statistical components of reviews, it becomes clear that consumers will share their opinions on various parts of movies. The review embedding includes an attention mechanism, and the conditional generative adversarial network is used to learn users' review preferences for various movie genres. Figure 3 illustrates the proposed model.

Wang et al. [18] employed a different strategy. Over the course of their research, the authors present an unsupervised network embedding-based method to jointly integrate direct and indirect neighborhood exploration for learning the user embedding to more precisely identify spam reviewers. In fact, direct relevance uses a truncated random walk to quantify indirect relevance for positive users, whereas direct relevance uses the degree of spammer agreements based on their direct co-rating associations to construct a user-based signed network. The following types of pairwise features are taken into consideration by the authors: product time proximity, product rating proximity, category rating proximity, and category time proximity.

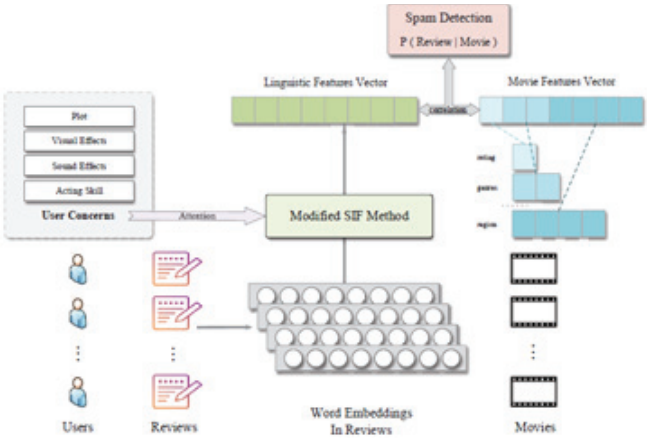


Figure 3 The fake review detection model illustrates by Gao et al. [17].

On the other hand, Xu et al. [19] employed behavioral data as hints in the context of using unsupervised learning algorithms to identify spammer groups. They use the CPM (clique percolation method) to generate candidate group spammers (k-clique cluster), suspicious reviewer graphs, and relational data to conduct suspicious reviewer analyses in three real-world datasets from Yelp [2] Finally, they ranked opinion spammer groups by group-based and individual-based spam indicators, with the highest ranked groups being the most likely to be opinion spammer groups.

- Semi-supervised approaches

One problem faced in fake review detection is a lack of labeled data; there is a limited source of open source datasets that can be considered for this purpose, the details of which will be described in next section. To address this issue, several writers offer a semi-supervised machine learning technique. Because it employs partially labeled data, this method actually lies between supervised and unsupervised machine learning. In other words, techniques of this kind use a large amount of unlabeled data and a small amount of labeled data to develop classifiers in order to lower the cost of gathering labeled instances and raise the accuracy of classification [20, 21].

Semi-supervised machine learning is an intriguing topic of research, despite the fact that there aren't many publications that use this approach. The list of various semi-supervised learning applications for spotting fake reviews is included. Tian et al. [22] attempted to address the scarcity 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 the the name "Ramp one-class SVM" comes from. The experiment followed these steps: preprocessing (removing stop-words and stemming, etc.), then 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. Additionally, some research attempted to compare the efficiency of the standard semisupervised algorithms [23]. The authors of this article compared six of the main semi-supervised learning algorithms (self-training, graph-based learning, co-training, multi view learning and low diversity separation generative methods) to the subsequent supervised classification techniques (SVM, NB and RF). First, after the traditional preprocessing task, they selected 1000 top features considered to be the strongest predictors using the chi-square test (with unigram and bigrams) by taking the high F-value. Then the classification was applied and evaluated based on the common metrics. To find the optimal hyper parameters they applied a grid search for each base classifier or the semisupervised approach, which improved the performance. In the same context, Hassan et al. [24] 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 model is described as follows:

Algorithm 1 EM Algorithm

INPUT: Labeled Instance set L, and Unlabeled instance set U.
OUTPUT: Deployable classifier, C

1: $C \Leftarrow train(L)$;
2: $PU = \emptyset$
3: **while** true **do**
4: $PU = predict(C, U)$;
5: **if** PU same as in previous iteration **then**
6: $returnC$;
7: **end if**
8: $C \Leftarrow train(L \cup PU)$;
9: **end while**

To label the unlabeled dataset, a classifier is first derived from the labeled dataset. Its name is PU for the projected set. The unlabeled dataset is then once again classified using a different classifier that was pulled from the recovered sets of the labeled and unlabeled datasets. Repeat this procedure

until the set PU stabilizes. The classification algorithm is then used to predict test dataset [4] after being trained with the combined training set.

