A Survey on Fake Review Detection using Machine Learning Techniques

Nidhi A. Patel
Computer Engineering Department
Sarvajanik Collage of Engineering and Technology
Surat, India
nidhi.patel00051@gmail.com

Prof. Rakesh Patel
Computer Engineering Department
Sarvajanik Collage of Engineering and Technology
Surat, India
rakeshpatel.ce@gmail.com

***Abstract* — Now a days the usage of Internet and online marketing has become very popular. Millions of products and services are available in online marketing that generate huge amount of information. Hence, it's difficult to find the best suitable services or products compatible to the requirement. Customers directly take decision based on reviews or opinions that are written by others based on their experiences. In this competitive world any person can write anything, this raise the number of fake reviews. Various companies are hiring people to write fake positive reviews about their services or products or unfair negative reviews about their competitors' services or products. This process gives wrong input to the new customers who wish to buy such items and hence we need a system to detect such fake reviews and remove them. In this paper we discuss various supervised, unsupervised and semi supervised data mining techniques for fake review detection based on different features.**

Keywords* — *Fake Review, Sentiment Analysis, Opinion Spam, Fake review detection technique, Machine learning.

I. INTRODUCTION

In recent years, the World Wide Web has drastically changed the way of sharing the opinions. Online reviews are comments, tweets, posts, opinions on different online platforms like review sites, news sites, e-commerce sites or any other social networking sites. Sharing reviews is one of the ways to write a review about services or products [1] [2]. Reviews are considered as an individual's personal thought or experience about products or services [7] [13]. Customer analyzes available reviews and takes decision whether to purchase the product or not [3]. Therefore online reviews are valuable source of information about customer opinions [5]. Fake or spam review refers to any unsolicited and irrelevant information about the product or service. Spammer writes fake reviews about the competitors' product and promotes own products [8] [10]. The reviews written by spammers are known as fake reviews or spam reviews [2]. Thus fake reviews detection has become critical issue for customers to make better decision on products trustworthy as well as the vendors to make their purchase [15].

The fake reviews are classified in two groups [11] [17].

- Untruthful reviews- These reviews promote or demote the products with positive or negative words respectively and misguide the customers.
- Reviews on brands- These reviews are not related to products, not on the different features of the product or services. Reviewer uses brand name repeatedly to promote a particular brand.

Jindal and Liu [11] [12] proposed three basic techniques for identifying fake reviews. These three approaches are as follows.

- Review Centric Approach- This approach identifies review as fake review based on the content of reviews written by reviewers. In this method, various features like review content similarity, use of capitals, all capital words, use of numerals, brand name, similarity between products and reviews, repeated use of good and bad words in review.
- Reviewer Centric Approach- This method depends on the behavior of reviewers. This approach considers information about users and all reviews that are written by them [1]. Features used in this method are account age, profile picture, URL length, IP address, number of written reviews by one reviewer, maximum rating per day etc [18] [27] [28].
- Product Centric Approach- This method mainly focuses on the product related information. In this method, sales rank of product, price of product etc are considered as features.

Initially fake review detection was introduced by Jinal et al. [12]. There are various ways to identify fake reviews. Machine learning technique is one of the ways to identify fake reviews [17]. Machine learning model learns and make prediction [2]. The basic steps involved in machine learning are data processing, feature extraction, feature selection, classification model generation. This process is shown in Fig. 1:

