A Survey On Opinion Spam Detection Methods

Uzair Aslam, Manoj Jayabalan, Hafiz Ilyas, Asim Suhail

Abstract: Since the past decade, fake Reviews also known as Opinion spam has plagued the e-commerce sector around the world. Opinion spam is considered extremely harmful as it can be used to control the sentiment of a product or service, which in turn can be used to damage the sales and reputation of a company. Throughout the years, extensive research has used Natural language processing for extracting textual features and use them with various machine learning algorithms for opinion spam detection. Majority of the reviewed literature has focused on supervised learning techniques using artificially crafted datasets. The purpose of this paper is twofold: to analyze the various machine learning techniques that have been proposed in the extant literature for detecting opinion spam and compare their accuracies, to provide further insights for future researchers in the field of opinion spam detection. This survey has concluded that semi-supervised techniques using multi-aspect features of reviews, reviewers, and products can provide a better result in spam detection. Furthermore, the lack of accurately labeled datasets presents a major challenge in the field of Fake review detection.

Index Terms: Opinion Spam Detection, Literature Review, Deceptive Reviews, Fake Reviews

1 INTRODUCTION

EVER since the discovery of the World Wide Web, E-commerce has penetrated every aspect of businesses throughout the world. With more and more companies using electronic commerce to market their products and services on the World Wide Web, it can easily be assumed that e-commerce is the only reason for business globalization. Companies around the world, market their products and services on the Internet and gain highly valuable insights about their business through Customer or consumer feedbacks. This feedback can sometimes be in the form of a product review, service review, comments or social media posts. These online reviews also referred to as word of mouth can have a persuasive yet dominating effect, which can influence customer preference for product purchases, brand reputation and promotion of services [1]. There is a gradual increase in a number of organization to incorporate opinion mining as a tool to gain valuable insights from online reviews to understand their customers and the performance of their products and services [2]. Similarly, the researchers in [22] suggest that when it comes to purchasing products or services such as from industries like healthcare, hotels, movies, tourism or even when buying stocks, the customers heavily rely on these online opinions or reviews for making decisions. However, online reviews can be both positive and negative, which can influence the customers in different ways. A negative review of a particular product could discourage potential customers from buying that product and can harm sales revenue for the seller. A positive review can have the opposite effect. The reviews can be manipulated by individuals and companies alike for personal gain such as getting more sales, or for harming the reputation of their competitors [22].

- *Uzair Aslam, Asia Pacific University of Technology & Innovation, Malaysia. E-mail: uzairaslam19@yahoo.com*
- *Manoj Jayabalan, Asia Pacific University of Technology & Innovation, Malaysia. E-mail: manoj@apu.edu.my*
- *Hafiz Ilyas, Asia Pacific University of Technology & Innovation, Malaysia. E-mail: ilyastariq_1996@yahoo.com*
- *Asim Suhail, Asia Pacific University of Technology & Innovation, Malaysia. E-mail: asimsohail@live.com*

These manipulated or fake reviews whether they are positive for self-gain or negative for the purpose of harming one's competitor is called opinion spam or review spam [4]. The negative reviews can significantly harm an organizations reputation and profit margins. Thus, some organization used as a weapon against competitors to harm their reputation and gain an advantage over them, this type of strategy is called Astroturfing. For instance, Samsung was accused and fined $340,300 by the Fairtrade commission Taiwan in 2013, for hiring people to write fake reviews about HTC products, which substantially harmed their product reputation and caused them a significant loss in sales [5]. Review spam has become so prominent that Spam individuals or spammer groups are publicly soliciting their services on websites like Fiver and Facebook, with a single review costing anywhere from $5 to $15[6]. A recent news report by Fox News said that multiloads of Amazon Sellers are manipulating reviews on their own products to increase sales and gain product popularity [7]. Even though websites like Amazon, Yelp and Dianping have strict Spam filters in place, most of these Spam filters can be further improved to accommodate the complex nature of spam, that allows them to pass through these filters undetected. This survey paper critically analyses the existing studies, methodologies, techniques, and algorithms applied for opinion spam detection. Furthermore, the purpose of this study is twofold: To provide future directions and insight to researchers for detecting opinion spam, and to determine the most efficient method or technique available by investigating the results and accuracy of present systems.

2 TYPES OF FAKE REVIEWS

Although researchers have been studying spam for many years, such as web spam and email spam, when it comes to opinion spam a whole new level of challenges arise. Unlike other types of web spam (Email spam, link spam, fake news) opinion spam is difficult to detect manually by the human eye. This makes it almost impossible to extract valuable, gold standard datasets which can be used to design detection algorithms and Systems [8]. Opinion spam can be classified into three distinct categories [8], [9]. Fake Reviews(I): These are the fictitious type of reviews, where the reviewers have not experienced the service or the product that they are writing about. Usually, there is a hidden agenda behind this type, i.e. either to influence user or consumers' opinion regarding a

certain product, service or to promote an idea or ideology. Brand Reviews (II): These reviews are not about a product nor a service rather they are opinions about brands and organization. Non-Reviews (II): This is the type of irrelevant content in the review section of any website that does not portray any sentiment and sometimes is a form of advertisements. The first type is the most difficult to detect manually and can sometimes be passed as truthful opinion. The Type I Reviews can be further classified into two categories, Positive Fake Reviews, and Negative Fake Reviews, and these are more destructive in nature. They can be classified as Deceptive Opinion spam and usually have a hidden agenda behind them [10], [11]. Initially, the spam reviews were duplicates of previously written content, since writing new content every time is quite time-consuming and expensive [12]. Similarly, the initial research was targeted at detecting duplicate or near-duplicate reviews using different machine learning algorithms such as Logistic Regression and Support Vector Machine [2], [4], [9], [10]. Whereas, the second and third types of spam review are quite rare, very unlikely to cause controversy and the damage they cause is quite minimal. The difference between these three types is given in Table I. As we can see the type I review looks genuine and it is difficult to tell just by looking whether it is fake or not, a much more detailed analysis is needed to check whether Type I reviews are spam or not [12].

