International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 12, Issue 4, (April 2016), PP.01-08

Opinion Spam Detection: A Review

Salma Farooq, Hilal Ahmad Khanday
[1]Department of Computer Science, IUST, Kashmir
[2]Assistant Professor, Department of Computer Science, University of Kashmir

ABSTRACT:-Online reviews are often the primary factor in a customer's decision to purchase a product or service, and are a valuable source of information that can be used to determine public opinion on these products or services. Opinions from social media are increasingly used by individuals and organizations for making purchase decisions and for marketing and product design. Because of their impact, manufacturers and retailers are highly concerned with customer feedback and reviews. Reliance on online reviews gives rise to the potential concern that wrongdoers may create false reviews to artificially promote or devalue products and services. This practice is known as Opinion (Review) Spam, where spammers manipulate and poison reviews (i.e., making fake, untruthful, or deceptive reviews) for profit or gain. Positive opinions often mean profits and fames for businesses and individuals, which, unfortunately, give strong incentives for people to game the system by posting fake opinions or reviews to promote or to damage the reputation of some target products, services, organizations, individuals, and even ideas without disclosing their true intentions, or the person or organization that they are secretly working for. Opinion spamming about social and political issues can even be frightening as they can warp opinions and mobilize masses into positions counter to legal or ethical mores. However, they must be detected in order to ensure that the social media continues to be a trusted source of public opinions, rather than being full of fake opinions, lies, and deceptions. Since not all online reviews are truthful and trustworthy, it is important to develop techniques for detecting review spam. In this paper, we survey the prominent machine learning techniques that have been proposed to solve the problem of review spam detection and the performance of different approaches for classification and detection of review spam. The primary goal of this paper is to provide a strong and comprehensive comparative study of current research on detecting review spam using various machine learning techniques

KEYWORDS:-Opinion Spam, review Spam, supervised spam detection, unsupervised spam detection, fake reviews

I. INTRODUCTION

The Web has dramatically changed the way that people express themselves and interact with others. They can now post reviews of products at merchant sites and express their views and interact with others via blogs and forums. Such content contributed by Web users is collectively called the user-generated content .It is now well recognized that the user generated content contains valuable information that can be exploited for many applications. Online reviews are often the primary factor in a customer's decision to purchase a product or service, and are a valuable source of information that can be used to determine public opinion on these products or services. Nowadays, almost everyone views online reviews before deciding on a restaurant, hotel, buying a product, or even choosing a travel destination. They are used by potential customers to find opinions of existing users before deciding to purchase a product. They are also used by product manufacturers to identify product problems and/or to find marketing intelligence information about their competitors. However, with its usefulness, it brings forth a curse opinion spam. Reliance on online reviews gives rise to the potential concern that wrongdoers may create false reviews to artificially promote or devalue products and services. They must be detected in order to ensure that the social media continues to be a trusted source of public opinions, rather than being full of fake opinions, lies, and deceptions. This practice is known as Opinion (Review) Spam, where spammers manipulate and poison reviews (i.e., making fake, untruthful, or deceptive reviews) for profit or gain. By extracting meaningful features from the text using Natural Language Processing (NLP), it is possible to conduct review spam detection using various machine learning techniques. Additionally, reviewer information, apart from the text itself, can be used to aid in this process. In this paper, we focus on customer reviews of products. In particular, we investigate *opinion spam* in reviews.

1.1 TYPES OF SPAM:

Spam detection in general has been studied in many fields. Web spam and email spam are the two most widely studied types of spam. Opinion spam is, however, very different. There are two main types of Web spam:

a). **Link spam***:* Link spam is spam on hyperlinks, which hardly exist in reviews. Although advertising links are common in other forms of social media, they are relatively easy to detect.
b). **Content Spam***:*Content spam adds popular (but irrelevant) words in target Web pages in order to fool search engines to make them relevant to many search queries, but this hardly occurs in opinion postings. Email spam refers to unsolicited advertisements, which are also rare in online opinions

The key challenge of opinion spam detection is that unlike other forms of spam, it is very hard to recognize fake opinions by manually reading them, which makes it difficult to find opinion spam data to help design and evaluate detection algorithms. In fact, in the extreme case, it is logically impossible to recognize spam by simply reading it. For example, one can write a truthful review for a good product and post it as a fake review for a bad product in order to promote it. There is no way to detect this fake review without considering information beyond the review text itself simply because the same review cannot be both truthful and fake at the same time.