Further, detecting fake reviews through considering spammer groups has also been treated in an unsupervised manner. Indeed, Zhang et al. [25] followed this lead and proposed a method that depended on context as many baseline spammer group detectors. In their study, the first thing done was to extract spammer group candidates using frequent items mining (FIM) and then proceed to manually label positive spammer group. After that they applied PU-learning to extract a reliable negative set (RN) from these steps which resulted in some labeled data and unlabeled data. Then, using the expectation maximization (EM) approach, an unlabeled dataset was added after an NB classifier was trained on the dataset.

Liu et al., [26] provided an interpretation of the graph-based technique. The authors suggest a unique method that combines representation learning using multimodal neural networks and a probabilistic graph model. They actually used both textual and rich features to train a neural network with an attention mechanism to learn the multimodal embedded representation of nodes (reviews, authors, and products), after which they incorporated the learned embedded representation into a probabilistic review graph for efficient spamicity computation. In order to conclude the prediction based on real-world datasets of restaurant and hotel reviews, they compared mPGM (the proposed model) with some baseline classifiers, such as SVM, linear regression, CNN, Bi-LSTM.

Last, but not least, Budhi et al. [27] offered 133 novel features from the combination of content- and behavior-based features, which is not far from the previous method (80 for content features, 29 behaviors and 24 product features). To improve the accuracy of the minority class and deal with unbalanced data, they used a sampling procedure (over- or under-sampling). The research looked at the effects of parallel processing on processing speed by employing machine learning and deep learning classifiers (MLP, linear regression (LR), DT, CNN, and SVM) with a 10-fold cross validation approach (several CPUs working together).

RQ2: What techniques and methods are most effective for preprocessing fake reviews in natural language processing?

In the field of preprocessing, recent state-of-the-art methods mainly focus on improving the efficiency of data cleaning, normalization, and transformation for machine learning and deep learning models. This includes techniques

such as parallel processing, feature selection, and automated feature engineering, as well as the use of unsupervised and semi-supervised methods for handling missing or noisy data. Additionally, preprocessing methods that are specifically designed for large-scale and high-dimensional data, such as deep learning-based feature extractors, have become increasingly popular. The state-of-the-art in fake review preprocessing involves various techniques aimed at detecting and filtering out fake reviews. This is a crucial step in ensuring the reliability and accuracy of customer feedback, which is used by businesses to make important decisions and improve their products or services.

Jnoub et al. [35] listed a set of preprocessing steps in their evaluation based on part of speech (POS) tagging by analyzing distribution of words in a given review, one can extract linguistic features that could indicate the authenticity of the review, n-gram term frequencies, stemming, stop word and punctuation marks filtering as the frequency and distribution of punctuation marks in a review can be used as a feature in a machine learning model to distinguish between real and fake reviews. Similar to Jnoub et al. [35], Jamshidi Nejad et al. [9] also exploited POS tagging combined with text normalization, transforming the informal terms to formal ones. Liu et al. [26] used a method consisting of a Skip-gram model for feature extraction. As the Skip-gram alone is not a guarantee for a successful fake reviews detection system they used the SIF (smooth inverse frequency) which is a natural language processing (NLP) method to weigh words in a text corpus based on their importance by down-weighting words that are too frequent across the entire corpus, such as stop words.

Moreover, researchers as Al Hafiz et al. [6], Aiyar et al. [8] and Y. Gao et al. [17] follow a common stage in the preprocessing field involving stop-words and punctuation removal, lower case conversion and stemming for cleaning data and n-gram extraction for converting the reviews into a numerical representation, such as a Bag-of-Words or (TF-IDF) and then extracting n-grams from the text and use them as features in a machine learning model, such as a support vector machine (SVM), random forest, or neural network. The model would then be trained on labeled data, where the reviews are either labeled as “fake” or “genuine”.

RQ3: What are the different important areas where fake reviews have been overwhelming?

