ISSN:0975-9646

Survey: Opinion Spam Detection Approaches and Techniques

P.N.V.S.Pavan Kumar[1], A.Suresh Babu [2], N.Kasiviswanath[3]

[1,3]*CSE Department, G.Pulla Reddy Engineering College(Autonomous)*
Kurnool, Andhra Pradesh,India

[2]*Jawaharlal Nehru Technological University*
Anantapuramu,Andhra Pradesh,India

Abstract— Online reviews play a significant role in today's e-commerce. Most of the customers now a days are depending on the reviews and ratings for taking decisions of what to buy and from where to buy. Thus ,Pervasive spam, fake and malicious reviews are affecting the decisions of customers while buying products. These reviews also affects stores rating and impression. Without proper protection, spam reviews will cause gradual loss of credibility of the reviews and corrupt the entire online review systems eventually. Therefore, review spam detection is considered as the first step towards securing the online review systems. We aim to give overview of existing detection approaches in a systematic way, define key research issues, and articulate future research challenges and opportunities for review spam detection. Opinion spam (or fake review) detection has attracted significant research attention in recent years ,the problem is far from solved. In this survey ,we present various methods of opinion spam detection.

Keywords— **Opinion Mining, Review spam, Machine Learning, Supervised Learning.**

I. INTRODUCTION

People exchange opinions about products or merchants in online blogs, forums, social media, or directly post reviews in various reputation systems provided by individual online retailers, mega-retailers (e.g., eBay, Amazon) or third-party sites (e.g., Bizrate, reseller rating.com, Google+ Local, Yelp, etc.). Recent surveys show that 83% of the consumers check out online reviews to know about the products or businesses they are buying from [10], and 80% of the consumers have changed purchase decision due to negative reviews [11]. People's attitudes and opinions are highly influenceable by others, which is known as the word-of-mouth effect in shaping decision making.

Opinion Mining or Sentiment analysis involves building a system to explore user's opinions made in blog posts, comments, reviews or tweets, about the product, policy or a topic. It aims to determine the attitude of a user about some topic. In recent years, the exponential increase in the Internet usage and exchange of user's opinion is the motivation for Opinion Mining. The Web is a huge repository of structured and unstructured data. The analysis of this data to extract underlying user's opinion and sentiment is a challenging task. An opinion can be described as a quadruple consisting of a Topic, Holder, Claim and Sentiment [5]. Here the Holder believes a Claim about the Topic and expresses it through an associated Sentiment. To a machine, opinion is a "quintuple", an object made up of 5 different things: [Bing Liu in NLP Handbook] (Oj, fjk, SOijkl, hi, tl), where Oj= the object on which the opinion is on, fjk = a feature of Oj, SOijkl = the sentiment value of the opinion, hi = Opinion holder, tl = the time at which the opinion is given.

There are several challenges in the field of sentiment analysis. The most common challenges are Word Sense Disambiguation (WSD), a classical NLP problem is often encountered. The opinion word unpredictable is used in different senses. Secondly, addressing the problem of sudden deviation from positive to negative polarity, as in "The movie has a great cast, superb storyline and spectacular photography; the director has managed to make a mess of the whole thing". Thirdly, negations, unless handled properly can completely mislead. "Not only do I not approve Supernova 7200, but also hesitate to call it a phone" has a positive polarity word approve; but its effect is negated by many negations. Fourthly, keeping the target in focus can be a challenge.

Different techniques are introduced and used for detecting review spam.[1] has given main 3 types of review spam, which are

a. Un-truthful review (False opinions) which is divided in two category. Positive Spam review(Undeserving opinion to promote product) Negative Spam Review(negative opinion to damage reputation
b. Reviews on brand only(reviews on some particular brands)
c. Non-reviews (contain no reviews) which is divided in Advertisements, Question or answers, Comment on other reviews or any Random text.

In this survey paper different techniques used to detect these type of spam are discussed. The rest of the article is organized as follows.

- Section II discuses various review spam detection approaches
- Section III gives a comparative analysis of Opinion Spam Detection(OSD) Techniques.
- Section IV concludes the survey paper.

II. OSD APPROACHES

Depending upon the approach used for spam detection it can be classified as:

A. Review centric approach
B. Reviewer centric approach
In this work main modules are

1. Customer Reviews

2. Review pre-processing
3. Stop-word-removal
4. Detecting Duplicate and near duplicate
5. Un-truth full spam review[1] Classification Technique

Here they used total 12 features extracted from reviews and give labels to each. For evaluation they have compared accuracy of four machine learning methods Gaussian, naïve bayes, Decision Tree, Multinomial naïve bayes, Logistic Regression.And they have shown that Logistic Regression and Gaussian have higher accuracy as compared to Decision tree and Multinomial naïve bayes.

C. Conceptual level Similarity Measure based Review Spam Detection [7]

Here the format of reviews they had used is pros and cons. According to them the review is not a spam in following two conditions

1. If the number of matched features is below some specified threshold i.e. partially related reviews
2. If the reviews are Unique Reviews, it has three steps:

1. Feature extraction-It involves feature extraction from reviews and storing them in feature database

2. Feature matrix construction-features extracted in step 1 are used to construct feature matrix.

3. Matching feature calculation between reviews-By calculating similarity score of

Different review pairs they are categorized as spam (duplicate/ near duplicate) or non-spam (partially related /unique) based on threshold value T. For evaluation purpose confusion matrix is created for pros and cons separately and compared human annotated result with automated result.

