Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019)
IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4

Detection of fake opinions on online products using Decision Tree and Information Gain

Sanjay K.S
Research scholar,Dept of Computer Applications
The Bangalore Social and Educational Institute
Management Studies,Bangalore,Karnataka,India
e-mail:sanjukumsi@gmail.com

Dr.Ajit Danti
Professor, Faculty of Engineering-CSE,
Christ(Deemed to be University),Bangalore,Karnataka,India
e-mail: ajit.danti@christuniversity.in

***Abstract*— Online reviews are one of the major factors for the customers to purchase any product or to get service from many sources of information that can be used to determine the public opinion on the products. Fake reviews will be published intentionally to drive the web traffic towards the particular products. These fake reviewers mislead the customers to distract the purchasers mind. Reviewers behaviors are extracted based the semantical analysis of his review content for the purpose of identifying the review as fake or not. In this work the reviews are extracted from the web for a particular product, along with the reviews of several other information related to the reviewers also been extracted to identify the fake reviewers using decision tree classifier and Information Gain .Significance of the features on the decision is validated using information gain. Experiments are conducted on exhaustive set of reviews extracted from the web and demonstrated the efficacy of the proposed approach.**

Key words: Fake, Stars, response, reply, reviews, opinion, Entropy, Information Gain, Decision Tree.

I. INTRODUCTION

Now a day's people have started to give fake reviews on the products. The reviews on a product may be positive or negative, the negative reviews will attract the customers more than a positive review. These fake reviews can affect any business which leads financial profit or loses. Generally reviews will appear in e-commerce websites like jabong ,Flipkart, Amazon etc.. Due to the financial reasons many fake reviews will be appeared in these websites. The company owners will intentionally motivate some of the people to write the fake reviews to improve their business towards another product. Day by day every one focus the online reviews before planning for hotel, to buy a product, or to stay in any home stay.

Now a days internet has become everything and it is too fast. People are interacting with the social network across the world because they can share all the information and their thoughts. Through the internet people can do the online shopping, human tendency to do survey on products which they want to purchase, lots of reviews on a particular products will be available on their company's website like flipkart , amazon . The online reviews can change decision of the customer and they can finalize a product by comparing with the different brands of products, the customer can select the product and satisfy their requirement only if the reviews are not fake. On the other hand if the reviews are fake then it misleads the customer.

II. LITERATURE SURVEY

There are many research work published in the area of fake review identification. 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 (Michael Crawford et. al, 2015). A framework has been proposed to detect fake product reviews or spam reviews by using Opinion Mining (Anusha Sinha et. al, 2018). An end-to-end trainable unified model has been proposed to leverage the appealing properties from Autoencoder and random forest, a stochastic decision tree model is implemented to guide the global parameter learning process (Manqing Dong et.al, 2018). 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 (Salma Farooq and Hilal Ahmad Khanday, 2016). To detect spam and fake reviews, and filter out reviews with expletives, vulgar and curse words, by incorporating sentiment analysis has proposed (Shashank Kumar, 2017). To classify movie reviews into groups of positive or negative polarity by using machine learning algorithms has proposed (Elshrif Elmurngi 2018). A method to recognizing the untruthful reviews that are given by the users which is having distinct semantic content based on sentiment analysis as the reviews of movies has proposed (Rashmi Gomatesh Adike and Vivekanand Reddy , 2016). A system used to classify tweets into different groups as spam and non spam tweets has proposed (Abha Tewari and Smita Jangale ,2016). To detect fake reviews for a product by using the text and rating property from a