1.2 TYPES OF SPAM REVIEWS:

Three types of spam reviews have been identified [1]:
a). Type 1 (Fake/Untruthful Reviews): These are untruthful reviews that are written not based on the reviewers' genuine experiences of using the products or services, but are written with hidden motives. They often contain undeserving positive opinions about some target entities (products or services) in order to promote the entities and/or unjust or false negative opinions about some other entities in order to damage their reputations. These are also called as bogus reviews.
b). Type 2 (Reviews about brands only): These reviews do not comment on the specific products or services that they are supposed to review, but only comment on the brands or the manufacturers of the products. Although they may be genuine, they are considered as spam as they are not targeted at the specific products and are often biased. For example, a review for a specific PNG product says "*I hate PNG, I never buy any of their products*".
c). Type 3 (Non-Reviews): These are not reviews. There are two main subtypes:
i). Advertisements
ii). Other irrelevant texts containing no opinions (e.g., questions, answers, and random texts).

II. SPAM DETECTION

It has been shown in [1] that Types 2 and 3 spam reviews are rare and relatively easy to detect using supervised learning. Even if they are not detected, it is not a major problem because human readers can easily spot them during reading. Fake reviews (Type 1) can be seen as a special form of deception [2] [3] [4].Researchers have identified many deception signals in text. For spam reviews of Type 2 and Type 3, we can detect them based on traditional classification learning using manually labeled spam and non-spam reviews because these two types of spam reviews are recognizable manually. The main task is to find a set of effective features for model building. However, for the first type of spam, manual labeling by simply reading the reviews is very hard, if not impossible, because a spammer can carefully craft a spam review to promote a target product or to damage the reputation of another product that is just like any other innocent review. For example, a spam review that praises a product that every reviewer likes (gives a high rating) is not very damaging. However, a spam review that criticizes a product that most people like can be very harmful. However, the problem is that there is no labeled training example. Using them to build spam detection models can predict those likely harmful reviews to a great extent.

In general, spam detection can be regarded as a classification problem with two classes: *spam* and *non-spam*. Machine learning models can be built to classify each review as spam or non-spam, or to give a probability likelihood of each review being a spam. To build a classification model, one needs labeled training examples of both spam reviews and non-spam reviews. That is where we have a problem. For the three types of spam, we can only manually label training examples for spam reviews of type 2 and type 3 as they are recognizable based on the content of a review. However, recognizing whether a review is an untruthful opinion spam (Type1) is extremely difficult by manually reading the review because one can carefully craft a spam review which is just like any other innocent review. Therefore, other ways have to be explored in order to find training examples for detecting possible Type 1 spam reviews.

2.1 DETECTING TYPE 2 AND TYPE 3 SPAM REVIEWS

Type 2 and Type 3 types of reviews spams are recognizable manually. The following parameters are used:
a) Review content: The actual text content of each review. From the content, we can extract *linguistic features* such as word and POS n-grams and other syntactic and semantic clues for deceptions and lies. However,

linguistic features may not be enough because one can fairly easily craft a fake review that is just like a genuine one. For example, one can write a fake positive review for a bad beauty product based on her true experience of using a good product.

b) Meta-data about the review: The data such as the star rating given to each review, user-id of the reviewer, the time when the review was posted, the time taken to write the review, the host IP address and MAC address of the reviewer's computer, the geo-location of the reviewer, and the sequence of clicks at the review site. From such data, we can mine many types of abnormal *behavioral patterns* of reviewers and their reviews. For example, from review ratings, we may find that a reviewer wrote only positive reviews for a brand and only negative reviews for a competing brand. Along a similar line, if multiple user-ids from the same computer posted a number of positive reviews about a product, these reviews are suspicious. Also, if the positive reviews for a hotel are all from the nearby area of the hotel, they are clearly not trustworthy.

c) Product information: Information about the entity being reviewed, e.g., the product description and sales volume/rank. For example, a product is not selling well but has many positive reviews, which is hard to believe.