2.1 Types of Spamming

Fake reviews can be classified as the worst type of false advertisement, as they have a direct effect on the sales of a product or service. Since consumers consider reviews on websites like Yelp or TripAdvisor as sentiments that are both truthful and based on experience, the fake reviews camouflaged as truthful can cause a lot of damage. Opinion spam or fake reviews can be authored by different types of people, for example, a friend or a colleague may write some fake reviews to help promote someone's business. In some cases, the ex-employees of a company write fake reviews to harm the reputation and undermine the services or products provided by that company. In any case, Spamming can be categorized into two types. Individual spamming and Group Spamming [37]. An individual Spammer is someone who writes fake reviews for some personal gain using a single Id to achieve a personal agenda. Which can be either to harm a former employer's reputation or simply writing reviews for some extra cash. Group Spamming can be further categorized into two types [13] Group of Spammers working together towards a common goal i.e. either to promote a certain product or service or undermine or damage the reputation of a product. The second type of group spamming is done by a single user who signs up with multiple user Ids for posting reviews about the same product. This is done to influence and take control of the customer sentiment about a product to either harm the product sales or to promote it. Apart from other types of spamming the most harmful and damaging of these is Group Spamming because it can take control and influence the entire sentiment about a product completely based on the sheer number of reviews. Crowdsourcing forums are becoming the center point for hiring large numbers of individuals to write and target opinion spam towards a common directive for a commission [2], [14], [15]. This has made Spam detection a much more difficult job since apart from simply duplicating content from previously written reviews, Spam authors are creating much more realistic and near-truthful content which is harder to detect. Furthermore, these authors have legitimate User Id's with multiple purchases and have authored many truthful reviews [14], [16].

TABLE ARIOUS TYPES OF REVIEWS FROM TYPE 1, 2 AND 3

Type	Sr no.	Review Content	Comments
I	1	The Phone always hangs whenever I try to use it. Very Bad!!	Review looks genuine
	2	Not enough memory on the phone. Always Hanging!!	
II	1	XYZ has superb products. I have always bought from them and recommend others to buy their products.	Reviews on Brands
	2	RSY is the worst company ever. Don't ever make the mistake of purchasing their products. Wastage of money.	
III	1	50% discount on MNO mobile phones today. Hurry!! Offer valid up until next month.	Advertisement

2.2 International Standards Governing the Publications of Online Reviews

According to a study done by Forbes, 97% of business owners around the world believe that a positive online footprint is crucial for their business in today's world of E-commerce [17]. With a plethora of opinion-based platforms, these days such as Yelp, Dianping, TripAdvisor, Facebook, Google and many more, one of the major concerns is which of the reviews on these platforms are Truthful and trustable. A new standard by the International Organization for Standardization has been published recently that have put things back in order and aims to bring back the trust in reviews [18]. ISO 20488: "Online consumer reviews- Principle and requirements for their collection, moderation, and publication" is aimed at companies and websites that host and publish reviews and is the first-ever International Standard for Online reputation published by the ISO's technical committee [19]. The standard has laid down some rules and guidelines for the collection of reviews, their moderation as well the publication of these reviews on company websites. The standard further guides how to consider certain reviews as fraudulent and fake and how to monitor and police these fraudulent reviews. Some of the guidelines for the publication of these reviews are as follow: All review content should either be rejected or approved for publication without any sort of editing. The review should be published along with the submission date on which the review submitted and the rating that was given. The sharing of personal information of the review author is under his/her own control. All reviews to be published in a timely manner without any sort of Bias. After the publication of the reviews, the website review administrator should allow the reviews to be flagged for being fraudulent or fake. The suppliers for the product in question should be allowed to respond to reviews posted on behalf of their products. Authors should be allowed to delete or remove their reviews from the website. After a review has been flagged and confirmed to be fraudulent and fake the following steps should be taken after its removal. Remove the review in question and marked where it was posted along with the author's name and the reason for removal i.e. Suspicious Activity. The internal Fraud

Mechanisms and Filters should be reviewed, and improvements should be made. Necessary steps should be taken to review the internal moderation steps and improving their accuracy. Review author should be prevented from posting further reviews in the future.

3 FEATURES IN SPAM DETECTION

For review spam detection, choosing the right features can enhance the detection accuracy. Throughout the years quite many features have been presented regarding review spam detection in the literature. These can be categorized into four predominant categories [2]. Review Features Spammer Features Group Features Topological Features

3.1 Review Features

The review features can be categorized into three different types [8]. The content of the review usually refers to the text of the review and it is the first thing that is to be considered when detecting spam [2]. The Content of a review is quite helpful since linguistic features such as POS n-grams and bag(multiset) of words can be extracted from it. Furthermore, other syntactic values such as deceit and deception can also be detected. However, just the using these linguistic features is not reliable because spam authors from crowdsourcing platforms can craft reviews so professionally that it becomes nearly impossible to detect them from a group of truthful opinions [8]. Meta-Data is any kind of information about the review that is describing the content and is not the actual content, for example, the time of the review, reviewers User Identification, Star Rating, IP(Internet Protocol) address and Mac address of the Pc from where the review generated. It is quite easy to mine the behavioral anomalies about the reviews and the reviewers through the metadata of the review. One such example is if a reviewer has written only positive reviews about one brand and negative reviews about its competitors, it is considered as spam. [2], [10]. Product Information can also be quite useful in detecting review spam for example if sales of a particular item are quite low and it has a lot of positive reviews, that raises an alarm on the number and trustworthiness of the positive review. In the study [2] researchers have categorized the review features into two types, Behavior-based, and Text-based. Behavior-based is comprised up of product information and metadata while text-based features are the content of the review.