review has proposed (Eka Dyar Wahyuni and Arif Djunaidy, 2016). An empirical study of efficacy of classifying product review by semantic meaning has proposed (Gurneet Kaur and Abhinash Singla, 2016). A studies of different approaches for identifying manipulated reviews and proposes a new approach to identify those manipulated reviews using Decision Tree has proposed (Rajashree S. Jadhav and Deipali V. Gore , 2014). To identifying whether a review is fake or truthful one by Naïve Bayes Classifier, Logistic regression and Support Vector Machines has been proposed (Kolli Shivagangadhar et.al, 2015). A decision tree model is used to classify a record is to find a path that from root to leaf by measuring the attributes test, and the attribute on the leaf is classification result has been proposed(Qing-yun dai et.al, 2016). A new decision tree algorithm IQ Tree for classification problem has been proposed (Bhanu Prakash Battula et.al, 2015). To extract a kind of "structure" from a sample of objects (Bhaskar N. Patel et.al, 2012) .An implementation of decision tree algorithm and for comparative study and to analysis the performance has been proposed (Pooja Sharma and Rupali Bhartiya, 2012).

III. RESEARCH METHODOLOGY

Now a day's online shopping is in leading, because of fake reviews the ratings of the branded products becoming down. The major task is to focus on identifying the fake reviews. The Decision Trees classifiers Technique is used in this work, reviews has been extracted and collected to identify the fake review by using six different conditions that is star ratings, Response, Reply, Useful Profile, Profile status, Template conditions.

In this paper online reviews of the product are extracted by using Web harvy crawler. Potential features are extracted from set of collected reviews, then define the rules for Decision Rule Classifier to segregate the fake reviews by using decision Classifier technique based on several criteria. Final step is to identify the best feature to determine review is fake or not. The overall architecture of the proposed Fake review Identification model is as shown below in Fig 1.

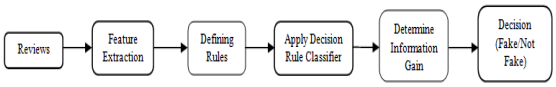

Fig 1: Block Diagram of the Proposed method

Following rules are defined to identify whether reviews are fake or not.

- R1(Response): It is to check whether the reviewer got any response for his review. Generally when a reviewer write a fake review on any particular product then company people will respond to the review quickly.
- R2(Useful profile): This is to check whether the reviewer profile is useful or not. If the reviewer profile is real profile then people will be liking the reviewer profile by stating "Use full information" .If the reviewer is giving the misguiding statements then will not be liking the profile.
- R3(Template): This is to check whether the reviewer uses standard Template or Not. If the review contains Template then definitely the review will be fake. If the reviewer wrote any negative or positive review within one sentence without stating the reason then it is fake.
- R4(Stars < 2): This is to check how much ratings the reviewer has given. Without any reason if the reviewer has given less than 2 ratings then it will be a fake review.
- R5(Reply): This rules is to check whether the reviewer has replied for the response from the company. If the reviewer is not a fake then definitely he will be replying for the response.
- R6(Thick): This rules is to check the thickness of the reviewer profile. The reviewer profile has full information (Thick) about his details which indicate reviewer is not a fake person else the reviewer profiles does not contain any detailed information (Thin), considered as fake review.

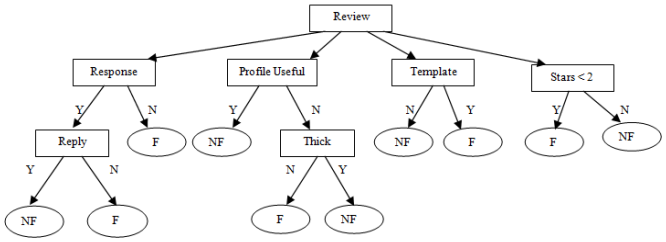

Fig 2: Decision Rule classifier tree.

Here: F=Fake, NF=Not Fake, Y=Yes, N=No.

Algorithm for decision rule classifier.

- Start
- Read the review and its features such as *Response, Useful profile, Template, Stars ratings, Reply, Thick.*
- If the response to the review is yes and there is no reply for response then it is a fake review otherwise it is not a fake review.
- If the reviewer profile is not useful and not thick then it is fake review.
- If the review uses template then it is a fake review otherwise it is not a fake review.
- If the star rating is less than 2 then it's a fake review otherwise it is not a fake review.
- Stop.

