Review Spam Detection using Machine Learning

Draško Radovanović, Božo Krstajić, *Member, IEEE*

Abstract **--- Prior to buying a product, people usually inform themselves by reading online reviews. To make more profit sellers often try to fake user experience. As customers are being deceived this way, recognizing and removing fake reviews is of great importance. This paper analyzes spam detection methods, based on machine learning, and presents their overview and results.**

***Keywords* — machine learning, review spam detection.**

I. Introduction

MACHINE learning is a field of computer science that allows computers to learn from data without being explicitly programmed. Supervised learning, a subfield of machine learning, needs labeled data to be able to learn. Data is labeled by human experts or some system whose behavior should be mimicked. During the training process, algorithm tries to find relationship between input (data) and output (labels). After the training, system can be used on unlabeled data. Algorithms used by methods in this paper belong to supervised learning algorithms.

As Internet continues to grow, online reviews are becoming more relevant source of information. Knowing that products' success depends on customer reviews, sellers often try to deceive buyers by posting fake comments. Sellers can post reviews themselves or pay other individuals to do it for them. This practice of posting fraudulent reviews is known as opinion or review spam.

Spammers can be hired to post positive reviews, or to write bad reviews to damage competitors' business. Canadian Competition Burau issued an official warning to their citizens in 2014, stating that they should be aware of fraudulent reviews and estimating that third of reviews found online are fake [1]. Poll taken on over 25 000 participants in 2009, says that over 70% consumers believe online reviews [2]. This shows that spam reviews present a major concern today. To tackle this problem many methods have been proposed during the last decade.

Majority of published papers can be categorized in three groups, based on the aim of proposed methods. Methods can be used for detecting spam reviews, individual spammers, or group spam [3]. Since methods that focus on group spam are not well researched, they are not presented in this paper. Description of data used in spam detection is given in section II. Frequently used techniques and their results are presented in section III. Experimental results are given in section IV.

D. Radovanović is with the Faculty of Electrical Engineering, University of Montenegro, Džordža Vašingtona bb, 81000 Podgorica, Montenegro (e-mail: draskor@ac.me).

B. Krstajić is with the Faculty of Electrical Engineering, University of Montenegro, Džordža Vašingtona bb, 81000 Podgorica, Montenegro (e-mail: bozok@ac.me).

II. Review spam description

Review spam can be divided in three groups as proposed in [4]:

1. Untruthful opinions
2. Reviews on brands only
3. Non-reviews

Untruthful opinions represent purposefully fake reviews. Reviews on brands only aren't focused on products, but rather on brands or manufacturers. Non-reviews include advertisements or other irrelevant reviews containing no opinions. Although types two and three fail to address specific products, they aren't fraudulent. These types of spam are also easy to spot manually and traditional classification approaches have no problem in detecting them. Untruthful reviews are shown to be much harder task for a machine as well as for a human observer. For those reasons this type of spam is considered in this paper.

Data used in techniques for review spam detection can be categorized in three types of data [5]:

1. Review content
2. Meta-data about the review
3. Product information

Review content represents the actual text of the review. Linguistic features used in detecting fraudulent behavior can be extracted from the review content. Those include POS (*position of speech)* and word n-grams, as well as other semantic and syntactic clues. The problem with using solely review content is that it is easy to create a fake review that is like a genuine one.

Review meta-data includes information about user id, its ratings, IP and MAC addresses of host computers, information about time it took to write a review, reviewer's location etc. Meta-data is useful in detecting abnormal behavior patterns. However, as opposed to review content data, it is not available on many websites.

Product information is information about merchandise being reviewed such as number of reviews, average ratings, product description, popularity and sales volume. For example, if certain product is not selling well, but has a lot of praising reviews, that can be suspicious.

Data used in training and testing systems for spam detection comes from variety of sources. Lack of standard

978-1-5386-3620-6/18/$31.00 ©2018 IEEE

datasets for this type of problem makes it hard to compare results from different papers. Some of the commonly used datasets are the ones obtained from Yelp.com [21] or by Amazon Mechanical Turk (AMT) [20]. AMT is a crowdsourcing platform that was used by researchers in [14] for creating a dataset of fake reviews. It is shown that reviews created via AMT do not accurately represent real-world data. Yelp offers real data from their site, however, labeling is based on their own spam filters which makes this dataset unreliable. Labeling reviews manually, as in many other machine learning problems, is not helpful. The main reason for this is that it is shown that machine learning techniques used today, although not sufficiently good, are better than humans in recognizing fraudulent reviews [6]. Many researchers have artificially created fake reviews used in their research, which further complicates comparison of their results with other papers.