3.2 Spammer Features

Similar to review features, a spammer or user features can also be categorized into three different types [2], [20]. Behavior-Based Features are those in which user behavior of the spammer is depicted, for example, number of reviews per day, number of purchases pertaining to the number of reviews, Positive and Negative review ratio, Review burst or number of reviews in a short time span. As similar research [21] defined features such as a Maximum number of reviews in 1 day, along with Reviewing Burstiness which means that if all the reviews are posted in a very short interval of time, they are considered as Spam. The ratio of those reviews that were in a burst sequence to that of the total reviews was also taken as an added Behavior-based feature [22]. Text-Based features in terms of Spammer include the number of words in a review and content similarity in reviews from the same user. Average Length of the words in a review along with the similarity of the maximum cosine for all review pairs was defined as text-based features for spammer detection in [20]. Social based features are used for the detection of spamming characteristics for individuals on social media websites like Twitter and Weibo. For example, people on Twitter or Weibo sometimes pay money to gain followers or fans. Though these fans do increase the post reach for the buyer. The sheer number of these zombie followers can attract genuine users to the profile and influence them to follow or like the profile [23].

3.3 Group Features

In recent years. Spamming has become more complex because instead of individual spamming, spammers are now more organized and have increased their effectiveness by working together in a team and pre-planning the spam attacks on a particular target. An Aggregation of the review features and User features can be used to define some valuable attributes which can, in turn, be used for Spammer Group identification. The study done in [24] proposed hierarchical Bayes which they called GLAD that takes pointwise and pairwise data as input and automatically detects spammer groups anomalies [20] proposed a relation-based model for the detection of Group spammers by defining a set of behavior-based features that were extracted from the collusion or collaboration of these spammer individuals.

3.4 Topological Features

Topological features are those that are derived by using graphical analytics and they try to describe each and every node in a graph [16]. These Features are based on the assumption that spammers have strong ties and collude with each other. In other words, when defining Topological features, it is important to assume that the spamming nodes in a graph have a higher centrality, influence, and connectivity.

4 OPINION SPAM DETECTION METHODS

From the existing studies, it has been deduced that opinion spam detection methods were focused and classified into three main categories i.e. methods based on review centric features, spammer features and spammer group features [2], [16]. In this section, the proposed methods and techniques are evaluated and discussed.

4.1 Opinion Spam Detection Methods Using Review Features

The most common approach seen towards the opinion spam detection were based on the review centric features. It utilizes metadata and content of the reviews to be modeled with machine learning algorithms. Researchers have shown interest in supervised learning, unsupervised learning, and semi-supervised learning. The most favored method is supervised learning which requires the system (classifier) to be trained using labeled data. However, this creates a lot of problems, as the major challenges in this field are the unavailability of techniques to accurately collect the data and label it [26]. Prior Studies such as [9], [10], [13] crafted their own artificial data using Amazon Mechanical Turk for generating fake reviews and applied supervised machine learning to train the system. Alternatively, Unsupervised learning focuses on training the system by using unlabeled datasets in order to identify hidden correlations between instances regardless of any class attributes [26]. Clustering is one of the commonly used examples of Unsupervised Learning. Clustering can detect the similarity between