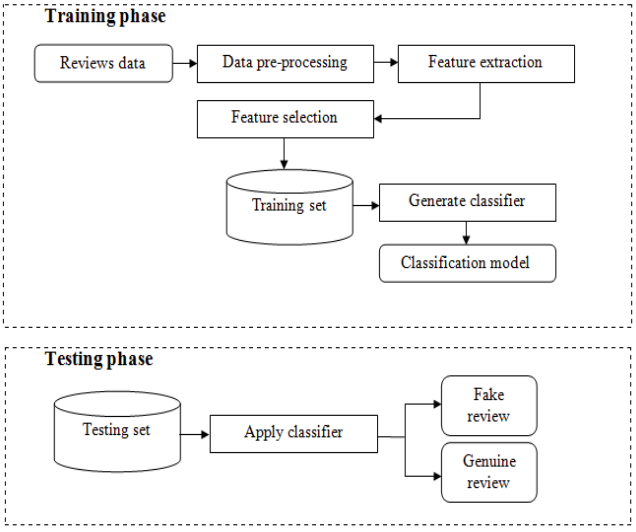

Fig. 1. Machine Learning based Fake Review Detection

Machine learning approach for fake review detection works as follows:

- Data collection: In this phase, review data will be gathered from various platforms like Amazon. These reviews could be for product or service like hotel reviews.

- Data pre-processing: In next step, data pre-processing is applied like punctuation marks removal, stemming, stop word removal etc. In punctuation marks removal, the whole text is divided into sentences, phrases or paragraphs. In the stemming process, stem will be created from every word in dataset. In stop word removal phase, frequently used group of words like determiners, articles and preposition will be detected and removed. After removing these words, only important words will be retained for the next step [6].

- Feature extraction and selection: In this step, features are extracted from the preprocessed data. The different types of features which are used to detect fake reviews are classified as linguistic features, relational features, and behavioral features. The classification is shown in Fig 2.

- Classifier model construction and testing: For training purpose, small set of labeled data is used. In this phase, classification model is generated by using the training review dataset. The reviews used for this purpose are already labeled as fake or genuine review. Once the classifier is trained, it will be tested using test dataset. The different machine learning algorithms which can be used for model construction are naive bayes classification, decision tree algorithm, support vector machine, k-nearest neighbor, logistic regression, etc.

The performance of fake review detection method depends on labeled data used for training purpose, correct selection of features and data mining techniques used for detection.

The rest of the paper is organized as follows: Section II summarizes the related fake review detection work. Section III discusses the machine learning based fake review detection techniques. In section IV, fake review detection based on important attributes such as features, classifiers are discussed. Section V discusses major challenges in fake review detection. The section VI concludes the paper.

II. Related Work

For fake review detection, there are numbers of machine learning algorithms. Using Machine learning techniques, fake reviews detection depends on behavioral features, linguistic and textual features and relational features. This is shown in Fig. 2.

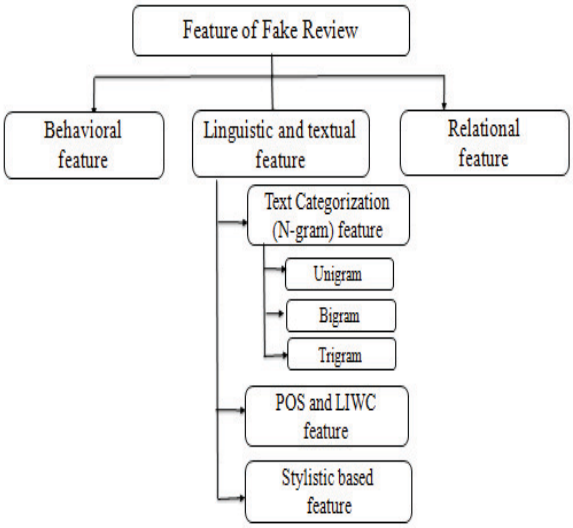

Fig. 2. Types of Fake Review Features

When spammer writes fake reviews, spammer reflects their thought, feeling and emotions. In behavioral feature, review spammers behave differently than genuine user. They may write large no of reviews in short time [2] [25], may use terms of extreme rating (very low or high) [20]. In this case, spammers write multiple fake reviews from different accounts rather than different time interval.

In Relational feature, graphical structure represents relationship among reviews, reviewers and products. The bi-partite graphical model represents the relationship between reviewers and products [21]. Tri-partite graphical model represents the relationship among reviews, reviewers and IP addresses of review spammers [22]. In network specific, different features are considered like number of product targeted by spam group, reviewer correlation in spam group, size of spam group and product reviewer ratio in spam group [23].