III. METHODS OVERVIEW

A. Review centric methods

To the best of our knowledge, the first method for review spam detection was proposed in 2007 by Jindal and Liu [8]. This paper was followed by [9] and [4] in which initial ideas were further investigated. In [4] method for detecting spam based on review duplication was proposed. The data was taken from amazon.com and included 5.18 million reviews and 2.14 million reviewers. They used a shingle method [10] for detecting near-duplicate reviews. Authors calculated a similarity score and labeled the ones who had a score over 90% as duplicates. After that 36 features describing reviews, reviewers and product information were extracted. They tried Naive Bayes, support vector machine (SVM) and logistical regression. Naive Bayes and SVM yielded bad results. Using logistical regression, they obtained AUC (Area Under ROC Curve) score of 0.63 using only text features and 0.78 using all features, showing that more features than just a review content is needed for a good classification model

Researchers observed that many spammers copy existing reviews entirely or change only a few words. Therefore, many researchers in this area have focused on methods for duplicates detection. These methods are used for finding textual or conceptual similarity between reviews. In [11] authors used conceptual similarity. They used data collected from a digital camera page and extracted its main features (photo quality, design, zoom, size etc.). Then for each review they extracted which features were mentioned and in which context. Using those features they calculated similarities between reviews and compared them. Using labels obtained by two human observers as ground truth, their method achieved only 43.6% accuracy. Method for measuring text similarity based on Kullback-Leibler was proposed in [12]. They used SVM algorithm for classification and reported similar results as [4].

Detecting spam using similarity between reviews can be a useful technique. However, it should be noted that spammers often copy genuine reviews. Using these techniques both genuine and fake review would be classified as spam. Besides duplicates detection, there are many other types of techniques in detecting spam by using review content. Nice breakdown of categories is given in [6]:

Bag of words approach considers words or sequence of words used in reviews as features. Sequences of words are called *n-grams* (where *n* denotes number of words in a sequence). Values of *n=1,2,3* are the most common.

Term frequency includes n-grams as well as the number of their occurrences. This additional information can improve bag of words approach.

POS tags are labels given to words based on their context. This process includes tagging a word based on its definition and relationships with adjacent words. Words are then marked as adverbs, verbs, etc. This information is collected and with additional features fed into a machine learning algorithm.

Stylometric features try to capture reviewers writing style. They include number of punctuation marks used, length of words and sentences etc.

Semantic features are focused on the meaning of words. They include synonyms and similar phrases. The idea of using these features is that spammers usually replace some words with similar ones, conveying the message, while making it harder to identify duplicate reviews.

LIWC software [13] is also commonly used in feature engineering for spam detection. LIWC analyses text and groups words in over 80 topical, linguistic and psychological categories. Including LIWC output along with other features has been shown to improve results.

Review meta-data analysis includes information such as review length, duration of writing, reviewer id etc. These features are used both in review centric and reviewer centric methods.

In [14] researchers developed three methods for detecting spam. They used content-based approach achieving almost 90% accuracy on their dataset. For feature extraction they tried POS, LIWC output and n-grams, as well as combining their results. Classification algorithms used were SVM and Naive Bayes. Using these features SVM outperformed Naïve Bayes.

Classifier	Features	Accuracy
SVM	POS	73.0%
	LIWC	76.8%
	Unigrams	88.4%
	Trigrams	89.0%
	Bigrams	89.6%
	LIWC+bigrams	89.8%
Human	Observer 1	61.9%
	Observer 2	56.9%
	Observer 3	60.6%

Table 1: Accuracy of methods tested in [14]

Results in table 1 show that combination of bigrams and LIWC yields best results (although only 0.2% better than using only bigrams). Human observers performed poorly, achieving accuracy of around 60%. Authors created publicly available dataset with 400 genuine and 400 fake reviews. Fake reviews in their dataset were obtained via AMT, while genuine reviews were obtained from TripAdvisor. All reviews in this dataset represent positive reviews. In [15] the same authors tackled review spam that contained negative reviews. They collected 400 truthful

reviews from six websites, and the same number of fake reviews was collected using AMT. N-grams were used as features and SVM as a classification algorithm. Achieved score was 86% accuracy. They also tested how system trained on negative reviews behaved when given positive reviews and vice versa. System trained on negative reviews had 81.4% accuracy when tested on positive reviews, and 75.1% in the other case. System trained on all samples gave 88.4% accuracy in the case of positive reviews and 86% in the case of negative reviews. Like in [14], three human observers were given task to try to detect which reviews were fake. Their results were 65%, 61.9% and 57.5% accuracy, showing that automatic methods outperform humans by huge margin.