The experiment done in fake review detection requires mostly a large dataset. However, each dataset uses a specific domain and the extracted features

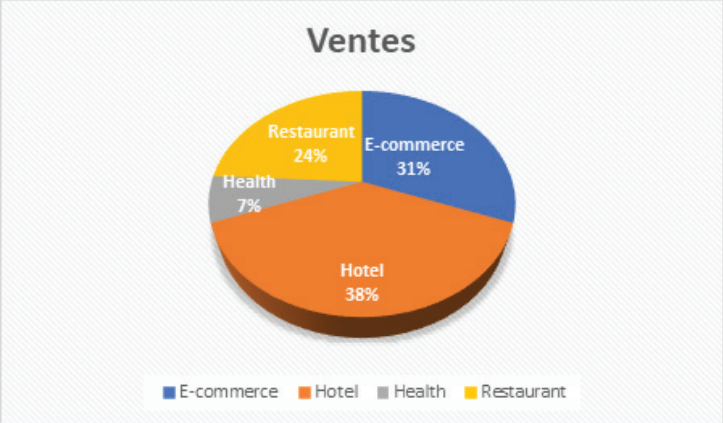


Figure 4 Areas where fake reviews have been overwhelming in last five years.

are based on this context which influence the results and make the nature of dataset as a factor of the evaluation metrics. Furthermore, some models obtained with high accuracy and precision perform less well when changing the area of their application. Figure 4 illustrates the areas interpreted by researchers in the 32 selected studies.

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 the hotel area which is explained by the application of the OTT and Yelp datasets [28]. OTT et al. [29] generated an open source dataset that contains hotel reviews from four sectors in the USA. Thus, the application of NLP, POS and n-grams, for instance, are specific for that purpose. On the other hand, Yelp focuses not only on the hotel area but also on the restaurant and health domains, which explains the high percentage of the hotel field application followed by the restaurant area where some interesting studies were performed by Luca et al. [30]. Also, other researchers, as described in the previous section, focus more on the e-commerce domain by managing their own dataset collected manually, even if it required more human interference, more effort and a lot of time to build a labeled dataset. In fact, e-commerce is the field most 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 buy the same product or service before proceeding to buy them. This reactivity forces stores and companies to improve their service or their product. Unfortunately, other stores or other company recruit

people to give false positive reviews so that they sell more or give a better service.

RQ4: What are the gaps in detecting fake reviews?

Researchers have a lot to contend with when trying to identify fake reviews. The sections that follow will identify and talk about some gaps in the retrieved literature.

First, the majority of studies concentrate on identifying fake reviews using OTT [29] or Yelp datasets [2] which are concerned mostly with the hotels field. Even while discussing and exchanging opinions about hotel reviews is a topic of great interest, there are still a lot of other issues that need to be looked into. For instance, there is no standard labeled dataset specializing in e-commerce reviews even if it has been considered to be a trending area in the last few decades. Therefore, experiments must be focused in one area at time which is helped by using 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 in detecting fake reviews in other popular and important platforms.

Taneja et al. [31] built their experiences on Cloud datasets in supervised manner by using ensemble voting (EV) which outperformed all other studies. However, the proposed model is not efficient when applied to other datasets. Therefore, it is important that researchers make their datasets available to the research community. This will be useful when developing new models, testing them, or assessing them. In addition, even though it takes a lot of time and resources, fresh public datasets are needed due to the reasons already discussed as well as the fact that several of the currently popular datasets use unclear wording. On the other hand, the features selected in each study have a major impact on the efficiency of the results. In fact, depending on where they are used, the features chosen may vary. This highlights the requirement for platform-dependent models that make use of all available platform features in order to maximize the recall and accuracy of fake review identification. As a result, there are some hazy areas that need further research. Detecting each form of fake review separately, as opposed to detecting all sorts using a single general model using the same feature values, is a new approach that has begun to attract academics' interest.

Second, a small number of studies provided evidence for the early detection of spammers. If it works, this strategy might be quite effective because it would deter spammers from publishing fake reviews in the future, but it needs more work and real investigations.

RQ5: Is there any experimentation in the 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, performance validation of these models is critical. Thus, several metrics are available, but the most popular used between researchers in fake review detection are: F1-score, recall and precision. Accuracy is another common model evaluation metric because it provides an accurate result if the records of instances in each of these classes (genuine and deceptive) are equal. The application of such measures is performed using the following formulas:

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

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

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

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

where,

TP stands for true positive which designs the number of positively classified positive fake reviews.
TN stands for true negative which designs the number of truly classified negative fake reviews.
FP stands for false positive which denotes the number of incorrectly classified fake reviews.
FN stands for false negative which is the number of missed fake reviews.