The linguistic feature is one of main features to detect fake reviews that depend on writing styles and languages. Linguistic and textual features include N-gram feature, POS feature, LIWC features and stylistic feature [2] [9]. N-gram feature contains unigram, bigram and trigram. In POS tag, each word of review, POS tagger use syntactic deception clues about review spamming. Most of the spammer writes imaginative reviews using pronouns or adverbs, verbs, while normal users write informative reviews using more adjective or noun. LIWC (Linguistic Inquiry and Word Count) is also used to identify the fake reviews. LIWC feature likes score of affective positive and negative feelings, score of punctuation marks [19]. The stylistic based feature depends on word similarity measure (for example, cosine similarity) semantic similarity between objects and review (like product, news articles etc.) [24] [26]. The stylistic based feature also includes percentage of repeated words, percentage of personal pronouns, percentage of emotional words, percentage of capitalized words, frequency of passive voices etc.

III. Machine Learning Based Fake Review Detection Techniques

Various techniques have been proposed in past to identify fake reviews based on types of data like labeled data (for example, supervised learning), unlabeled data (for example, unsupervised learning), and partially labeled data (for example, semi-supervised learning) that is described below.

A. Supervised Learning techniques

Wael et al. [6] use supervised learning algorithm for fake review detection. Before applying the classification method, different preprocessing steps are performed; these steps include stemming, removal of punctuation marks and stop word removal. They use linguistic feature to identify fake reviews. Linguistic feature contains POS and bag-of-words. Bag-of-words features consist of individual word or group of words that are found in given text. Then different classification algorithms are applied like decision tree, random forest, support vector machine, naive bayes and gradient boosted trees. Here naive bayes and support vector machine give better result.

Jitendra et al. [2] applied different features based on content similarity and sentiment polarity for identifying fake and genuine reviews. Here authors use sentiment score based on sentiment polarity between positive and negative reviews, linguistic and unigram as feature. They then applied three algorithms 1) support vector machine, 2) naive bayes and 3) decision tree.

Snehasish et al. [3] uses supervised machine learning algorithm. In this method, fake reviews are differentiated from genuine reviews using four linguistic clues like level of detail, understandability, cognition indicators and writing style. Level of detail contains different operationalized features like informativeness, contextual detail, lexical diversity, function words and perceptual detail. Informativeness was calculated by POS (part of speech) like noun, verb, adjective, verbs, adverbs, pronouns etc. Contextual detail contains spatial and temporal references while perceptual detail contains feeling and visual words, proportion of aural. Lexical diversity words contain non content words that overcome the level of detail in the reviews. Writing style depends on use of upper case, lower case, question marks, all punctuation, tenses, and emotions. Tenses was measured based on group of future, past and present tense words. Cognition indicator is based on tentative word, motion words, exclusion words as well as causal words etc. The authors use different supervised learning algorithms like logistic regression, C4.5, back propagation network, naive bayes, support vector machine using polynomial kernel, support vector machine using linear kernel, support vector machine with radical basis faction kernel, voting, k-nearest neighbor and random forest.

The fake and authenticate reviews are compared against two baselines. Baseline 1 contain different feature like character per word, length of review in words, first person singular words, lexical diversity, brand references, first person plural words, negative emotion word and positive emotion words. Second baseline contains verbs, adverbs, adjective, words per sentence, character per word, modal verbs, all punctuation, first person plural words, first person singular words, spatial words, function words, temporal words, emotiveness, visual words, feeling words, aural words, negative emotion words and positive emotion words. Second baseline gives more accurate result compared to first baseline.