instances from unlabeled data and group them together. A combination of supervised and unsupervised learning is called semi-supervised which uses both labeled and unlabeled data to train a classifier. Semi-supervised Learning has proven to be very effective in detecting opinion spam. Writing reviews with different content every time is a very time-consuming task, to tackle this, most spammers simply copy from previously written reviews and post for multiple products. Supervised learning techniques can be used to detect fake reviews by classifying the data into two categories i.e. spam and not spam.The Researchers Jindal and Liu [4] studied content duplication for opinion spam detection and their findings indicated that spammers usually create templates of existing reviews to later use them in various other products. Prior to the research done by Jindal and Liu[4], the text characteristics that can play a vital part in indicating suspicious activities, such as the creation of opinion spam and fake reviews had not been addressed [10]. The researchers crafted a review dataset by crawling 5.8 million reviews from Amazon for 6.7 million products and 2.14 million reviewers. Since there is no way of detecting fake reviews manually, the researchers considered the nearly duplicate-to-duplicate content of reviews as opinion spam or fake reviews. Jindal and Liu used a 2-gram detection method to detect nearly-similar to similar content in reviews. They identified the duplicates by considering the similarity score or Jaccard distance of over 90% using their 2-gram method. In this method, the similarity score between any two reviews is given by the ratio of the intersection of the union of 2 gram for the reviews to their 2-gram. After detecting spam using the near-duplicates, the researchers came to the conclusion that detecting review spam using just duplicates was insignificant. Further extended with the classification model to train the classifier on the duplicate reviews to detect potential spam reviews. Another research [28] focused on similarity using content-based features in three categories such as similarity ratio of the review with other reviews posted by that author, Review similarity with other reviews for the same item and similar content of reviews posted on other items. Furthermore, the authors employed reviewer centric features in three categories using burst patterns in the writing of reviews which includes, the frequency at which a reviewer posts reviews, the frequency at which a review is posted on a product and frequency at which a reviewer writes for a product. Although supervised methods work great with labeled data, the number of fake reviews in a dataset can be very small which can lead to biasness. For this reason, the researchers employed another method by setting a threshold to detect spam reviews. This was done by calculating the Jaccard similarity for each pair of reviews based on their bigrams, where, If the Jaccard similarity is equal to 0.7 then that would classify the review into the spam category. Using the variants for the weights of the 6 defined features, the review that had opinion spam score near to the value of 1 was categorized as fake. The problem here arises that even though the accuracy of both these models was quite high, sometimes a single genuine author can write 2 similar reviews unconsciously using the same wording or text. The study [10] came up with a novel method of detecting opinion spam by integrating computational linguistics and psychology. They developed and applied three different methods for detecting deceptive opinion spam. Their study produced a new dataset collected by using Amazon Mechanical Turk where they hired a group of people to intentionally write fake reviews for a group of hotels. These reviews were supposed to only include positive sentiment about the hotels. They further collected real-time reviews from Trip Advisor for the same group of hotels and considered these as Truthful reviews. This collected data resulted in being the first-ever publicly available dataset for opinion spam detection which included 400 deceptive reviews and 400 truthful reviews and has been used multiple times in several studies for this field. In a later study [29] the researchers repeated the same process for crafting a dataset of 800 reviews from AMT and Amazon for the same set of hotels focusing primarily on negative sentiments. To benchmark their model's performance the researcher assigned 3 human judges to manually ascertain the spam in a subset of their dataset. Their finding led them to the conclusion that the manual detection of spam through humans was only 60% accurate. Based on the dataset crafted by [10], the researchers in [30] proposed that writing styles, genre, and the readability may vary in genuine and deceptive reviews and by using context-free grammar parse trees the results and accuracy can be improved bi-fold. However, after employing behavioral Features along with Linguistic features, researchers in [35] were able to accurately detect fake reviews in the Yelp dataset with an accuracy of up to 84%. This showed that even though classifiers trained on linguistic features can only detect fake reviews in a synthetic dataset accurately, the same framework is not as much accurate when it comes to real-world dataset i.e. Yelp filtered reviews. Using a combination of Behavioral + Linguistic features increases this accuracy exponentially. A study done in [31] proposed another detection method, which used a feedback neural network for classification based on three different features of the review. The study crafted a dataset from the e-commerce website called Tmall. For feature collection, their model tested the emotional polarity and the text duplication while formalizing and quantifying the metadata. The study employed feature engineering along with quantification and normalization methods for detection through the metadata feature. These extracted features were then fed to the feedforward Neural Network as input which in turn classifies the reviews and gives them as output. Acting on the assumption that spammers tend to use positive and negative words to either undermine or hype up a product, the researchers took advantage of the abstraction and extraction ability of their feedforward network to detect the emotional polarity patterns for fake reviews. The study utilized a Rectified Linear unit as an activation function for the hidden layers in the model. This proved to be highly efficient as their model was able to extract the required information about the reviews. After applying this to real-world data their precision turned out to be 83%. The study [32] argued that past studies had focused too much on detection techniques but failed to pay more attention to feature engineering. They suggested that proper focus on feature selection and engineering can significantly increase the accuracy of the existing models drastically. The researchers made use of a dataset from Yelp which had approximately 5044 different restaurants from 4 different states of United States. It contained 608,589 reviews from 260,277 users in the interval from July 2010 till November 2014. The study used an approach known as stacking where multiple classifiers are combined. The study employed features such as review count, review gap, rating entropy, rating deviation, time of review and user tenure. The researchers proposed a model where instead of using raw

(standard) values for the features, they utilized univariate and multivariate distributions (transformed values) of features and fed them into different classifiers. These features were transformed and then fed into various classifiers such as logistic regression, Naïve Bayes, K, nearest neighbors, SVM, AdaBoost, Random forest and Classification and regression trees, to observe the performance. The researchers used k- 5-fold cross-validation to check their performance. Their results showed that Logistic regression outperformed all other classifiers with an AUC of 0.817. Apart from the CART, all other classifiers outperformed those in previous studies trained on raw feature values. The major challenge in Review spam is the absence of datasets that are accurately labeled, and since Supervised learning requires a labeled dataset it may not always be the right fit. However, Unsupervised learning resolves this issue as it does not require labeled data. In [33] the researchers proposed a novel method of detecting deceptive reviews, they did this by integrating a semantic language model with their developed text mining model. The model calculates the level of deceptiveness in a review by using the semantic language features, they estimate the results of identical and overlapping semantic contents in different reviews. Their model classifies a review as spam if the semantic content of this review matches the content of another review of a different author. Reviews that had a cosine similarity above a set value were manually checked to see if they were opinion spam. Human judges were used to label the reviews as spam if they had a cosine similarity. For those that did not indicate any cosine similarity were considered to be nondeceptive. Another study [34] used both supervised learning and unsupervised techniques to detect review spam. For the supervised method, the researchers utilized the gold standard dataset collected in [10] and proposed a new model. Where previous studies have trained their models on a combination of n-grams and linguistic features only, [34] proposed to use all features (POS tags, Text categorization, Linguistic Features, Genre identification and sentiment as a feature) together to get better results. By combining the different set of features such as n-grams, sentiment score, and linguistic features the researchers trained all three classifiers. The results showed that when trained on all three sets of features the classifiers outperform most of the previous researchers. They utilized both review centric and reviewer centric features for this method such as Textual features and rating related features of the reviews, the user as well as the product on which the review was posted.