To evaluate their approach, researchers found a real problem in collecting real life data as there is the critical issue of availability of labeled datasets. Labeled datasets are required for training supervised classifiers or evaluating the performance of existing detection methods. Furthermore, the fact that spammers are rapidly expanding doubles the need for an up-to-date adequate dataset. Consequently, most of them use real life Yelp datasets [28] or OTT datasets also called gold standard datasets. The gold standard dataset provided by Ott et al. [29] was used by many researchers in state-of-the-art studies. It included, first, spam reviews generated by Amazon Mechanical Turk (AMT), which refers to a crowdsourced anonymous online workforce

Table 2 OTT dataset details.

Numbers of Reviews	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

known as Turkers who built the first text-based spam review dataset as well as many supervised classification-based works. Ott et al. [29] did, in fact, hire a group of people from AMT to write fake reviews for the same hotels. Second, genuine TripAdvisor.com reviews 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; from genuine reviews, there are 800 fake reviews and 800 true reviews. 400 are written with a negative sentiment polarity, while 400 are written with a positive sentiment polarity. Similarly, for fake reviews, 400 contain positive sentiment and 400 contain negative sentiment. Table 2 provides more information about the number of reviews in this dataset.

Therefore, the authors who used the supervised or the semi-supervised manner are those who manipulated the gold standard dataset since they needed labeled data, similar to Rout et al. [4], Martinez et al. [5] and Hassan et al. [32] 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, similar to Liu et al. [12], Architha et al. [13] and Neisari et al. [33], which showed their effectiveness on single and multidomain contexts with accuracy between 88% and 90%. Noekhah et al. [11] tried in turn to use two datasets, a crowdsourced one from Amazon.com and the second from an OTT dataset [29, 34]. The results of this experience showed that the nature of the data affects the evaluation of the proposed method. Table 3 resumes the results obtained for the supervised methods proposed by the selected articles.

On the other hand, the Yelp dataset [28] from Yelp.com is largely used in fake reviews detection to test the effectiveness of the proposed methods. Although Yelp does not disclose the details of their spam filtering algorithm, the data list is available on the Yelp website. It includes datasets from the YelpChi, YelpNYC, and YelpZip subcategories. YelpChi is the smallest

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

Article	Dataset	Accuracy
Rout et al. [4]	OTT dataset	88.67%
Martinez et al. [5]	OTT dataset	85%
Hassan et al. [32]	OTT dataset	88.75%
Noekhah et al. [11]	OTT dataset	95%
	Amazon.com	93%

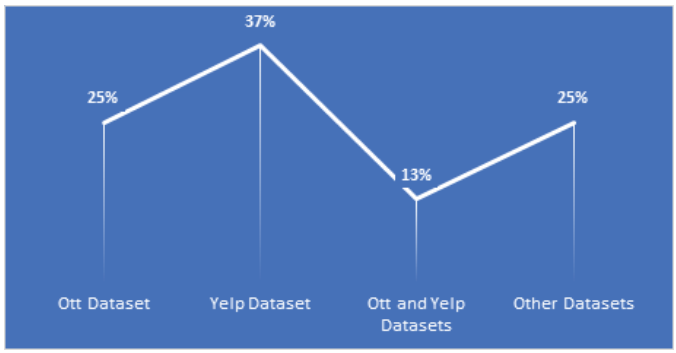


Figure 5 Databases used in the articles studied.

dataset, with reviews for a limited number of restaurants and hotels in the Chicago area. In New York City, YelpNYC and YelpZip are compiled. YelpNYC contains reviews for restaurants in New York City, whereas YelpZip contains restaurant reviews from a variety of areas with continuous zip codes beginning with NYC.