III. STRATEGIES USED FOR REVIEW SPAM DETECTION

The ultimate goal of opinion spam detection in the review context is to identify every fake review, fake reviewer, and fake reviewer group. The three concepts are clearly related as fake reviews are written by fake reviewers and fake reviewers can form fake reviewer groups. The detection of one type can help the detection of others. However, each of them also has its own special characteristics, which can be exploited for detection.
There are two different strategies that have been discussed so far:

a) Supervised Spam Detection
b) Unsupervised Spam Detection

3.1 SUPERVISED SPAM DETECTION:

In general, opinion spam detection can be formulated as a classification problem with two classes: fake and non-fake. Supervised learning is naturally applicable. However, as we described above, a key difficulty is that it is very hard, if not impossible, to recognize fake reviews reliably by manually reading them because a spammer can carefully craft a fake review that is just like any innocent review [5]. Due to this difficulty, there is no reliable fake review and non-fake review data available to train a machine learning algorithm to recognize fake reviews. Despite these difficulties, several detection algorithms have been proposed and evaluated in various ways. Due to the fact that there is no labeled training data for learning, [5] exploited duplicate reviews. In their study of 5.8 million reviews and 2.14 million reviewers from amazon.com, a large number of duplicate and near-duplicate reviews were found, which indicated that review spam was widespread. Since writing new reviews can be taxing, many spammers use the same reviews or slightly revised reviews for different products.

These duplicates and near-duplicates can be divided into four categories:

a) Duplicates from the same user-id on the same product
b) Duplicates from different user-ids on the same product
c) Duplicates from the same user-id on different products
d) Duplicates from different user-ids on different products

The first type of duplicates can be the results of reviewers mistakenly clicking the review submit button multiple times (which can be easily checked based on the submission dates). However, the last three types of duplicates are very likely to be fake. Thus the last three types of duplicates were used as fake reviews and the rest of the reviews as non-fake reviews in the training data for machine learning. Three sets of features have been employed:

3.1.1 REVIEW CENTRIC FEATURES:

These are features about each review and include the following features:
a). Number of feedbacks, number of helpful feedbacks and percent of helpful feedbacks that the review gets. Intuitively, feedbacks are useful in judging the review quality.
b). Length of the review title and length of review body. Since longer reviews tend toget more helpful feedbacks and customer's attention, a spammer might want to use this to his/her advantage.
c). Position of the review in the reviews of a product sorted by date, in both ascending and descending order. It has been found that reviews which are written early tend to get more user attention, and thus can have bigger impact on the sale of a product.
d). Textual features:

Percent of positive and negative opinion-bearing words in the review, e.g., "awesome", "great", "terrible" and "poor". Type 2 reviews would use these words excessively to praise or to criticize the brand or the manufacturer.
Cosine similarity of the review and product features. This feature is useful for detecting Type 3 reviews, particularly advertisements.
Percent of times brand name is mentioned in the review. This feature has been used for reviews which praise or criticize the brand.
e). Rating related features:
Rating of the review and its deviation from product rating. Feature indicating if the review is good, average or bad.
Binary features indicating whether a bad review was written just after the first good review of the product and vice versa.

3.1.2 REVIEWER CENTRIC FEATURES:

These are the features about each reviewer and include the following:
a). Ratio of the number of reviews that the reviewer wrote which were the first reviews of the products to the total number of reviews that he/she wrote, and ratio of the number of cases in which he/she was the only reviewer
b). Rating related features: average rating given by reviewer standard deviation in rating and a feature indicating if the reviewer always gave only good, average or bad rating
3. Binary features indicating whether the reviewer gave more than one type of rating, i.e. good, average and bad.
c). The ratio of the number of cases in which he/she was the only reviewer.

3.1.3 PRODUCT CENTRIC FEATURES:

The product related features are as follows:
a). Price of the product.
b). Sales rank of the product.
c). Average rating and standard deviation in ratings of the reviews on the product.