IV. EXPERIMENTAL RESULTS

Huge data from the web will be collected using the WEB Crawler to get the online reviews. The web Crawlers extracts all the reviews from the web and stores in the form of text documents. In this work Web harvy tool is used for collection of data and experimentation.

Dataset is prepared by collecting the reviews of the web users on the product called "LED TV" from the website "www.amazon.com",200 online reviews are collected for experimentation. From 200 reviews, 90 reviews has been detected as a fake reviews. Decision of the proposed approach is validated by manual review analysis, in which 94 reviews are detected as fake. Hence the success rate of 96 % has been achieved as shown in the Table 1.

TABLE 1: Experimental results for fake reviews

Number of Reviews collected	**200**
Product	LED TV
Website	Amazon
Fake Reviews Identified Using proposed method	90
Fake Reviews Identified Manually	94
Success rate	96 %

Experimental result shows the efficiency of the proposed system. The Table 2 shows the sample fake reviews of the product.

Table 2: Sample experimental results for fake reviews of a product.

SL.No	**Details**	**Review 1**	**Review 2**
1	PRODUCT	LED TV	LED TV
2	REVIEWS	Not a worth product	Waste of Money.
3	STARS	1	1
4	RESPONSE	N	Y
5	REPLY	N	N
6	USEFUL PROFILE	1%	1%
7	PROFILE STATUS	Thin	Thin
8	TEMPLATE	Yes	Yes

A. Evaluation

The Proposed approach is evaluated using Entropy and Information Gain to identify the best feature to decide the fake review among six features.

Table 5: Different Accuracy Measures for Identification of reviews

Sl.No	**Feature**	**Entropy**	**Information Gain**
1	Response(R1) and Reply(R5)	0.3	0.3
2	Profile Useful (R2) and Thickness(R6)	0.3	0.3
3	Template(R3)	0.3	0.2
4	Stars < 2(R4)	0.3	0.2

Sample Illustration of Entropy and Information Gain:

The above table shows the two methods to identify the best feature among these. The Entropy value will be same for all the measure here because of same leaf nodes at left and right. The Information Gain value is 0.3 which is High for Response, Reply and Profile Useful and Thickness and Information Gain 0.2 which is value is low for Template and Stars < 2 i,e 0.2 when compare to other. This Indicates that the *Response, Reply and Profile Useful, Thickness* will be the best feature to find out the accurate fake review.

Information needed to specify the exact physical state of a system, given its macroscopic. Entropy is an expression of the disorder, or randomness of a system, or of the lack of information about it. The concept of entropy plays a central role in information theory.Entropy is degree of randomness of elements determined using the equation (1).

$$\text{Entropy, } E(x) = -\sum p(x) \log p(x) \qquad (1)$$

Where P(x) is the probability of x.

Information Gain is nothing but Identification of best feature in the working set , that gives us less impurity. the amount of information gained about a random variable from observing another random variable. However, in the context of decision trees, the term is sometimes used synonymously with mutual information, which is the conditional expected value of the Kullback–Leibler divergence of the univariate probability distribution of one variable from the conditional distribution of this variable given the other one.

$$IG(x) = \sum_{ch=1}^{n} (E(x) - W * E(c$$

Where,

E(x): entropy of node x

W: Weighted avg.

E(child): Entropy of Childs.

Information gain with highest value is considered as best or most significant features in the identification of review is fake or not.

V. CONCLUSION

In this paper, Identification of fake reviews from the online reviews has been classified. Decision rule classifier is applied for various potential features such as *RESPONSE,USEFUL PROFILE, TEMPLATE,STAR RATING,REPLY,THICK* etc to identify whether review is fake or not. Decision of the classifier is validated by the information gain to identify most significant features. Proposed approach is experimented as reviews extracted from Amazon website on different products. The efficiency of the proposed approach has achieved 96 % success rate.

REFERENCES

[1]. Anusha Sinha, Nishant Arora, Shipra Singh, Mohita Cheema, Akthar Nazir "Fake Product Review Monitoring Using Opinion Mining" International Journal of Pure and Applied Mathematics, Volume 119 No. 12 2018, ISSN: 1314-3395.