4.2 Opinion Spam Detection Methods using Reviewer Features

It has been assumed that opinion spammers usually are more active during certain times of the day, and post spam reviews in those intervals. The number of anomalies is far greater in that interval than any other given time. The researchers in [11] found a delay in time between purchases by truthful and genuine buyers and their reviews. This was done by analyzing the user behaviors during review bursts spam attacks and during normal purchasing session. But this was not the case with spammers, as they started posting reviews as soon as they were given the task on any crowdsourcing platforms. The researchers proposed that spam attack detection can be achieved if the time duration of the product reviews is divided into multiple time intervals. Doing so will increase the chances of finding anomalies Even though [11] provided a better insight using burst patterns for spam detection, it should be made clear that Spam bursts patterns could, in fact, be subjected to other conditions, such as promotion or sale on various products. For this reason, a deeper analysis is required in burst patterns for distinguishing fake and genuine reviews. The researchers in [35] proposed a framework in which both review centric and reviewer centric features were used for the detection of fake reviews in the Chinese language on Dianping website. The researchers noted that the total number of fake reviews that were spawned from the same Ip address were far greater in number than those that were truthful. They categorized the Filtered reviews as Positive instances and the unfiltered ones were considered as unlabeled data. The researchers than argued that although the filtered reviews are guaranteed fake, the unlabeled data might have positive instances (fake reviews) which would have passed the filter altogether. The study proposed using the collective classification model along with the positive-unlabeled learning method to gain higher accuracy than previous studies. Relying on the dependencies between reviewer, review and IP address they proposed using the algorithm Multi-Typed Heterogeneous Collective Classification. But since the MHCC algorithm treats the unlabeled data as simply negative, based on the assumption that there might still be positive data of fake reviews in the unlabeled data, this would mean that MHCC was trained with wrongly labeled data. To counter this the researchers proposed to augment the MHCC algorithm with the collective positive-unlabeled learning framework. This would allow the alteration of initial labels and the training data could be used for testing as during each iteration the model would give out a new set of labels. [36] presented a creative model by taking advantage of the alluring properties of both the Autoencoder neural network and neural random forest. Utilizing the dataset from [13] and a public dataset from Amazon the researchers first pre-processed the data and got a set of 7920 reviews, out of which 3363 were spam. The researchers then employed statistical analysis to extract a number of quality features for their model. These features included rating signal, History records, Products comment info, User's review info, and the feedback signal. Employing these features with the proposed autoencoder decision forest the researchers drew the accuracy and proportion for prediction using each of these distinct features. Even though [36] presented a fairly accurate model for detecting opinion spam, the model is based on extracting quality features using "autoencoder", though it is known to be great for feature extraction, it does have its cons.

4.3 Spammer Group Detection

Nowadays in most cases, spam users collaborate with other spammers to launch spam attack towards a product or service, with the intention of either harming the reputation of that product or taking control of the sentiment for that product. This type of group spam has been clearly outlined and categorized into two types [8]. A single individual with multiple User Ids or a group of individuals working towards a common target either for some form of commission or hidden agenda. These spammer groups may have been hired separately and may or may not interact with each other, but they always work towards a common directive [2]. In [13] the researchers created the first-ever labeled dataset for group spam which had more than 2000 instances of labeled non-spam and spam groups. In [12] the researchers proposed a novel unsupervised learning

method of detecting spammer groups by using a Fuzzy inference system and storing and analyzing data using Hadoop. The researchers argued that although previous studies have categorized review spam detection as a discrete classification problem, it may not be that simple, as there is always a level of uncertainty involved in the detection process due to some hidden factor. To tackle this problem of uncertainty in spam detection they proposed a novel model called FIS (Fuzzy interface system) which uses Fuzzy logic for the detection of suspected groups that are involved in opinion spam. Four Fuzzy Linguistic variables were used which included Fuzzy start Time-lapse, Reviewer Text Similarity, Collective Time Lapse, and peer text similarity. Furthermore, the study proposed a ranking algorithm which they called a Fuzzy ranking evaluation algorithm. This algorithm was used to determine the extent to which a group is suspicious. In [12] they further used the Fuzzy FSL deduction algorithm to define the various fuzzy rules where they came up with eighty-one fuzzy rules. They proposed that if the cosine similarity between reviews was 1 or close to 1 this would constitute these reviews to be in the spamming noise category. The extracted fields such as opinion content, the date on which the review was posted, user identification, the content of opinion and product identification were fed into their novelty algorithm as inputs. The accuracy of their fuzzy inference system for spam group detection was 80.77% with the precision of about 80.82%, which meant that for every 100 sets of suspect spammer groups their model was able to identify and detect up to 81 of these groups. A novel approach was proposed in [37] where they used a semi-supervised method to make use of both labeled and unlabeled data for the detection of spammer groups. They called this model the semi- spammer group detection model. The classifier was used to estimate the probability of groups belonging to a certain class i.e. Spammer or normal. The probability distribution of each of the classes along with the mean and standard deviation of each class was determined using labeled data. Next, they used unlabeled data by deploying the expectation maximization algorithm to improve the Naïve Bayes classifier. Since the parameter variance was exceptionally large from the Naïve Bayes for parsed data, they proposed using expectation maximization algorithm which re-estimates the parameters by repeating the E and M steps until they converge to a single value for estimation. For the Semi SGD model, they utilized a variation of the expectation-maximization model i.e. EM-λ. This was done to utilize the unlabeled data and control its influence by adding in a weighting factor λ for estimation. By setting the value of λ to 0.6 and 100 normal groups as well as 50 spammer groups as training data and putting the list of groups with their spammer probabilities set in descending order they concluded that groups lying on the top positions are more suspect of being spammer groups. While comparing their results with supervised and unsupervised methods, they concluded that the semi-supervised outperformed both techniques and gave a higher accuracy.