These features have been helpful since spams could be concentrated on cheap/expensive or less selling products.
Experimental results conducted by various researchers have shown some tentative but interesting results:

a). Negative outlier reviews (ratings with significant negative deviations from the average rating of a product) tend to be heavily spammed.
b). Positive outlier reviews are not badly spammed.
c). Reviews that are the only reviews of some products are likely to be fake. This can be explained by the tendency of a seller promoting an unpopular product by writing a fake review.
d). Top-ranked reviewers are more likely to be fake reviewers. Amazon.com gives a rank to each reviewer based on its proprietary method. Analysis showed that top-ranked reviewers generally wrote a large number of reviews. People who wrote a large number of reviews are natural suspects. Some top reviewers wrote thousands or even tens of thousands of reviews, which is unlikely for an ordinary consumer.
e). Fake reviews can get good feedbacks and genuine reviews can get bad feedbacks. This shows that if the quality of a review is defined based on helpfulness feedbacks, people can be fooled by fake reviews because spammers can easily craft a sophisticated review that can get many positive feedbacks.
f). Products of lower sales ranks are more likely to be spammed. This indicates that spam activities seem to be limited to low selling products, which is intuitive as it is difficult to damage the reputation of a popular product, and an unpopular product needs some promotion.

These results are tentative as it has not been confirmed whether the three types of duplicates are definitely fake reviews or not, and that many fake review are not duplicates and are considered as Non-fake reviews.

In [6], another supervised learning approach has been attempted to identify fake reviews. In their case, a manually labeled fake review corpus was built from Epinions reviews. In Epinions, after a review is posted, users can evaluate the review by giving it a helpfulness score. They can also write comments about the reviews. The authors manually labeled a set of fake or non-fake reviews by reading the reviews and the comments. For learning, several types of features have been proposed, which are similar to those in [1] with some additions,

e.g., subjective and objectivity features, positive and negative features, reviewer's profile, authority score computed using PageRank [7] etc. For learning, they used naive Bayesian classification which gave promising results. The authors also experimented with a semi-supervised learning method exploiting the idea that a spammer tends to write many fake reviews.

In [4], supervised learning was also employed. In this case, the authors used Amazon Mechanical Turk to crowd source fake hotel reviews of 20 hotels. Several provisions were made to ensure the quality of the fake reviews. Several classification approaches have been tried which have been used in related tasks such as genre identification, psycholinguistic deception detection, and text classification. All these tasks have some existing features proposed by researchers. Their experiments showed that text classification performed the best using only unigram and bigrams based on the 50/50 fake and non-fake class distribution. Furthermore, using 50/50 fake and non-fake data for testing may not reflect the true distribution of the real-life situation. The class distribution can have a significant impact on the precision of the detected fake reviews.

3.2 UNSUPERVISED SPAM DETECTION

Due to the difficulty of manually labeling of training data, using supervised learning alone for fake review detection is difficult. So, some un-supervised spam detection methods have been proposed.

3.2.1 GENERATIVE MODEL

Opinion Spam detection has usually been modeled as an instance of unsupervised Bayesian clustering with two clusters, spam and non-spam. The Bayesian setting conveniently allows treating spamicity of authors/reviews as latent variables. Specifically, the spam/non-spam category of a review is modeled as a latent variable This can be seen as the category/class variable reflecting the cluster memberships of every review. The Latent Spam Model (LSM- Arjun, Vivek) belongs to the class of generative models for clustering [8] [9]. Each review of an author is represented with a set of observed linguistic and behavioral features which are emitted conditioned on the latent spam/non-spam category variable and associated distributions. This is achieved using posterior inference techniques (e.g., Markov Chain Monte Carlo) for probabilistic model-based clustering. The stationary distributions of class/category assignments is used for generating clusters of spam (fake) and non-spam (non-fake) reviews.

Features Observed: Linguistic n-grams have been shown to be useful for deception detection [4]. The behavioral features are constructed from various abnormal behavioral patterns of reviewers and reviews. The following features have been observed:

a). Author Features: The following author features have been proposed wherein the values close to 0 or 1 indicate non-spamming/spamming respectively:

i). Content Similarity: Spammers typically post fake experiences. However, as crafting a new fake review every time is time consuming, they often post reviews which are duplicate/near-duplicate versions of their previous reviews [1]. It is naturally useful to capture the maximum content similarity (using cosine similarity) across any pair of reviews by an author/reviewer. The maximum similarity is used to capture the worst spamming behavior.