Dataset used in [14] and [15] contains fake reviews written for the purpose of these researches, and not from real websites. This includes bias and does not represent real-world data which was observed by researchers in [16]. Authors tested methods presented in [14] on real data from Yelp.com, as well as on AMT dataset. Using AMT dataset they achieved 88.8% accuracy, while on Yelp dataset algorithm correctly classified only 67.6% reviews. Furthermore, researchers in [17] created a method for synthetizing fake reviews from genuine ones. Authors claim that even the best algorithms in spam detection had error rate higher than 30%. Method presented in [14] was tested on these reviews and achieved accuracy of only 59.5%, as opposed to 89.8% reported on AMT dataset. These results show that data in AMT dataset is flawed and can't be used to obtain reliable results.

B. Reviewer centric methods

Spammers often have similar behavior patterns which can make their detection easy. Spammer behavior is analyzed, and some useful features are extracted and described in [18] and [19]:

Daily Number of Reviews: Large number of reviews written in one day by a single user is indication of a spammer. Most spammers (75%) write more than 5 reviews a day, while 90% of non-spammers write 3 or less reviews per day, and 50% writes one review per day.

Positive Review Percentage: Positive reviews are defined as reviews with four or five stars rating. Analyzing data from non-spammers, it was shown that percentage of positive reviews was uniformly distributed among users. On the other hand, about 85% of spammers had 80% or more positive reviews.

Review Length: As spammers are paid by number of spam messages they post, they tend to write shorter reviews to maximize their profit. The average review length of 92% of regular users is over 200 words while only 20% of spammers write reviews over 135 words.

Reviewer deviation: Considering that spammers are usually giving product ratings that are either high or low, it is expected that their ratings are different than average ratings. In [18], authors calculated absolute rating deviation of a review from other reviews of the same product. Then they calculated expected ratings deviation for users across all their reviews. Approximately 70% of non-spammer users had deviation less then 0.6, while over 80% of spammers had deviation greater than 2.5.

Early rating deviation: When product is published, sellers try to promote it from the very start to get attention. Because of that, spammers are the most active right after the product is published Calculating average rating of a product and using two features: review rating and weights of that rating indicating how early it was given, researches in [19] showed that it is possible to use these features in detecting spam reviews.

Maximum content similarity: This feature is based on the fact that spammers usually post the same reviews multiple times making only small changes. Using cosine similarity between reviews of the same author, it was shown that over 70% of spammers achieve score of 0.3 or higher, while 30% of non-spammers achieve cosine score over 0.18.

Researchers in [18] used behavioral features on Yelp dataset. They used before described *Maximum number of reviews (MNR), Positive review percentage (PR), Review length (RL), Reviewer deviation (RD)* and *Maximum content similarity (MCS)*. They also used *n-grams* as proposed in [14]. Classification was done using SVM and 5-fold cross-validation. Using bigrams on hotel reviews they achieved accuracy of 64.4%. Behavior features (BF) yielded 83.2%, while combination of bigrams and BF got 84.8%. These results show that on Yelp dataset methods using behavior features achieve much better results than content-based methods. Authors also tested the effect of excluding one feature from proposed method.

Excluded feature	Accuracy
MNR	83.1%
PR	80.1%
RL	79.7%
RD	84%
MCS	82.9%

Table 2: Accuracy of proposed method excluding one feature

This shows that omitting one behavioral feature does not significantly affect accuracy results.

In [19] authors used behavioral features *Early rating deviation, Reviewer deviation* as well as *Targeting products* and *Targeting groups. Targeting products* feature is based on checking similarities between different reviews of the same reviewer. *Targeting groups* feature is based on the fact that sometimes spammers write multiple reviews for products of one manufacturer during a short period of time. Dataset was downloaded from Amazon website and logistic regression model was used for training. Methods were evaluated using human judgement, which was shown to be unreliable by many papers. Nevertheless, this paper identified some important characteristics of spammers' behavior which can be further investigated.

Behavioral methods, although not as well researched as content-based techniques, were shown to be a powerful tool in spam detection. Beside the ones described in this paper, there are many other features collected by websites that could be used to improve accuracy of spam detection systems. Some of them are IP and MAC addresses of a host computer, its geo-location, click behaviors etc. However, these features are intended for internal use and are rarely available to researchers, so their contribution to spam detection algorithms is yet unknown.

IV. Experimental results

In practice, most websites use proprietary algorithms for spam detection. Decision making process is usually done with combination of spam filtering software and human analysis. As software for detection of untruthful reviews is not publicly available, results obtained by using a popular program for detecting bot spam are presented.

Akismet [22] is the most popular plugin for spam detection in blog comments. Its reported stats as of today are over 400 billion removed spam messages and accuracy over 99%. Although its algorithm is not publicly available, some information about its decision-making process is known. When comment is posted, Akismet compares its content to known spam messages in its database. If match is found, that comment is marked as spam.