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 used both Yelp and OTT datasets, especially in semi-supervised methods. Indeed, Tian et al. [22] validated their results 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 RQ3) outperforms other methods by realizing 92.13% accuracy in the OTT dataset with positive reviews, 90.25% in the OTT dataset with negative reviews and 74.37% in Yelp. Similarly, the experiment of Shahariar et al. [14], with their methods based on CNN and LSTM, gave better results for CNN and LSTM than OTT

and Yelp with 94.56% accuracy. Moreover, in their testing semi-supervised algorithms and with the use of both datasets, the Lighart et al. [23] results showed that the self-training model with multinomial NB as a base classifier and bigrams as an input feature achieves the best accuracy of 93

The Yelp dataset was used alone, especially by researchers who adopt unsupervised methods. Xu et al. [19] conducted 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. Tanega et al. [31] 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 et al. [8] applied their n-gram assisted YouTube spam detection by extracting 13,000 comments using public YouTube API and, as a result, they performed 84.37% of accuracy by applying the naïve bias method and 88.75% of accuracy using a support vector machine algorithm. The rest of the studies are divided between those who use their own datasets [35] and those who used Chinese platforms [17].

4 Discussion

Despite the considerable effort academics have made in this direction, the detection of fake reviews remains a challenging task. Indeed, the nature of the reviews promotes the use of the natural language processing (NLP), sentiment analysis, n-grams and part of speech tagging in the features extraction. Thus, each study adopts different kinds of features, either content-based features, behavioral features, or product-based features.

On the other hand, the selected studies analyzed different machine learning techniques and tools based. It was 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 technique 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 the 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 that are based on graphical methods, 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 [28], Amazon and OTT dataset [29] because of the hard task that can be provided to build a dataset oneself.

5 Gaps and Future Works

While the body of work already in existence has made outstanding progress in the area of fake review identification using machine learning approaches, there are still many open questions and directions for future research. This section discusses these gaps and proposes potential directions for future works to enhance the effectiveness and robustness of fake review detection systems. One significant challenge in the field of fake review detection is the absence of large, standardized, and diverse datasets that accurately represent the complexity of fake review patterns across various domains and platforms. Indeed, most studies widely used datasets such as Yelp or OTT datasets, which may not fully encapsulate the spectrum of fake review characteristics found in other contexts. In addition, The choice of dataset has a substantial impact on experimental outcomes and the reported performance of detection models. The specific attributes of the dataset, such as the labeling of genuine and fake reviews, the variability in writing styles, and the intricacies of domain-specific language, can significantly influence the efficacy of machine learning algorithms. Consequently, the reported precision, recall, and other metrics may vary extensively based on the unique attributes of the chosen dataset.

Moreover, fake review spammers are likely to employ adversarial techniques to evade detection systems. Researchers should explore the vulnerabilities of existing detection models to adversarial attacks and develop robust models that can withstand such challenges. Adapting techniques from adversarial machine learning could contribute to the creation of more resilient fake review detection systems.

On the other hand, urgent interventions are required for real-time identification to stop the propagation of false reviews. Future work should focus on creating algorithms that can quickly and accurately identify bogus reviews in real-time, allowing platforms to take rapid action against fraudulent content.

6 Conclusion

The extensive usage of fake reviews may undermine the reliability of a reputation system and deceive customers when making purchases. Nevertheless, it is difficult to identify them because of 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 an SLR 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' approaches by developing new methods based on adopted features. Thus, one of the key areas for the future is to develop robust traits that are difficult for spammers to alter, 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

[1] Chen, Cheng, Wu, Kui, Srinivasan, Venkatesh and Zhang, Xudong. (2011). Battling the Internet Water Army: Detection of Hidden Paid Posters. 10.1145/2492517.2492637.

[2] Mukherjee, A., Venkataraman, V., Liu, B., and Glance, N. (2013, June). What yelp fake review filter might be doing?. In Proceedings of the international AAAI conference on web and social media (Vol. 7, No. 1).

[3] Kitchenham, Barbara and Charters, Stuart. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. 2.

[4] J. K. Rout, A. K. Dash and N. K. Ray, "A Framework for Fake Review Detection: Issues and Challenges," 2018 International Conference on

Information Technology (ICIT), 2018, pp. 7–10, doi: 10.1109/ICIT.2018.00014.

[5] Martinez-Torres, Rocio and Toral, S.L. (2019). A machine learning approach for the identification of the deceptive reviews in the hospitality sector using unique attributes and sentiment orientation. Tourism Management. 75. 393–403. doi: 10.1016/j.tourman.2019.06.003.