ii). Maximum Number of Reviews: Posting many reviews in a single day reflects abnormal reviewing pattern and can be used as a behavioral feature. This feature simply computes the maximum number of reviews posted in a day for an author. It is normalized by the maximum value in the dataset.

iii). Reviewing Activity: The study in [10], [11], [12] reports that opinion spammers are usually not long time members of a site. Genuine reviewers, however, use their accounts from time to time to post reviews over a considerably long period of time. It is thus useful to exploit the activity freshness of an account to detect spamming. The activity of an author is computed by measuring the difference of his first and last review posting dates. This feature is normalized by the maximum value in the available dataset. This activity feature indicates that authors posting reviews over a reasonably long time frame are less likely to be spamming than those who just created their accounts to some post specific (probably deceptive/spam) reviews and do not ever use that account afterwards.

b). Review Features: Following are the review features which can be used as indicators wherein values close to 0 or 1 indicate non-spamming/spamming respectively:
i). Extreme Rating: Opinion spamming typically projects entities incorrectly either in a very positive or a very negative light [5]. Thus, on a 5-star rating scale, spammers are likely to give extreme ratings (1 or 5 stars) in order to promote or to demote entities.

ii). Rating Deviation: It has been observed that the ratings of spammers deviate from the average ratings given by other genuine reviewers. On a 5-star scale, the absolute rating deviation of a review from the general rating consensus (average rating of the entity) can be between 0 and 4.

iii). Early Time Frame: Lim et al. [10] noted that spammers often review early to inflict spam as the early reviews can greatly impact the people's sentiment on the entity. To capture this spamming characteristic, it is measure whether a review is posted within some early time frame.

3.2.2 SPAM DETECTION BASED ON ATYPICAL BEHAVIORS

This sub-section describes some techniques that try to discover atypical behaviors of reviewers for spammer detection. For example, if a reviewer wrote all negative reviews for a brand but other reviewers were all positive about the brand, and wrote all positive reviews for a competing brand, then this reviewer is naturally suspicious. The first technique is from [10], which identified several unusual reviewer behavior models based on different review patterns that suggest spamming. Each model assigns a numeric spamming behavior score to a reviewer by measuring the extent to which the reviewer practices spamming behavior of the type. All the scores are then combined to produce the final spam score. Thus, this method focuses on finding spammers or fake reviewers rather than fake reviews. The spamming behavior models are:

(a) **Targeting products**: To game a review system, it is hypothesized that a spammer will direct most of his efforts on promoting or victimizing a few target products. He is expected to monitor the products closely and mitigate the ratings by writing fake reviews when time is appropriate.

(b) **Targeting groups**: This spam behavior model defines the pattern of spammers manipulating ratings of a set of products sharing some attribute(s) within a short span of time. For example, a spammer may target several products of a brand within a few hours. This pattern of ratings saves the spammers' time as they do not need to log on to the review system many times. To achieve maximum impact, the ratings given to these target groups of products are either very high or very low.

(c) **General rating deviation**: A genuine reviewer is expected to give ratings similar to other raters of the same product. As spammers attempt to promote or demote some products, their ratings typically deviate a great deal from those of other reviewers.

(d) **Early rating deviation**: Early deviation captures the behavior of a spammer contributing a fake review soon after product launch. Such reviews are likely to attract attention from other reviewers, allowing spammers to affect the views of subsequent reviewers.

The second technique has also been proposed that focuses on finding fake reviewers or spammers [13], [14],[15]. Here the problem was formulated as a data mining task of discovering unexpected class association rules. Unlike conventional spam detection approaches such as the above supervised and unsupervised methods, which first manually identify some heuristic spam features and then use them for spam detection. This technique is generic and can be applied to solve a class of problems due to its domain independence. Class association rules are a special type of association rules with a fixed class attribute. The data for mining class association rules (CARs) consists of a set of data records, which are described by a set of normal attributes $A = \{A1, A2,.., An\}$, and a class attribute $C = \{c1, c2,.., cm\}$ of m discrete values, called *class labels*. A CAR rule is of the form: $X \mid ci$, where X is a set of conditions from the attributes in A and ci is a class label in C. Such a rule computes the conditional probability of $\Pr(ci \mid X)$ (called *confidence*) and the joint probability $\Pr(X, ci)$ (called *support*). For the spammer detection application, the data for CAR mining is produced as follows: Each review forms a data record with a set of attributes, e.g., *reviewer-id*, *brand-id*, *product-id*, and a class. The class represents the sentiment of the reviewer on the product, *positive*, *negative*, or *neutral* based on the review rating. In most review sites (e.g., amazon.com), each review has a rating between 1 (lowest) and 5 (highest) assigned by its reviewer. The rating of 4 or 5 is assigned positive, 3 neutral, and 1 or 2 negative. A discovered CAR rule could be that a reviewer gives all positive ratings to a particular brand of products. The method in [5] finds four types of unexpected rules based on four unexpectedness definitions. The unexpected rules represent atypical behaviors

