Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.9854903

COMPUTATIONAL INTELLIGENCE-BASED METHODS FOR SPOTTING FAKE REVIEWS: A SURVEY

[1]Digvijay Singh, [2]Dr. Minakshi Memoria
[1]Research Scholar, Department of CSE, Uttaranchal University, Dehradun, Uttarakhand, India
[1]Assistant Professor, Department of CSE, Dev Bhoomi Uttarakhand University, Dehradun, Uttarakhand, India
[2]Associate Professor, Dept of CSE, Uttaranchal University, Dehradun, Uttarakhand, India
[1] digvijay.dbgi@gmail.com, [2] minakshimemoria@gmail.com

Over the last decade, online marketplaces are evolving at a high rate and customers are relying more on the internet for their purchases. A more positive and favourable review about a product or service draws more customers, generating more profit. At the same time, Reviews without experience with the product have also been written to attract customers. Hence, spotting these bogus reviews is a crucial and important research field. The capacity to recognize fake reviews is influenced by both the review's fundamental elements and the reviewers' actions. Extensive research has been done in this area to identify fraudulent reviews and reviewers are generally involved in this. However, the group of bogus reviewers works together as a team to concentrate on particular products and submit false reviews of those products in large numbers. The goal of this work is to present a robust, thorough comparative and comprehensive analysis for applying machine learning to identify false reviews and reviewers.
Keywords: Fake review, Fake reviewers, Spam opinion, feature engineering.

1. INTRODUCTION

Because of the increasing growth of e-commerce, which makes it easier to buy and sell goods and other services online, customers are gradually more adopting online stores for shopping in order to save time. These online platforms also provide them the liberty to post their opinion about their purchases [1]. These internet reviews can significantly improve the shopping experience for a new customer. Among these past customers' opinions, favourable opinions influence the buyers to go for that goods or brands. Also, favourable ones frequently result in big financial advantages while negative ones frequently result in financial or sales losses [2,3]. Due to this, the majority of business owners depend on consumer opinion to amend their business tactics and maintain the quality of their products or services [4,5]. Spam material is described as insignificant or uninvited data that is combined with opinions and utilized for marketing, promotion, information dissemination, and monetary gain [6]. The service provider solicits feedback from clients regarding their use of its goods and services to determine whether an individual customer is satisfied with their purchase.

Customers can rate services according to their personal experiences, whether the service was good or bad. Blindly believing these comments, however, is risky for both service providers and users. The terminology "fake opinion" refers to information provided in reviews that is

false or erroneous to mislead readers and hamper product sales [7]. Spam reviews are basically classified into three categories [8]:

FIG 1. Review Classification

It is also important to determine fraudulent reviewers who are creating bogus reviews while attempting to locate review spam. Spammers/fake reviewer can be categorized into two types: 1) Individual who is using different IDs and also registered on various portals with different or the same name, 2) A group of imposters who collaborates, registers at many websites, and posts spam on these websites.

Some more fundamental behaviors of fraudulent reviews [9]:

Very less information about the reviewer: Individuals who have less personal data and fewer social connections are typically considered to be dishonest or spammers.

Review content likeness: Spammers routinely create and submit online reviews that are identical or nearly identical to one another. Such scores can reveal fake reviews.

Brief review: Review information in capital letters and with grammatical errors is typically submitted by scammers because they are continuously focused on getting quick returns. They pay considerably closer attention to the trademarks on the goods.

Posting and uploading of reviews in a short time span: Examining the date and time that each review was posted is the simplest technique to identify spam. These reviews might be spam if they were all posted simultaneously.

Behavioral characteristics have been taken into account in the phony review detection method in other papers. [10] has taken into consideration a number of reviewer behavior factors, including average rating and ratio of reviews on Amazon.

2. RELATED WORK

The procedure of fraudulent review detection seeks to analyze, spot, and categorize fraudulent and honest reviews of products on e-commerce portals. Fake review analysis has gained popularity as a study area over the past decades. Because fake/spam review recognition has such a big effect on clients and online commercial enterprises, several investigators have done research on the topic. Feature engineering is the method of creating or extracting significant characteristics from text data. The two methods utilised in fake review detection research are linguistic (review-centric features approach) and behavioural (spammer reviewer features), which only take into account the content of a specific review. The next section lists and explains both features.