5 Discussion

As the domain of opinion spam is quite young when compared to other research areas, such as email and web spam. Only a handful of studies have tackled the problem of detecting review spam using machine learning. Throughout our review, it was noted that most studies were focused on using supervised learning techniques for opinion spam detection. The condition for supervised learning is the availability of labeled data, which in case of review spam is quite difficult if not impossible to craft. As discussed above, most of the datasets that were used in surveyed literature were created synthetically for the sole purpose of research [10], [29], [30]. The issue arises when developing or training classifiers using these synthetic datasets since these datasets do not in any way represent the behavior or qualities of a real-world spammer or spam. For example, researchers in [20] tried the same framework as used by [10], [29], [30] for AMT generated reviews on a ground truth dataset by Yelp, the final results and the features extracted from this ground truth dataset varied by a huge margin. When the classification results for both these datasets are compared in [20] it is seen that the classifier achieved 87% accuracy when it was used to evaluate the synthetic dataset generated by AMT, while only 65% accuracy was achieved when evaluating the ground truth dataset. This sudden drop in accuracy clearly shows that the AMT generated reviews are by no means a representation of the real-world reviews, and that the real-world reviews significantly differ in features as compared to the synthetic ones. A study done by [38] proposed the idea that even though several types of research have indicated that Recurrent Neural Networks are great for generating probabilistic language models, they have fallen short in terms of truly impersonating man written texts. However, this is not the case when it comes to domain-specific texts such as short length reviews which can easily be generated to mimic human-written texts. The researchers thus suggested that Deep neural networks could be used to generate opinion spam by spammers in the near future and might already be in use for such a purpose. To counter such an issue, they developed an automated review writing model based on the Recurring Neural Network (RNN), their findings were that regular language models have limited performance and efficiency when the training data is composed of long textual sequences, whereas RNN resolves this issue by building a memory model. One of the most important conclusions of this study showed that apart from opinion spam written by humans, machine-generated reviews are harder to detect even with the most advanced and best-trained machine learning algorithms. To test this theory the researchers applied SVM's trained on similarity features (cosine similarity of Unigrams), Semantic features (frequency of positive and negative words and sentiments), syntactic features (frequency of POS tags) and LIWC features, however, none of the classifiers could detect and distinguish the machine-generated reviews from the real ones and passed them all as truthful. This shows that spammers are getting smarter and there is a need for smart detection systems to counter that spamming. Most traditional models fell short of identifying and detecting machine generated reviews as spam and let them pass through the filter. Unless one has access to a machine-generated data corpus to further train the models, this approach seems difficult. Researchers in [20], [35] agree that web platforms such as Yelp and Dianping have strict spam filters in place which filter out fake reviews and opinion spam almost flawlessly. Future researchers should not only train their models on synthetic datasets such as those created by [10], [30], rather they should train classifiers based on these ground truth datasets from Yelp or Dianping. Another common issue that were noticed that since most of the past researchers have only focused on batch type offline learning scenarios, these scenarios may not be representative of the real-world

environment, as in the real world the characteristics of a review change periodically over time. The researchers in [39] noticed that since opinion spam is an adversarial classification issue, where spammers are constantly trying to evade detection by the filters, studies should focus on online learning models. The researchers designed several research questions for the review spam domain and performed several experimentations to do an analysis and get insights on these issues. The study organized and based the experimentation on 4 different scenarios such as (offline learning with non-chronologically ordered opinions), and (Using reviews that are sorted on their posting time in an offline learning environment). Both these scenarios were repeated using reviews for online environments. The study employed 2 different datasets, one from Yelp, which was representative of the real-world reviews and the other from [10] which represented the artificially crafted reviews. These classifiers were trained using data in controlled environments for each scenario of online as well as offline reviews. The researchers concluded that F scores were low in the online scenario as compared to the offline ones since online scenarios are trained using a small number of reviews, hence they are prone to more errors initially. However, with the continuous training, the error tends to decrease with the feedback that they receive. The study also concluded that performance is directly affected by the type of product or service the reviews belong too. Their results showed that the performance of the base models is seriously affected when the same is applied for online (time-ordered) reviews which makes them ineffective in a real-world scenario.

5.1 Key Take-Away points

Throughout the reviewed literature it has been noticed that one aspect that has a more significant impact on the accuracy of detection models is feature selection or feature engineering. While most studies such as [39] did focus mainly on this aspect of the problem the issue still persists as most studies [10], [29], [20], [30] have used the same frameworks, same datasets, same classifiers and even similar performance metrics, they have attained different accuracies and results by simply opting for different features than the previous studies. The most noteworthy conclusion throughout the study for various features is that a combination of these features provides enhanced performance and results in opinion spam detection. For example, in [9] the researchers proposed that using review centric features in a combination of reviewer centric features drastically increased the performance of the detection process. Similarly, it can be seen that studies that opted for a combination of textual features with reviewer features yielded slightly better performance. As mentioned in [12] the truthful reviews in the real world significantly outnumber fake reviews, causing a class imbalance in the datasets. This issue has plagued the classifiers in poor performance throughout the reviewed literature as the majority class (truthful reviews) gets favored in the training of a classifier. [41] highlighted class imbalance issue in opinion spam detection and suggested the use of oversampling or random undersampling to overcome it, they showed promising results in dealing with imbalanced datasets. Although both these methods work great in solving the issue of imbalance, ensemble techniques are another way to reduce the effects of class misbalancing. Multiple studies have used the gold standard dataset crafted by [10] and it has been noticed that the maximum accuracy achieved for this dataset was by using a combination of Linguistic inquiry & word count or LIWC and Bigrams. Similarly, for studies that opt for other datasets need to explore a combination of review centric and reviewer centric features for better performance of detection models.

5.2 Future Directions

As it was seen in [35] that websites like Dianping have filters that can detect and filter out spam, but sometimes a few fake reviews can pass through these filters as truthful ones and can cause noise in the dataset. The same classifier used for a synthetic dataset in [13] proved inaccurate on a ground truth dataset by Yelp due to the existence of spamming noise along with other factors in real-world datasets. Ensemble techniques such as bagging, or boosting can prove to be highly useful in moderating the effects of noisy data while classification. Not only are these techniques useful for enhancing performance on noisy data, rather they can help with imbalanced data which is a common occurrence in this domain.