of reviewers. Below, an example behavior is given for each type of unexpectedness definition. The unexpectedness definitions are quite involved and can befound in [1], [5].

Confidence unexpectedness: Using this measure, one can find reviewers who give all high ratings to products of a brand, but most other reviewers are generally negative about the brand.

Support unexpectedness: Using this measure, one can find reviewers who write multiple reviews for a single product, while other reviewers only write one review.

Attribute distribution unexpectedness: Using this measure, one can find that most positive reviews for a brand of products are written by only one reviewer although there are a large number of reviewers who have reviewed the products of the brand.

Attribute unexpectedness: Using this measure, one can find reviewers who write only positive reviews to one brand and only negative reviews to another brand.

The advantage of this approach is that all the unexpectedness measures are defined on CARs rules, and are thus domain independent. The technique can thus be used in other domains to find unexpected patterns. The weakness is that some atypical behaviors cannot be detected, e.g., time-related behaviors, because class association rules do not consider time. It is important to note that the behaviors studied in published papers are all based on public data displayed on review pages of their respective review hosting sites. As mentioned earlier, review hosting sites also collect many other pieces of data about each reviewer and his/her activities at the sites. These data are not visible to the general public, but can be very useful, perhaps even more useful than the public data, for spam detection. For example, if multiple user-ids from the same IP address posted a number of positive reviews about a product, then these user-ids are suspicious. If the positive reviews for a hotel are all from the nearby area of the hotel, they are also doubtful. Some review hosting sites are already using these and other pieces of their internal data to detect fake reviewers and reviews. The idea is that fake reviews will distort the overall popularity ranking for a collection of entities. That is, deleting a set of reviews chosen at random should not overly disrupt the ranked list of entities, while deleting fake reviews should significantly alter or distort the ranking of entities to reveal the “true" ranking. This distortion can be measured by comparing popularity rankings before and after deletion using rank correlation.

IV. LINGUISTIC TRACES OF DECEPTION:

Studies on psycholinguistic deception, however, state that lying/deceptive communications usually have fewer personal/first person pronouns. It is worthwhile here to understand the difference. Writing fake opinions/reviews on the Web is a distinctive cognitive/psychological process and not the same as conventional lying. Traditional lying/deceptive communications refers to statements of untrue facts. It involves the psychological process of “detachment” resulting in the use of fewer first-person pronouns. , studies have shown that when people lie they tend to detach themselves and like to use words such as *you*, *she*, *he*, *they*, rather than *I*, *myself*, *mine*, etc. Liars also use words related to certainty more frequently to hide “fake” or to emphasize “truth”. This phenomenon has been attested by researchers [16], [17] that liars tend to avoid statements of ownership to “dissociate” themselves resulting in less usage of first-person/personal pronouns. Fake reviews/opinions on the Web differ from conventional lies in two keys aspects. First, fake reviewers actually like to use more first-person pronouns such as *I*, *myself*, *mine*, *we*, *us*, etc., to make their reviews sound more convincing and to give readers the impression that their reviews express their “own” true experiences. We call this “attachment” as opposed to “detachment”. Second, fake reviews may not be traditional lies of facts. For instance, an author of a book can pretend to be a reader of the book and write a review, or fake reviewers reviewing a product they never used, etc. Thus, we see that deceptive opinion spam on the Web has subtle differences and complexities than traditional lying or deception as studied in the psycholinguistic literature. Fake review detection is thus a challenging problem. The various models have shown promising results on multiple domains/datasets. Additionally, if richer internal/private data from websites (e.g., IP addresses, geo-location, session/network/click logs, mouse gestures, etc.) are available, more behaviors can be modeled which can significantly improve the detection accuracy.