Stylometric characteristics or features: These characteristics or features are crucial for recognizing the text writing patterns of assessors and spotting deceit. There are two different

categories of physical characteristics. Lexical characters are the first category of characters. These characters include the ones that appear in each word of the review text (T), the proportion of numeral letters to (T), the percentage of uppercase and lowercase characters to (T), the percentage of special characters like >, %, [,], /, #, -,;,-, *, &, @, $ etc. Another one includes feature based on lexical words comprising the count of words, a the proportion of words in the statement, a token's dimension percentage, a ratio of the characters in words to T, and a ratio of small words (between three and five characters). A syntactic-based feature called stylistic illustrates the reviewer's writing style in the review comment. The frequency of punctuation marks (.?!: ;, ") is a syntactic or grammar structure[12].

Number reviews/day: 90% of honest individuals can post only one opinion after making a purchase, however around 70% of fraudsters write more than five reviews daily. In order to help in identification of spammers, this considers the amount of reviews that users may submit [10,11]. Language investigation and word count is an analytical implements that users can use to build dictionaries that reflect their personal interests.

When parts of speech (POS) were added, LIWC's ability to identify spam reviews in [1] performed better. A complex linguistic trait is thought to be LIWC. [13] recommended a study on the impact of reviewers' social links on online customer review fraud detection. Yelp's review dataset (135,413) was collected and pre-processed for use in their experiment, with 103,020 of them being recommended and 32,393 not. The backpropagation neural network technique was then used to accomplish the classification of reviews into authentic and fraudulent using the behavioural and social interaction attributes of users.

The initial study for spotting the spam/fake reviews was published by Jindal et al. [14]. Untruthful, brand-specific, and unconnected reviews were the three categories of spam that the writers identified. They employed the logistic regression supervised machine learning method to group duplicates or near to duplicate and nearly identical Amazon product reviews as fake or not, and achieved the resulting area under the curve was 78%. (AUC).

The distribution footprints of reviewers were the focus of Feng et al.'s [15] investigation, which also indicated a relationship between spreading abnormalities and dishonest behaviours. The phony hotel reviews obtained from the Trip Advisor website were scrutinized. Fitzpatrick et al. [16] state that one specific approach to handle a typical recognition challenge where it is able to employ both written and oral cues is a deceptive review analysis. The comprehensiveness, size, way of writing and correlates of cognitive qualities of the review text were all investigated, according to Banerjee et al. [17].

3. REVIEW FEATURE AND FEATURE ENGINEERING FOR FAKE REVIEW DETECTION

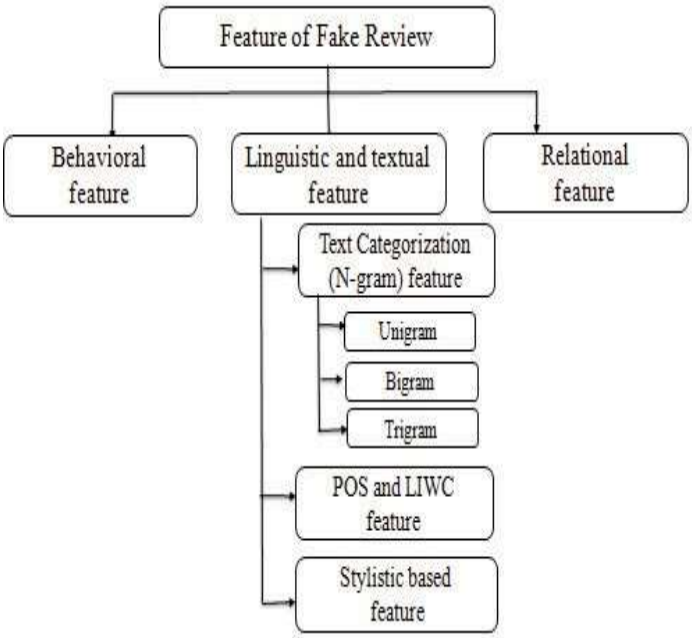

Fig 2: Various Features of Fake Reviews

3.1 Sentiment Score

A method of evaluating and identifying the text's polarity (positive, negative, or neutral) is the sentiment score. When evaluations reflect a considerable amount of negative emotion, negative fraudsters are trained to include more negative than good phrases in their reviews. The emotion score should be calculated for each review text since positive fraudsters, on the other hand, are always trained to use more positive terms. To determine the sentiment score in each review, use the following formula:

$$S(r) = (P(WOR) - N(WOR)) / T(WOR)$$

where P(WOR) represents the number of favourable words, N(WOR) represents the number of unfavourable words, and T(WOR) represents the whole amount of words (positive and negative) in the review text. S(r) represents the review's emotion (S).