B. Semi-Supervised Learning techniques

Hernandez et al. [4] first introduced PU-learning technique to detect fake review. Positive Unlabeled (PU) learning technique is combination of some positive label and unlabelled dataset. PU-learning technique is semi-supervised technique, which only uses two class classifiers positive as deceptive and unlabeled without having negative as truthful training example. In this algorithm, first unlabeled data are considered as negative class. In next step, classifiers are trained based on initial set of positive instances. Then classifiers are applied only on unlabeled instances and generate labeled instances. After, classified positive and negative instances, the positive instances as deceptive reviews are eliminated from unlabeled instances and rest of them are considered as negative instances. Again classifiers are applied into negative instances. This process is repeated until the stop criteria, which classify fake and genuine reviews. Here two classifiers are applied in PU-learning, support vector machine and naive bayes.

Rohit et al. [5], use PU-learning algorithm using various classifiers. In this, author uses six different classifiers to detect fake reviews. These are decision tree, naive bayes, random forest, support vector machine, logistic regression and k-nearest neighbor classifiers. Here logistic regression classifier gives best performance compared to all six different algorithms.

Hernandez et al. [16] compared tradition PU-learning with modified PU-learning algorithm. Using modified PU-learning technique, author analyzed that it is possible to detect less number of instances from unlabeled set. In every iteration, only new negative instances are considered that are generated by output of previous iteration and classifier is only applied on that new negative instances. Thus in every iteration, negative instances are reduced and final instances are correctly identified as fake or genuine reviews. In this paper, authors also detect positive and negative fake reviews. They used naive bayes and support vector machine classifier with both unigrams and bigrams features and reviews are classified into fake and non-fake reviews.

C. *Unsupervised Learning techniques*

Main advantage of unsupervised learning approach is that, without any labeled dataset, we can classify fake and genuine reviews.

Jitendra et al. [2] uses unsupervised learning approach. Author uses different features based on review data, reviewer data and product information based on difference in behavioral pattern of reviews. Here author uses Amazon cell phone reviews dataset to identity fake and genuine reviews.

TABLE 1. PARAMETRIC EVOLUTION OF DIFFERENT FAKE REVIEW DETECTION TECHNIQUES

Title/ Author/ Publication	Approach	Feature(s)	Classifiers	Dataset	Limitation/ Future scope
Using Supervised Learning to Classify Authentic and Fake Online Reviews [3] **Authors:** S. Banerjee, A. Chua, J. Kim **Publication:** 2015 ACM	Supervised	Linguistic clues • Writing style • Level of details • Structure of Word • Cognition Indicators	Supervised machine learning algorithm like Random forest, Support vector machine, Naive bayes, etc.	15 Asia Hotel Reviews	- Only used for labeled data. - Not suitable for Unlabeled dataset.
Deceptive Review Detection Using Labeled and Unlabeled Data [2] **Authors:** J. Rout, S. Singh, S. Jena, and S. Bakshi **Publication:** 2016 Springer	Supervised Unsupervised	Sentiment Score, Linguistic feature and Unigram Reviewer data, Review Data and Product information	Support vector machine, Naive bayes, Decision tree	20 Chicago Hotel Review Amazon cell phone and electronic product dataset	- Number of features are limited
The impact of applying different pre processing steps on review spam detection [6] **Authors:** E. Wtaiwi, G. Naymat **Publication:** 2017 Elsevier	Supervised	Linguistic clues N-gram	Decision tree Random forest Naïve bayes Support vector machine	Chicago Hotel Review	-
Using PU-Learning to Detect Deceptive Opinion Spam [4] **Authors:** P.Rosso, D.Cabrera, M. Gomez **Publication:** ACL 2013	Semi-supervised	-	Support vector machine Naive bayes	Chicago Hotel Reviews	- Limited classifier - Required positive instances (labeled data) Future scope : Adding unlabeled instances into positive instances
Review Spam Detection Using Semi-supervised Technique [5] **Authors:** R.Narayan,J. Rout and S. Jena **Publication:** 2018 Springer	Semi-supervised	-	Decision tree, Naive bayes, Random forest, Support vector machine, Logistic regression, K-nearest neighbor	Chicago Hotel Reviews	- Required positive instance (labeled data) Future scope : Same work extended for unsupervised learning technique.
Detecting positive and negative deceptive opinions using PU-learning [16] **Authors:** P.Rosso, D.Cabrera, M. Gomez **Publication:** 2015 Elsevier	Semi-supervised	Unigram and Bigram	Support vector machine Naive bayes	Hotel review	- Required positive instance (labeled data)