D. Toward A Language Modeling Approach for Consumer Review Spam Detection

This paper [8] is to show the trustworthiness of reviews by detecting the review spam. Their experimental result shows that the KL divergence and the probabilistic language model is effective for the detection of untruthful reviews.In their work they have used

- The pre processing techniques like POS (Part of speech Tagging), stop-word-removal, stemming on the data crawled from web.
- And they have developed their POS tagger based on the word-net lexicon and the publically available Word-Net API. And used the unsupervised probabilistic language model (for untruthful review detection which is type 1 review spam), and a supervised classifier (for non-review detection which is type 3 spam).
- For non-review spam detection they identify features which were used in detecting web spam [4] Which are Syntactical, Lexical and Stylistic features. For classification task they have used SVM (Support Vector Machine) and Logistic Regression.
- For un-truth full type of reviews they build the computational model using KL (Kullback-Leibler) divergence which is a well-known measure co

E. Text Mining and Probabilistic Language Modeling for Online review Spam Detection

[12]Has detected type 1 and type 3 spam reviews. Main focus on type 1 spam review.

In this study they have detected the fake reviews and the final decision was on the Visitors that the review is fake or not. In their work they have divided their work in following modules.

Mod 1: In this the user selects the detection scope.

Mod 2: If reviews are not available locally then use API (Application Programming Interface) to retrieve reviews.

Mod 3: traditional document preprocessing procedures, which are stop-word removal, Part-of-Speech (POS)tagging, and stemming were applied on data.

Mod 4: after the reviews were preprocessed, the high order concept association mining module was invoked to extract the prominent concepts and their high-order associations for each product category. These association relationships were used to bootstrap the performance

Mod 5: non review detection is performed by a supervised SVM classifier.

Mod 6: untruthful review here type-1 spam review detection is carried out by an unsupervised probabilistic language model.

For the non-review spam detection they have used SVM (Support Vector Machine) and LR to classify the reviews. For that they have used the features same as in web spam detection technique for SVM. And for un-truth full reviews they developed their model and used different techniques.

The results: they have used the methods for untruth full reviews are SVM, VS (Vector Space),I-match,

LM(unigram Language Model), SLM(Semantic Language Model). The result shows that SLM gives the highest result and SVM gives poor result. And for non-review based spam they used KNN (Nearest neighbor classifier), LR(Logistic Regression), and SVM (Support Vector Machine). Results shows that SVM is performing well and it has the highest result among them.

III. COMPARATIVE ANALYSIS OF OSD TECHNIQUES

When developing a new review spam detection framework, it is important to under-stand what approaches and techniques have been used in prior studies. Based on our survey, most of the previous studies have focused on supervised learning techniques. However, in order to use supervised learning, one must have a labeled dataset, which can be difficult (if not impossible) to acquire in the area of review spam. Despite the prevalence of opinion spam, existing methods are not keeping pace due to the unavailability of large-scale ground truth datasets in the real world commercial setting which impedes research of opinion spam detection. Existing work typically relies on pseudo fake reviews rather than real fake ones. For example, Jindal and Liu (2008) treated duplicate and near-duplicate Amazon product reviews as fake reviews. Li et al. (2011) manually labeled fake reviews by reading the reviews and comments, which are unreliable.Ott et al. (2011) used Amazon Mechanical Turk (AMT) to crowdsource

anonymous online workers to write fake hotel reviews. The review dataset that they compiled had only 800 reviews which is too small to support reliable statistical analysis. In addition to that, the motivations and the psychological states of mind of hired Turkers and the professional spammers in the real world can be quite different as the results shown in (Mukherjee et al. 2013).

Companies such as Dianping and Yelp have developed effective fake review filtering systems against opinion spam.Mukherjee et al. (2013) reported the first analysis of Yelp's filter based on reviews of a small number of hotels and restaurants in Chicago. Their work showed that behavioral features of reviewers and their reviews are strong indicators of spamming. However, the reviews they used were not provided by Yelp but crawled from Yelp's business pages. Due to the difficulty of crawling and Yelp's crawling rate limit,they only obtained a small set of (about 64,000) reviews.

Below figure Fig.1 shows the table for comparison of previous works and results for review spam detection .