6 CONCLUSIONS

A lot of research has been done on the detection of fake and deceptive reviews and filter it from genuine truthful ones. For this study, we have surveyed most of the existing literature regarding opinion spam detection that uses machine learning and natural language processing. The objective of this study was to better understand the existing research on the methodologies and machine learning techniques used so far and to provide future insights to Researchers. The study has reviewed research work done in 3 different categories of detection methods, Review spam detection, Spam user detection, and Spammer group detection using supervised, unsupervised or semi-supervised learning. It has been noted that even though most of the literature is focused on the review centric features and that too using supervised learning, better accuracy can be attained by taking other features such as reviewer and reviewer groups centric features into account. Topological features such as social media activity of these spammer individuals can further enhance the detection results. From the reviewed literature, it is clear that the major challenge in the field of opinion spam detection is the unavailability of the labeled dataset. Although many studies have crafted their own synthetic datasets, it is noticed from the literature that these datasets do not represent the ground truth, real-world reviews as they were written not by spammers but by turkers for research. Furthermore, these datasets differ from real-world datasets, in features and do not contain the spamming noise or class imbalance found in the real-world data. Since websites like Yelp and Dianping have built-in filters which work flawlessly in detecting fake reviews, data from these web platforms can prove significant in building an accurate model for detecting opinion spam. Since manually labeling real-world spam is not only difficult but impossible, training AI to learn from the human-written reviews and generate similar review can be one way to craft a dataset, mirroring real-time opinion spam. This approach was tried out by (Yao et al., 2017) where the researchers handled this problem by training a recurrent neural network to learn from the review pattern and generate similar to nearly similar reviews. Although this approach did generate nearly similar ground truth data, the problem arose when they bypassed all known machine learning classifiers as truthful. Unless one has access to a machine-generated data corpus to further train the models, this approach seems difficult. Throughout the extent,

literature semi-supervised techniques have out-shined those studies that utilized supervised or unsupervised techniques. The focus should also be directed towards training neural networks and deep neural networks for the classification of fake reviews. Even though some machine learning algorithms perform better than neural networks on smaller data sets, with the ever-increasing data volume for opinions on websites such as Yelp and Dianping, neural networks will be essential in the classification of such huge volumes of data.

REFERENCES

[1] W. Duan, B. Gu, and A. B. Whinston, "Do online reviews matter? - An empirical investigation of panel data," Decis. Support Syst., vol. 45, no. 4, pp. 1007–1016, 2008.

[2] A. Heydari, M. A. Tavakoli, N. Salim, and Z. Heydari, "Detection of review spam: A survey," Expert Syst. Appl., vol. 42, no. 7, pp. 3634–3642, 2015.

[3] N. Hu, I. Bose, N. S. Koh, and L. Liu, "Manipulation of online reviews: An analysis of ratings, readability, and sentiments," Decis. Support Syst., vol. 52, no. 3, pp. 674–684, 2012.

[4] N. Jindal and B. Liu, "Analyzing and detecting review spam," Proc. - IEEE Int. Conf. Data Mining, ICDM, pp. 547–552, 2007.

[5] J. M. Chang, "Samsung Fined For Paying People to Criticize HTC's Products," ABC News, 2013. [Online]. Available: http://abcnews.go.com/Technology/samsung-fined-paying-people-criticize-htcs-products/story?id=20671547. [Accessed: 07-Feb-2018].

[6] Jillian D'Onfro, "Google Maps spam fighters," CNBC, 2018. [Online]. Available: https://www.cnbc.com/2018/04/13/google-maps-spam-fighters.html. [Accessed: 24-May-2018].

[7] T. Daniel, "Shopping on Amazon, how to tell if reviews are fake | Boston 25 News," FoxNews, 2018. [Online]. Available: http://www.fox25boston.com/news/shopping-on-amazon-how-to-tell-if-reviews-are-fake-1/694913717. [Accessed: 07-Feb-2018].

[8] B. Liu, "Opinion Spam Detection," in Sentiment Analysis and Opinion Mining., no. May, Morgan & Claypool Publishers, 2012, pp. 123–135.

[9] N. Jindal and B. Liu, "Opinion spam and analysis," Proc. Int. Conf. Web search web data Min. - WSDM '08, p. 219, 2008.

[10] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding Deceptive Opinion Spam by Any Stretch of the Imagination," pp. 309–319, 2011.

[11] S. Xie, G. Wang, S. Lin, and P. S. Yu, "Review spam detection via temporal pattern discovery," Kdd2012, p. 823, 2012.

[12] K. Dhingra and S. K. Yadav, "Spam analysis of big reviews dataset using Fuzzy Ranking Evaluation Algorithm and Hadoop," Int. J. Mach. Learn. Cybern., vol. 0, no. 0, pp. 1–20, 2017.

[13] A. Mukherjee, B. Liu, J. Wang, N. Glance, and N. Jindal, "Detecting group review spam," Proc. 20th Int. Conf. companion World wide web - WWW '11, p. 93, 2011.

[14] X. Wu, W. Fan, J. Gao, Z. M. Feng, and Y. Yu, "Detecting Marionette Microblog Users for Improved Information Credibility," J. Comput. Sci. Technol., vol. 30, no. 5, pp. 1082–1096, 2015.

[15] A. Fayazi, K. Lee, J. Caverlee, and A. Squicciarini, "Uncovering Crowdsourced Manipulation of Online Reviews," Proc. 38th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR '15, pp. 233–242, 2015.