IV. ANALYSIS OF EXISTING FAKE REVIEW DETECTION TECHNIQUES

From past research work, fake reviews can be detected by different techniques like classification, clustering, or combined both of them. For correctly identifying fake reviews, different techniques are used based on features and classifiers. These features and classifier are illustrated below.

TABLE 1 shows summary of different strategies that are used to identify fake and genuine reviews.

A. Approach:

To identify fake reviews based on types of data like unlabeled data (for example unsupervised learning), labeled data (for example supervised learning) and partially labeled data (for example semi-supervised learning).

B. Feature:

The different features are used to identify fake and genuine reviews like linguistic feature, sentiment score, relational feature etc.

C. Classifier:

Using different classification algorithms like logistic regression, k-nearest neighbor, random forest, naive bayes and support vector machine, the reviews are classified as fake and genuine reviews.

D. Dataset:

Authors either use publically available dataset or create own dataset.

V. Major Challenges In Fake Review Detection Techniques

The challenges which are involved in different fake review detection are as below.

1. Review features like ratings, brand names reference are hard for human, machines not to mention [2].
2. When only one review is available for a particular item, it is difficult to identify rating behaviours [1].
3. When fake reviews are intentionally fabricated like genuine review, it would be hard to decide genuine review.

VI. Conclusion

Due to rapid development of the internet, the size of the reviews of the items / products increases. These huge amounts of information are generated on Internet; there is no analysis of quality of reviews that are written by consumer. Anyone can write anything which conclusively leads to fake reviews or some companies are hiring people to post reviews. Some of the fake reviews that have been intentionally fabricated to seem genuine, capability to identify fake online reviews are crucial. In this paper, we have discussed different fake reviews detection techniques that are based on unsupervised, supervised as well as semi supervised methodologies. In this paper, we have seen different features in detail like linguistic features, behavioral and relational features. We have also compared different techniques to identify fake reviews. We have also discussed major challenges of fake review detection.

References

[1] A. Rastogi, M. Mehrotra, "Opinion spam Detection in Online Reviews", *Journal of information and Knowledge Management,* vol. 16, no. 04, pp. 1-38, 2017.

[2] J. Rout,S. Singh, S. Jena, and S. Bakshi, "Deceptive review detection using labeled and unlabeled data", *Multimedia Tools and Applications*,vol.76, no. 3, pp. 3187-3211, 2016.

[3] S. Banerjee, A. Chua, J. Kim, "Using Supervised Learning to Classify Authentic and Fake Online Reviews ", *Proceeding of the 9th International Conference on Ubiquitous Information Management and Communication*",ACM, 2015.

[4] P.Rosso, D.Cabrera, M. Gomez, "Using PU-Learning to Detect Deceptive Opinion Spam", pp.38-45, 2013.

[5] R.Narayan,J. Rout and S. Jena, "Review Spam Detection Using Semi-supervised Technique", *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, pp. 281-286, 2018.

[6] W. Etaiwi,G. Naymat, "The impact of applying preprocessing steps on review spam detection", *The 8th international conference on emerging ubiquitous system and pervasion networks,* Elsevier, pp. 273-279, 2017.

[7] W. Zhang,R. Y. K. Lau and Li. Chunping, "Adaptive Big Data Analytics for Deceptive Review Detection in Online Social Media", *Thirty Fifth International Conference on Information Systems,* Auckland 2014,pp.1-19,2014.

[8] C. Lai, K. Xu, R. Y. Lau, Y. Li, and L. Jing, "Toward a Language Modeling Approach for Consumer Review Spam Detection," *2010 IEEE 7th International Conference on E-Business Engineering*, pp. 1–8, 2010.