[2]. Abha Tewari and smita Jangale "Spam Filtering Methods and machine Learning Algorithm - A Survey" International Journal of Computer Applications, Volume 154 – No.6, November 2016, ISSN- 0975 – 8887.

[3]. Bhanu Prakash Battula, KVSS Rama Krishna and Tai-hoon Kim " An Efficient Approach for Knowledge Discovery in Decision Trees using Inter Quartile Range Transform" International Journal of Control and Automation , Vol. 8, No. 7 (2015), pp. 325-334, ISSN: 2005-4297 IJCA.

[4]. Bhaskar N Patel, Satish G. Prajapati and Dr. Kamaljit I. Lakhtaria "Efficient Classification of Data Using Decision Tree" Bonfring International Journal of Data Mining, Vol. 2, No. 1, March 2012. ISSN 2277 – 5048.

[5]. Eka Dyar Wahyuni and Arif Djunaidy "Fake review detection from a product review using modified method of iterative computation framework" MATEC Web of Conferences 58, 03003 (2016), DOI: 10.1051/ matecconf/20165803003.

[6]. Elshrif Elmurngi and Abdelouahed Gherbi "Detecting Fake Reviews through Sentiment Analysis Using Machine Learning Techniques" International Conference on data Analysis, june-2018, ISBN: 978-1-61208-603-3.

[7]. Gurneet Kaur and Abhinash Singla "Sentimental Analysis of Flipkart reviews using Naïve Bayes and Decision Tree algorithm" International Journal of Advanced Research in Computer Engineering & Technology, Volume 5 Issue 1, January 2016, ISSN: 2278 – 1323.

[8]. Kolli Shivagangadhar, Sagar H, Sohan Sathyan, Vanipriya C.H " Fraud Detection in Online Reviews using Machine Learning Techniques" International Journal of Computational Engineering Research, Volume, 05 , Issue 05 , May – 2015, ISSN (e): 2250 – 3005.

[9]. Michael Crawford,Taghi.M,Khoshgoftaar,Joseph.D. Prusa Aaron N. Richter and Hamzah Al Najada "Survey of review spam detection using machine learning techniques" Journal of Big Data (2015) 2:23 ,Springer journal , DOI 10.1186/s40537-015-0029-9.

[10]. Manqing Dong , Lina yao, Xianzhi Wang Boualem Benatallah, Chaoran Huang Xiaodong Ning "Opinion Fraud Detection via Neural Autoencoder Decision Forest" a School of Computer Science and Engineering, University of New South Wales, Sydney 2052, Australia, 2018, www.elsevier.com.

[11]. Pooja Sharma and Rupali Bhartiya " Implementation of decision tree Algorithm to analysis of performance" International Journal of Advanced Research in Computer and Communication Engineering", Vol. 1, Issue 10, December 2012, ISSN : 2278-1021.

[12]. Qing-yun dai,Chun-ping Zhang and Hao wu " Research of Decision Tree Classification Algorithm in Data Mining" International Journal of Database Theory and Application, Vol.9, No.5 (2016), pp.1-8, ISSN: 2005-4270 IJDTA.

[13]. Rajashree S. Jadhav and Deipali V. Gore "A New Approach for Identifying Manipulated Online Reviews using Decision Tree" International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1447-1450,ISSN-0975-9646.

[14]. Rashmi Gomatesh Adike and Vivekanand Reddy "Detection of Fake Review and Brand Spam Using Data Mining Technique" International Journal of Recent Trends in Engineering and Research, Volume 02, Issue 07; July - 2016, ISSN: 2455-1457.

[15]. Salma Farooq and Hilal Ahmad Khanday "Opinion Spam Detection: A Review" International Journal of Engineering Research and Development, Volume 12, Issue 4, e-ISSN: 2278-067X, p-ISSN: 2278-800X, April 2016.

[16]. Shashank Kumar "Research on Product Review Analysis and Spam Review Detection" Research gate- Conference Paper, February 2017, 317932754.