[16] Z. Wu, L. Zhang, Y. Wang, and J. Cao, "Identifying Spam in Reviews," Encycl. Soc. Netw. Anal. Min., no. December, 2017.

[17] R. Erskine, "Study: 97% Of Business Owners Say Online Reputation Management Is Important--Here's How To Keep Up," Forbes, 2018. [Online]. Available: https://www.forbes.com/sites/ryanerskine/2018/07/30/study-97-of-business-owners-say-online-reputation-management-is-important-heres-how-to-keep-up/#12d217856c02. [Accessed: 06-Oct-2018].

[18] C. Naden, "Putting the trust back into online reviews," International Organization for Standardization, 2018. [Online]. Available: https://www.iso.org/news/ref2295.html?utm_medium=email&utm_campaign=ISO Newsletter July 2018&utm_content=ISO Newsletter July 2018+CID_78edac57d13087be5b508833969a3117&utm_source=Email marketing software&utm_term=Read more. [Accessed: 03-Oct-2018].

[19] BSI, "BS ISO 20488:2018 Online consumer reviews. Principles and requirements for their collection, moderation and publication," Bsigroup.Com, 2018.

[20] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "What Yelp Fake Review Filter Might Be Doing?," Seventh Int. AAAI …, pp. 409–418, 2013.

[21] A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, and R. Ghosh, "Spotting opinion spammers using behavioral footprints," Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '13, p. 632, 2013.

[22] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting Burstiness in Reviews for Review Spammer Detection," Proc. Seventh Int. AAAI Conf. Weblogs Soc. Media, pp. 175–184, 2013.

[23] H. Liu, Y. Zhang, H. Lin, J. Wu, Z. Wu, and X. Zhang, "How Many Zombies Around You?," 2013 IEEE 13th Int. Conf. Data Min., pp. 1133–1138, 2013.

[24] R. Yu, X. He, and Y. Liu, "GLAD: Group anomaly detection in social media analysis," Trans. Knowl. Discov. from Data, vol. 10, no. 2, p. 18:1-18:21, 2015.

[25] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," Proc. 21st Int. Conf. World Wide Web - WWW '12, p. 191, 2012.

[26] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," J. Big Data, vol. 2, no. 1, 2015.

[27] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," J. Big Data, vol. 2, no. 1, p. 23, Dec. 2015.

[28] Y. Lin, T. Zhu, H. Wu, J. Zhang, X. Wang, and A. Zhou, "Towards online anti-opinion spam: Spotting fake reviews from the review sequence," ASONAM 2014 - Proc. 2014 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min., no. Asonam, pp. 261–264, 2014.

[29] M. Ott, C. Cardie, and J. T. Hancock, "Negative Deceptive Opinion Spam," Naacl-Hlt, no. June, pp. 497–501, 2013.

[30] S. Banerjee and A. Y. K. Chua, "Applauses in hotel reviews: Genuine or deceptive?," Proc. 2014 Sci. Inf. Conf. SAI 2014, pp. 938–942, 2014.

[31] N. Luo, H. Deng, L. Zhao, Y. Liu, X. Wang, and Z. Tan, "Multi-Aspect Feature Based Neural Network Model in Detecting Fake Reviews," 2017 4th Int. Conf. Inf. Sci. Control Eng., pp. 475–479, 2017.

[32] N. Kumar, D. Venugopal, L. Qiu, and S. Kumar, "Detecting Review Manipulation on Online Platforms with Hierarchical Supervised Learning," J. Manag. Inf. Syst., vol. 35, no. 1, pp. 350–380, 2018.

[33] R. Y. K. Lau, S. Y. Liao, R. C.-W. Kwok, K. Xu, Y. Xia, and Y. Li, "Text mining and probabilistic language modeling for online review spam detection," ACM Trans. Manag. Inf. Syst., vol. 2,

no. 4, pp. 1–30, 2011.
[34] J. K. Rout, S. Singh, S. K. Jena, and S. Bakshi, “Deceptive review detection using labeled and unlabeled data,” Multimed. Tools Appl., vol. 76, no. 3, pp. 3187–3211, 2017.
[35] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao, “Spotting Fake Reviews via Collective Positive-Unlabeled Learning,” Proc. - IEEE Int. Conf. Data Mining, ICDM, vol. 2015–Janua, no. January, pp. 899–904, 2015.
[36] M. Dong, L. Yao, X. Wang, B. Benatallah, C. Huang, and X. Ning, “Opinion Fraud Detection via Neural Autoencoder Decision Forest,” pp. 1–10, 2018.
[37] L. Zhang, Y. Yuan, Z. Wu, and J. Cao, “Semi-SGD: Semi-Supervised Learning Based Spammer Group Detection in Product Reviews,” Proc. - 5th Int. Conf. Adv. Cloud Big Data, CBD 2017, pp. 368–373, 2017.
[38] Y. Yao, B. Viswanath, J. Cryan, H. Zheng, and B. Y. Zhao, “Automated Crowdturfing Attacks and Defenses in Online Review Systems,” 2017.
[39] E. F. Cardoso, R. M. Silva, and T. A. Almeida, “Towards automatic filtering of fake reviews,” Neurocomputing, vol. 0, pp. 1–11, 2018.
[40] S. Feng, R. Banerjee, and Y. Choi, “Syntactic stylometry for deception detection,” Proc. 50th Annu. Meet. Assoc. Comput. Linguist. Short Pap. 2. Assoc. Comput. Linguist., no. July, pp. 171–175, 2012.
[41] H. Al Najada and X. Zhu, “ISRD: Spam review detection with imbalanced data distributions,” Proc. 2014 IEEE 15th Int. Conf. Inf. Reuse Integr. IEEE IRI 2014, pp. 553–560, 2014.