[9] M. I. Ahsan, T. Nahian, A. A. Kafi, M. I. Hossain, and F. M. Shah, "Review spam detection using active learning," *2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2016.

[10] M. Ott, Y. Choi, C. Cardie and J.T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination", ACM, pp.309-319,2011.

[11] N. Jindal and B. Liu., "Opinion spam and analysis", *Proceedings of the international conference on Web search and web data mining - WSDM 08* (2008), ACM, pp. 219 –230,2008.

[12] N. Jindal and B. Liu, "Review spam detection", *Proceedings of the 16th international conference on World Wide Web - WWW 07* (2007), ACM, pp. 1189–1190, 2007

[13] S. Shojaee, A. Azman, M. Murad, N. Sharef and N. Sulaiman, "A Framework for Fake Review Annotation", *2015 17th UKSIM-AMSS International Conference on Modelling and Simulation,* IEEE, pp. 153-158,2015.

[14] J. Koven, H. Siadati, and C. Y. Lin, "Finding Valuable Yelp Comments by Personality, Content, Geo, and Anomaly Analysis," *2014 IEEE International Conference on Data Mining Workshop*, pp. 1215–1218, 2014.

[15] S. Banerjee and A.Y.K. Chua. 2014. "Applauses in hotel reviews: Genuine or deceptive ?", *2014 Science and Information Conference* (2014), pp. 938–942,2014.

[16] P.Rosso, D.Cabrera, M. Gomez, "Detecting positive and negative deceptive opinions using PU-learning", Elsevier, pp.1-11, 2014.

[17] R. Dewang, A.Singh," Identify of Fake review using new set of lexical and syntactic feature", *Proceedings of the Sixth International Conference on Computer and Communication Technology 2015* , ACM, pp. 115-119 ,2015.

[18] J. Fontanarava, G. Pasi, and M. Viviani, "Feature Analysis for Fake Review Detection through Supervised Classification," *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 658–666, 2017.

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

[20] S. Feng, L. Xing, A Gogar and Y. Choi,"Distributional footprints of deceptive product reviews". *In Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 98-105, 2012.

[21] Akoglu, L, R Chandy and C Faloutsos , "Opinion fraud detection in online reviews by network effects". *In Proceedings of the 7th AAAI International Conference on Weblogs and Social Media (ICWSM'13),* pp. 2-11, 2013.

[22] Li, H, Z Chen, B Liu, X Wei and J Shao, "Spotting fake reviews via collective positive-unlabeled learning". *In Proceedings of the 2014 IEEE International Conference on Data Mining (ICDM),* pp. 899–904. New York: IEEE, 2014.

[23] Wang, Z, T Hou, D Song, Z Li and T Kong, "Detecting review spammer groups via bipartite graph projection*", The Computer Journal* , 59(6), pp. 861–874, 2015.

[24] Wang, JZ, Z Yan, LT Yang and BX Huang, "An approach to rank reviews by fusing and mining opinions based on review pertinence", pp.3–15, 2015.

[25] Y. Li, X. Feng, and S. Zhang, "Detecting Fake Reviews Utilizing Semantic and Emotion Model," *2016 3rd International Conference on Information Science and Control Engineering (ICISCE)*, pp. 317–320, 2016.

[26] S. L. Christopher and H. A. Rahulnath, "Review authenticity verification using supervised learning and reviewer personality traits,"

2016 International Conference on Emerging Technological Trends (ICETT), 2016.

[27] P. Liu, Z. Xu, J. Ai, and F. Wang, "Identifying Indicators of Fake Reviews Based on Spammers Behavior Features," *2017 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, pp. 396–403, 2017.

[28] M. Singh, L. Kumar, and S. Sinha, "Model for Detecting Fake or Spam Reviews," *Advances in Intelligent Systems and Computing ICT Based Innovations*, pp. 213–217, Jan. 2017.