3.2 N-Grams

N-gram features are a technique for choosing N nearby words as traits from the text's contents. [2]. A unigram is formed when N = 1 (single word) is assigned at a time. When two neighbouring words (N = 2) are assigned at the same time, it is known as a bigram, and when four neighbouring words (N = 4) are assigned at the same time, it is known as four-gram.

3.3 Feature Extraction (TF-IDF)

The frequency-inverse document frequency (TF-IDF) technique is utilised in text classification systems for feature mining and demonstration. It's used in data extraction and natural language processing. In addition, the TF-IDF statistical method is used to rank the importance of each term or word in the dataset. It is divided into two parts: term frequency, which determines the frequency of specific words in texts in order to determine how similar they are, and similarity. The TF formula is as follows:

$$TF(w)d = \frac{nw(d)}{|d|}$$

Set D denotes a collection of documents, and d is one among them. The number of frequently recurring words, w, contained in document d is represented by nw (d). Document d is made up of phrases and words, w. As a result, the following formula is used to determine the volume of document d:

$$|d| = \quad nw\,(d),\ w \in d$$

The above calculation was used to determine how many times a specific word appeared in the manuscript. The second component was calculated by dividing the total number of documents in the corpus by the number of documents containing that specific word. The formula for determining IDF is as follows:

$$IDF(w)d = \frac{1 + \log |D|}{|\{d : D|w \in d\}|}$$

Thus, the following formula may be used to calculate the TF-IDF for word w associated with document d and corpus D:

$$TF.IDF = TF(w)_d \times IDF\ (w)_D$$

The classifier can see from the TF-IDF that some words are used a lot in the document.

3.4 Performance parameters

Performance of the classification model is measured using different performance parameters such as Accuracy, Error rate, Sensitivity, and Specificity [18-20].

- Accuracy: is the measure of the percentage of correctly identified instances by model. It is calculated as

$$\text{Accuracy} = tp{+}tn\ tp{+}fn{+}fp{+}tn$$

- Error rate: is the measure of the percentage of incorrectly identified instances by the model.

$$\text{Error Rate} = \frac{fn{+}fp}{tp{+}fn{+}fo{+}tn}$$

- True positive rate/Sensitivity is the proportion of positive samples accurately detected by the classification model.

$$\text{Sensitivity} = tp\ tp{+}fn$$

- True negative rate/Specificity: This is the percentage of accurately predicted negative samples by the classification model.

$$\text{Specificity} = tn\ fp{+}tn$$

Where TP stands for true positive, it refers to the number of bogus reviews that the model detects. TN stands for true negative and represents the total number of records that the model recognizes as not phony. The number of false positives (FP) is the number of records that the model recognizes as phony but the reviews are not. FN stands for false negative, which is the number of records that are not phony but are detected as such by the model.

4. REVIEWER-CENTRIC FEATURES

4.1 Behaviour trait of the user

Information about user account behaviour, including how many reviews have been published, their impact on customer feedback overall, etc. One illustration of such a feature is the negative ratio.

4.1.1 Ratio of Negative

Spam account user profiles utilize very critical or destructive language while writing reviews that criticize rival companies, or they may give such companies lower ratings.

4.2 Linguistics Features for users

This element contains information about the review's linguistics, such as the words and types used by particular users and how similar they are. For convenience, the content similarity factor can be used to refer to the average and maximum content similarity features of this type.

4.2.1 Factor of Text Similarity

Maximum, Average Similarity: When spammers write their opinions, there is always a certain level of similarity. This is because of the simple fact that spammers frequently do not alter their template and seek to leave as many reviews as they can, allowing some recognition through this design.

5. CONCLUSION

This study comprehensively examined the most significant work in machine learning-based fake review identification to date. First, we looked over prior researchers' feature extraction methodologies. Then, we went over a few well-known machine-learning algorithms and neural network models for false review detection. Traditional statistical machine learning improves the performance of text classification models by improving feature extraction and classifier building. Deep learning, on the other hand, enhances the presentation learning approach, the algorithm's structure, and new knowledge to improve performance. Finally, we emphasised the importance of further research in this field as well as potential future directions.

We conclude that supervised machine learning was used in the majority of past research to detect bogus reviews. However, in order to determine if a review is false or not, supervised machine learning requires access to a labelled dataset, which can be difficult in the fake review detection business. Due to the difficulties in acquiring tagged datasets, we discovered that the most often used datasets in the current studies are built utilising a crowdsourcing design.

We believe that researchers who have a thorough grasp of the essential elements of this discipline will find this survey to be helpful.