Setup for this experiment included creating five WordPress websites on a local machine and installing Akismet on each of them. To post a comment, unlogged user is required to enter his name and email address. Six types of comments are defined based on their content:

1. Comments without alphanumerical characters
2. Random combination of characters
3. Random combination of words
4. Common phrases
5. Links
6. Text with links

Conducted tests show that comment is marked as spam if it contains random combination of non-alphanumerical characters (dashes, dots, commas…).

Random combination of characters or words that were tried were not marked as spam immediately. However, if the same comment was posted more than five times in a short time interval it was marked as spam in all next cases.

Common phrases and links were posted tens of times consecutively without being marked as spam. Common phrases are defined as short sentences or combination of words that are likely to be found on web.

Comments containing text with links were marked as spam if posted more than five times in a relatively short amount of time. For example, comments "http://www.example.com" and "Example test" were not marked as spam when posted tens of times. However, comment "Example test: http://www.example.com" was marked as spam when posted over five times consecutively.

Experiment included posting these comments multiple times on the same website as well as on different websites. Comments on different websites were posted using the same email address, as well as different email addresses. Further, cases when times between posting comments were five seconds, one minute and ten minutes were tested. It was shown that in all these cases the same results were obtained. After a comment is marked as spam, using slight variations of it will also be blocked by Akismet. If one comment of a user was marked as spam manually, all further comments posted on the same website by that user were marked as spam automatically. However, other users were still able to post comments with that same content. This shows that manual labeling spam comments is used for detecting spam users, rather than determining if comment's content represents spam.

V. Conclusion

In this paper a brief overview of spam detection methods published during the last decade was presented. It was shown that using different datasets yields extremely different results. Moreover, the lack of a proper gold standard dataset was recognized as a major problem in spam detection. Although linguistic approaches dominate in number of research papers, spammer detection techniques have shown promising results. Therefore, future research should be focused on combining content-based and reviewer-based methods for achieving the best results.

Literature

[1] http://www.competitionbureau.gc.ca/eic/site/cb-bc.nsf/eng/03782.html

[2] Nielsen blog, July 7, 2009. Global advertising: Consumers trust real friends and virtual strangers the most. http://blog.nielsen.com/nielsenwire/consumer/global-advertising-consumers-trust-real-friends-andvirtual-strangers-the-most

[3] A. Heydari, Mhd. A. Tavakoli, N. Salim, and Z. Heydari, Detection of review spam: A survey, Expert Systems with Applications 42, no. 7 (2015): 3634-3642

[4] N. Jindal and B. Liu. 2008. Opinion spam and analysis. In Proceedings of the international conference on Web search and web data mining, pages 219–230. ACM

[5] B. Liu. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, pages 1–167, 2012.

[6] M. Crawford, T.M. Khoshgoftaar, J.D. Prusa, A.N. Richter, H. Al Najada, "Survey of review spam detection using machine learning techniques", Journal Of Big Data, 2, pp. 1-24, 2015.

[7] Mukherjee A, Venkataraman V, Liu B, Glance NS (2013) What yelp fake review filter might be doing? Boston, In ICWSM.

[8] N Jindal, B Liu, Proceedings of the 16th international conference on World Wide Web, 1189-1190

[9] N. Jindal and B. Liu. Analyzing and Detecting Review Spam. ICDM2007.

[10] A. Z. Broder. On the resemblance and containment documents. In *Proceedings of Compression and Complexity of Sequences 1997*, IEEE Computer Society, 1997

[11] Algur et al. (2010). Conceptual level similarity measure based review spam detection. In International conference on signal and image processing (pp. 8)

[12] Lai et al. (2010). Toward a language modeling approach for consumer review spam detection. In IEEE 7th international conference (pp. 8)

[13] Pennebaker et al. (2007). The development and psychometric properties of LIWC2007. Austin, TX, LIWC.Net.

[14] Ott et al. (2011). Finding deceptive opinion spam by any stretch of the imagination. In 49th annual meeting of the association for the computational linguistics (pp. 11).

[15] M. Ott, C. Cardie, & J. T. Hancock, Negative deceptive opinion spam, Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies, June 2013 (pp. 497–501).

[16] Mukherjee et al. (2013). Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews, UIC-CS-03-2013. Technical Report.

[17] Sun et al. (2013). Synthetic review spamming and defense. In Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. ACM.

[18] Mukherjee et al. (2013). What yelp fake review filter might be doing. In Seventh international AAAI conference on weblogs and social media.

[19] Lim et al. (2010). Detecting product review spammers using rating behaviors. In Proceedings of the 19th ACM international conference on Information and knowledge management. ACM

[20] http://myleott.com/op-spam.html

[21] https://www.yelp.com/dataset

[22] https://akismet.com/