V. CONCLUSION

This paper studied opinion spam in reviews. The paper identified three types of spam. Detection of such spam is done first by detecting duplicate reviews. Then the methods for detecting Type 2 and Type 3 spam reviews by using supervised learning with manually labeled training examples was discussed. Results have shown that the logistic regression model is highly effective. However, it is difficult to detect type 1 opinion

spam because it is very hard to manually label training examples for Type 1 spam. The current study, however, only represents an initial investigation. Much work remains to be done. This paper also discussed the various suggested ways to utilize linguistic and behavioral clues to detect deceptive opinion spam (fake reviews) in an unsupervised Bayesian inference framework. The LSM model treats opinion spam detection as a clustering problem. Learning exploits distributional divergence on linguistic and behavioral dimensions between spammers (fake reviewers) and other (non-spammers). The fully Bayesian approach facilitates modeling spamicity of authors and reviews as latent variables precluding the need of any labeled data. Thus, in this paper, opinion spam detection was discussed keeping in consideration both the supervised and unsupervised methodologies. This topic needs further research as very little has been done in this perspective and there is dire need to further improve the detection methods, and also look into spam in other kinds of media, e.g., forums, blogs or even more. A lot needs to be unveiled yet in this regard.

REFERENCES

[1]. Jindal, N., & Liu, B., "Identifying Comparative Sentences in Text Documents",SIGIR, 2006.

[2]. Ott, M., Cardie, C., & Hancock, J. T.,"Negative Deceptive Opinion Spam", In Proceedings of NAACL-HLT, (pp. 497-501), 2013.

[3]. Ott, M., Cardie, C. and Hancock, J., "Estimating the Prevalence of Deception in Online Review Communities", Proceedings of the 21st international conference on World Wide Web, (WWW), 2012.

[4]. Ott, M., Choi, Y., Cardie, C. Hancock, J., "Finding Deceptive Opinion Spam by Any Stretch of the Imagination", Association of Computational Linguistics (ACL), 2011.

[5]. Jindal, N., and Liu, B. , "Opinion Spam and analysis", Proceedings of the International Conference on Web search and web data mining (WSDM),2008.

[6]. Li, F., Huang, M., Yang, Y. and Zhu, X. , "Learning to Identify Review Spam", In Proceedings of International Joint Conference of Artificial Intelligence (IJCAI), 2011.

[7]. PageL. et al., "The PageRank Citation Ranking: Bring Order to the Web", Tech. report, Stanford Digital Library Technologies, Jan. 1998.

[8]. Berkhin P., "Survey of clustering data mining techniques", Technical Report, Accrue Software, San Jose, CA, 2002.

[9]. TatemuraB., Wu Y., "Tomographic Clustering to Visualize Blog Communities as Mountain Views", In Proc. Of WWW Conference, Japan, 10- 14 May, 2005.

[10]. Lim, E. P., Nguyen, V. A., Jindal, N., Liu, B., & Lauw, H. W. ,"Detecting Product Review Spammers Using Rating Behaviors", Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM), 2010.

[11]. Lauw,H.W.,Lim,E.P.,Wang,K, "Bias and Controversy: Beyond the Statistical Deviation",ACM SIGKDD international conference on Knowledge discovery and data mining (KDD). 2006.

[12]. Lauw, H.W., Lim, E.P., Wang,K. , "Summarizing Review Scores of Unequal Reviewers", In Proceedings of the SIAM conference in Data Mining (SDM), 2007.

[13]. Jindal, N. & Liu, B., "Review Analysis", Tech. Report, 2007.

[14]. Li, K., & Zhong, Z. , "Fast Statistical Spam Filter by Approximate Classifications",SIGMETRICS, 2006.

[15]. Liu, B. , "Web Data Mining" Springer, 2007.

[16]. Knapp, M. L., Hart, R. P., & Dennis, H. S., "An Exploration of Deception as a Communication Construct",Human Communication Research, 1, 15-29,1974.

[17]. Luca, M., & Zervas., G. "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud", Harvard Business School NOM Unit Working Paper, 2013.