References

[1] S. N. Alsubari, S. N. Deshmukh, M. H. Al-Adhaileh, F. W. Alsaade and T. H. Aldhyani, "Development of integrated neural network model for identification of fake reviews in e-commerce using multidomain datasets," *Applied Bionics and Biomechanics*, vol. 2021, pp. 1–11, 2021.

[2] Y. Li, X. Feng and S. Zhang, "Detecting fake reviews utilizing semantic and emotion model," in *2016 3rd Int. Conf. on Information Science and Control Engineering*, Beijing, China, pp. 317–320, 2016.

[3] X. Hu, J. Tang, H. Gao and H. Liu, "Social spammer detection with sentiment information," in *2014 IEEE Int. Conf. on Data Mining*, Shenzhen, China, pp. 180–189, 2014.

[4] F. Long, K. Zhou and W. Ou, "Sentiment analysis of text based on bidirectional LSTM with multi-head attention," *IEEE Access*, vol. 7, pp. 141960–141969, 2019.

[5] V. W. Feng and G. Hirst, "Detecting deceptive opinions with profile compatibility," in *Proc. of the 6th Int. Joint Conf. on Natural Language Processing*, Nagoya, Japan, pp. 14–18, 2013.

[6] S. J. Delany, M. Buckley and D. Greene, "SMS spam filtering: Methods and data," *Expert Systems with Applications*, vol. 39, no. 10, pp. 9899–9908, 2012.

[7] S. Sarika, M. S. Nalawade1 and S. S. Pawar, "A survey on detection of shill reviews by measuring its linguistic features," Int. J. Emerg. Trends Technol. Comput. Sci., vol. 3, no. 6, pp. 269–272, 2014.

[8] L. Bing, Web Data Mining. Book. Springer, Berlin Heidelberg New York, 2008.

[9] N. Hussain, H. T. Mirza, I. Hussain, F. Iqbal and I. Memon, "Spam review detection using the linguistic and spammer behavioral methods," *IEEE Access*, vol. 8, pp. 53801–53816, 2020.

[10] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos *et al.,* "Exploiting burstiness in reviews for review spammer detection," in *Proc. of the Int. AAAI Conf. on Web and Social*, Media, Massachusetts, USA, pp. 175–184, 2013.

[11] A. Heydari, M. A. Tavakoli, M. N. Salim and Z. Heydari, "Detection of review spam: A survey," *Expert Systems with Applications*, vol. 42, no. 7, pp. 3634–3642, 2015.

[12] S. Shojaee, M. Murad, A. B. Azman, N. M. Sharef and S. Nadali, "Detecting deceptive reviews using lexical and syntactic features," in *2013 13th Int. Conf. on Intelligent Systems Design and Applications*, Salangor, Malaysia, pp. 53–58, 2013.

[13] K. Goswami, Y. Park and C. Song, "Impact of reviewer social interaction on online consumer review fraud detection," *Journal of Big Data*, vol. 4, no. 1, pp. 1–19, 2017.

[14] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proc. of the 2008 Int. Conf. on Web Search and Data Mining*, Palo Alto, California, USA, pp. 219–230, 2008.

[15] S. Feng, L. Xing, A. Gogar and Y. Choi, "Distributional footprints of deceptive product reviews," in *Proc. of the Sixth Int. AAAI Conf. on Weblogs and Social*, Media, Dublin, Ireland, pp. 98–105, 2012.

[16] E. Fitzpatrick, J. Bachenko and T. Fornaciari, "Automatic detection of verbal deception," *Computa- tional Linguistics*, vol. 43, no. 1, pp. 269–271, 2015.

[17] S. Banerjee, A. Y. Chua and J. J. Kim, "Using supervised learning to classify authentic and fake online reviews," in *Proc. of the 9th Int. Conf. on Ubiquitous Information Management and Communication*, BaliIndonesia, pp. 1–7, 2015.

[18] Verma L, Srivastava S, Negi P C (2016) A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data Journal of medical systems, 40(7), 178

[19] Böger, B., Fachi, M. M., Vilhena, R. O., de Fátima Cobre, A., Tonin, F. S., & Pontarolo, R. (2020). Systematic review with meta-analysis of the accuracy of diagnostic tests for COVID-19. American journal of infection control.

[20] Guthrie, W., Wallis, K., Bennett, A., Brooks, E., Dudley, J., Gerdes, M., & Miller, J. S. (2019). Accuracy of autism screening in a large pediatric network. Pediatrics, 144(4), e20183963.