[6] Siagian, Al Hafiz and Aritsugi, Masayoshi. (2020). Robustness of Word and Character N-gram Combinations in Detecting Deceptive and Truthful Opinions. Journal of Data and Information Quality. 12. 1–24. 10.1145/3349536.

[7] J. Wang, H. Kan, F. Meng, Q. Mu, G. Shi and X. Xiao, "Fake Review Detection Based on Multiple Feature Fusion and Rolling Collaborative Training," in IEEE Access, vol. 8, pp. 182625–182639, 2020, doi: 10.1109/ACCESS.2020.3028588.

[8] Aiyar, Shreyas and Shetty, Nisha. (2018). N-Gram Assisted Youtube Spam Comment Detection. Procedia Computer Science. 132. 174–182. doi: 10.1016/j.procs.2018.05.181.

[9] S. Jamshidi Nejad, F. Ahmadi-Abkenari and P. Bayat, "Opinion Spam Detection based on Supervised Sentiment Analysis Approach," 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE), 2020, pp. 209–214, doi: 10.1109/ICCKE50421.2020.9303677.

[10] X. Wang, X. Zhang, C. Jiang and H. Liu, "Identification of fake reviews using semantic and behavioral features," 2018 4th International Conference on Information Management (ICIM), 2018, pp. 92–97, doi: 10.1109/INFOMAN.2018.8392816.

[11] Noekhah, Shirin, Salim, Naomie and Zakaria, Nor. (2019). Opinion spam detection: Using multi-iterative graph-based model. Information Processing & Management. 57. 21. doi: 10.1016/j.ipm.2019.102140.

[12] Liu, Yuxin, Wang, Li, Shi, Tengfei and Li, Jinyan. (2021). Detection of spam reviews through a hierarchical attention architecture with N-gram CNN and Bi-LSTM. Information Systems. 103. 101865. doi: 10.1016/j.is.2021.101865.

[13] K. Archchitha and E. Y. A. Charles, "Opinion Spam Detection in Online Reviews Using Neural Networks," 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer), 2019, pp. 1–6, doi: 10.1109/ICTer48817.2019.9023695.

[14] G. M. Shahariar, S. Biswas, F. Omar, F. M. Shah and S. Binte Hassan, "Spam Review Detection Using Deep Learning," 2019 IEEE 10th

Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2019, pp. 0027–0033, doi: 10.1109/IEMCON.2019.8936148.

[15] Mewada, Arvind and Dewang, Rupesh. (2021). Deceptive reviewer detection by analyzing web data using HMM and similarity measures. Materials Today: Proceedings. doi: 10.1016/j.matpr.2020.12.1126.

[16] Guanxiong Huang, Hai Liang, Uncovering the effects of textual features on trustworthiness of online consumer reviews: A computational-experimental approach, Journal of Business Research, Volume 126, 2021, Pages 1–11, ISSN 0148-2963, https://doi.org/10.1016/j.jbusres.2020.12.052. (https://www.sciencedirect.com/science/article/pii/S0148296320308881X).

[17] Y. Gao, M. Gong, Y. Xie and A. K. Qin, "An Attention-Based Unsupervised Adversarial Model for Movie Review Spam Detection," in IEEE Transactions on Multimedia, vol. 23, pp. 784–796, 2021, doi: 10.1109/TMM.2020.2990085.

[18] Ziyang Wang, Wei Wei, Xian-Ling Mao, Guibing Guo, Pan Zhou, Sheng Jiang, User-based network embedding for opinion spammer detection, Pattern Recognition, Volume 125, 2022, 108512, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2021.108512. (https://www.sciencedirect.com/science/article/pii/S0031320321006889).

[19] G. Xu, M. Hu, C. Ma and M. Daneshmand, "GSCPM: CPM-Based Group Spamming Detection in Online Product Reviews," ICC 2019 - 2019 IEEE International Conference on Communications (ICC), 2019, pp. 1–6, doi: 10.1109/ICC.2019.8761650.

[20] C. Yılmaz, "SPR2EP: A Semi-Supervised Spam Review Detection Framework," 2018, Accessed: 00, 2020. [Online]. Available: https://hdl.handle.net/11511/54601.

[21] D. A. Navastara, A. A. Zaqiyah and C. Fatichah, "Opinion Spam Detection in Product Reviews Using Self-Training Semi-Supervised Learning Approach," 2019 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA), 2019, pp. 169–173, doi: 10.1109/ICAMIMIA47173.2019.9223407.