Dataset	Features used	Learner	Performance metric	Score	Method complexity
5.8 million reviews written by 2.14 reviewers crawled from amazon website	Review and reviewer features	LR	AUC	78 %	Low
5.8 million reviews written by 2.14 reviewers crawled from amazon website	Features of the review, reviewer and product characteristics	LR	AUC	78 %	Medium
5.8 million reviews written by 2.14 reviewers crawled from amazon website	Text features	LR	AUC	63 %	Low
6000 reviews from Epinions	Review and reviewer features	NB with Co-training	F-Score	0631	High
Hotels through Amazon Mechanical Turk (AMT) by Ott et al.	Bigrams	SVM	Accuracy	89.6 %	Low
Hotels through Amazon Mechanical Turk (AMT) by Ott et al.	LIWC + Bigrams	SVM	Accuracy	89.8 %	Medium
Hotels through Amazon Mechanical Turk (AMT) by Ott et al. + gathered 400 deceptive hotel and doctor reviews from domain experts	LIWC + POS + Unigram	SAGE	Accuracy	65 %	High
Yelp's real-life data	Behavioral features combined with the bigram features	SVM	Accuracy	86.1 %	Medium
Hotels through Amazon Mechanical Turk (AMT) by Ott et al.	Stylometric features	SVM	F-measure	84 %	Low
Hotels through Amazon Mechanical Turk (AMT) by Ott et al.	n-gram features	SVM	Accuracy	86 %	Low
Dataset collected from amazon.com	Syntactical, lexical, and stylistic features	SLM	AUC	.9986	High
Their own crawled Arabic reviews from tripadvisor.com, booking.com, and agoda.ae	Review and reviewer features	NB	F-measure	.9959	Low

Fig. 1: comparison of previous works and results for review spam detection

IV. CONCLUSIONS

As review text is an important source of information and tens of thousands of text features can easily be generated based on this text, high dimensionality can be an issue. Additionally, millions of reviews are available to be used to train classifiers, and training classifiers from a large, highly dimensional dataset is computationally expensive and potentially impractical. Despite this, feature selection techniques have received little attention. Many experiments have avoided this issue by extracting only a small number of features, avoiding the use of n-grams, or by limiting number of features through alternative means such as using term frequencies to determine what n-grams are included as features. Further work needs to be conducted to establish how many features are required and what types of features are the most beneficial. Feature selection should not be considered optional when training a classifier in a big data domain with potential for high feature dimensionality. Additionally, we could find no studies that incorporated distributed or streaming implementations for learning from Big Data into their spam detection frameworks.

REFERENCES

[1] Nitin Jindal and Bing Liu. "Opinion Spam and Analysis." Proceedings of First ACM International Conference on Web Search and Data Mining (WSDM-2008), Feb 11-12, 2008, Stanford University, California, USA.

[2] Lau RY, Liao SY, Kwok RCW, Xu K, Xia Y, Li Y (2011) Text mining and probabilistic language modeling for online review spam detecting. ACM Trans Manage Inf Syst 2(4):1–30

[3] Ott M, Choi Y, Cardie C, Hancock JT (2011) Finding deceptive opinion spam by any stretch of the imagination. In:Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics.

[4] Li F, Huang M, Yang Y, Zhu X (2011) Learning to identify review spam. In: IJCAI Proceedings-International Joint Conference on Artificial Intelligence, vol 22, No. 3., p 2488.

[5] Shojaee S, Murad MAA, Bin Azman A, Sharef NM, Nadali S (2013) Detecting deceptive reviews using lexical and syntactic features. In: Intelligent Systems Design and Applications (ISDA), 2013 13th International Conference on (pp. 53–58). IEEE,Serdang, Malaysia

[6] Soo-Min Kim and Eduard Hovy, "Determining the Sentiment of Opinions", *Proceedings of the Conference on Computational Linguistic*, Article No. 1367, 2004.

[7] Qian T, Liu B (2013) Identifying Multiple User ids of the Same Author. In: EMNLP., pp 1124–1135.

[8] Siddu P.Algur, Amit P.Patil, P.S Hiremath S.Shivashankar Conceptual level Similarity Measure based Review Spam Detection 2010 IEEE

[9] C.L. Lai, K.Q. Xu, Raymond Y.K. Lau, Y. li, L. Jing Toward A Language Modeling Approach for Consumer Review Spam Detection
International Conference on E-Business Engineering 2010.

[10] "The company behind the brand: In reputation we trust:' Weber Shandwick's online survey. 2012.

[11] "2011 Cone online influence trend tracker."http://www.coneinc.com/20 11coneonlineinfluencetrendtracker. 2011.

[12] J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.

[13] Raymond Y. K. Lau, S. Y. Liao, Ron Chiwai Kwok, Kaiquan Xu, Yunqing Xia, Yuefeng Li "Text Mining and Probabilistic Language Modeling for Online Review Spam Detection" ACM Trans. Manag. Inform. Syst. 2, 4,Article 25 (December 2011)

[14] Jindal N, Liu B (2007) Review spam detection. In: Proceedings of the 16th international conference on World Wide Web (pp. 1189–1190). ACM, Lyon, France

[15] Jindal N, Liu B (2008) Opinion spam and analysis. In: Proceedings of the 2008 International Conference on Web Search and Data Mining (pp. 219–230). ACM, Stanford, CA.

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

[17] Hammad ASA (2013) An Approach for Detecting Spam in Arabic Opinion Reviews. Doctoral dissertation, Islamic University of Gaza.

[18] Li J, Ott M, Cardie C, Hovy E (2014) Towards a general rule for identifying deceptive opinion spam. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1566–1576, Baltimore, Maryland,USA, June 23-25 2014. ACL.