[22] Tian, Yingjie, Mirzabagheri, Mahboubeh, Tirandazi, Peyman, Hosseini Bamakan, Seyed Mojtaba. (2020). A non-convex semi-supervised approach to opinion spam detection by ramp-one class SVM. Information Processing & Management. 57. 102381. doi: 10.1016/j.ipm.2020.102381.

[23] Ligthart, Alexander, Catal, Cagatay and Tekinerdogan, Bedir. (2021). Analyzing the effectiveness of semi-supervised learning approaches for opinion spam classification. Applied Soft Computing. 101. 107023. doi: 10.1016/j.asoc.2020.107023.

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

[25] L. Zhang, Z. Wu and J. Cao, "Detecting Spammer Groups From Product Reviews: A Partially Supervised Learning Model," in IEEE Access, vol. 6, pp. 2559–2568, 2018, doi: 10.1109/ACCESS.2017.2784370.

[26] Liu, Yuanchao, Pang, Bo and Wang, Xiaolong. (2019). Opinion Spam Detection by Incorporating Multimodal Embedded Representation into a Probabilistic Review Graph. Neurocomputing. 366. doi: 10.1016/j.neucom.2019.08.013.

[27] Budhi, Gregorius, Chiong, Raymond, Wang, Zuli and Dhakal, Sandeep. (2021). Using a Hybrid Content-Based and Behaviour-Based Featuring Approach in a Parallel Environment to Detect Fake Reviews. Electronic Commerce Research and Applications. 47. 101048. doi: 10.1016/j.elerap.2021.101048.

[28] Luca, Michael and Zervas, Georgios. (2013). Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. SSRN Electronic Journal. 62. doi: 10.2139/ssrn.2293164.

[29] Ott, Myle, Choi, Yejin, Cardie, Claire and Hancock, Jeffrey. (2011). Finding Deceptive Opinion Spam by Any Stretch of the Imagination.

[30] Luca, Michael and Zervas, Georgios. (2013). Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. SSRN Electronic Journal. 62. doi: 10.2139/ssrn.2293164.

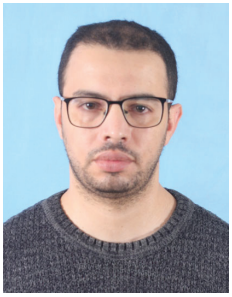
[31] Harsh Taneja, Supreet Kaur, An ensemble classification model for fake feedback detection using proposed labeled CloudArmor dataset, Computers & Electrical Engineering, Volume 93, 2021, 107217, ISSN 0045-7906, https://doi.org/10.1016/j.compeleceng.2021.107217. (https://www.sciencedirect.com/science/article/pii/S004579062100210X).

[32] R. Hassan and M. R. Islam, "A Supervised Machine Learning Approach to Detect Fake Online Reviews," 2020 23rd International Conference on Computer and Information Technology (ICCIT), 2020, pp. 1–6, doi: 10.1109/ICCIT51783.2020.9392727.

[33] Neisari, Ashraf, Rueda, Luis and Saad, Sherif. (2021). Spam Review Detection Using Self-organizing Maps and Convolutional

Neural Networks. Computers & Security. 106. 102274. doi: 10.1016/j.cose.2021.102274.
[34] Ott, Myle, Claire Cardie and Jeffrey T. Hancock. "Negative Deceptive Opinion Spam." NAACL (2013).
[35] N. Jnoub and W. Klas, "Declarative Programming Approach for Fake Review Detection," 2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA, 2020, pp. 1–7, doi: 10.1109/SMAP49528.2020.9248468.

Biographies



Mohammed Ennaouri is a senior researcher in the High National School for Computer Science and Systems Analysis (ENSIAS), Rabat, Morocco. He received his Master's degree in Internet of things: software and analytics in 2021. He is currently working as a teacher in secondary school. His research interest includes fake news, machine learning algorithms and recommender Systems.

Ahmed Zellou Received his Ph.D. in Applied Sciences at the Mohammedia School of Engineers, Mohammed V University, Rabat, Morocco 2008, his habilitation to supervise research work in 2014. He becomes full professor in 2020. His research interests include interoperability, mediation systems, distributed computing, data, indexing, recommander systems, data quality, and semantic web where he is the author/coauthor of over 100 research publications.