Fake Review Detection using Classification

Neha S. Chowdhary
Department of Computer Applications
Veermata Jijabai Technological Institute

Anala A. Pandit, PhD
Department of Computer Applications
Veermata Jijabai Technological Institute

ABSTRACT

In today's world, where Internet has become a household convenience, online reviews have become a critical tool for businesses to control their online reputation. Reviewing has changed the face of marketing in this new era. Nowadays, most companies invest money in mining the reviews to gain insights into customer preferences as well as to gain competitive intelligence and are hiring individuals to write fake reviews. The fraudsters' activities mislead potential customers and organizations reshaping their businesses and prevent opinion-mining techniques from reaching accurate conclusions. Thus, it has become essential to detect fake reviews to bring to surface the true product opinion. This paper focuses on product reviews and detecting spam fake reviews among them using supervised learning techniques using synthetic fake reviews (to cover all types) as a training set. Term frequency and user review frequency are two features whose impact on classification model is studied in this paper. It classifies the reviews to test the accuracy of the model. The results have been encouraging with an accuracy of over 98%.

General Terms

Review spam, Opinion mining, fake review detection, Review spam, fake reviews, opinion spam

Keywords

Review spam, Opinion mining, fake reviews, Naïve Bayes classification, Opinion Spamming, Random Forest Classifier, Classification Model Evaluation Measures

1. INTRODUCTION

The Internet has vastly changed not only the customers' perspective on buying online but also the business processes. One could say, there are two worlds: one before ecommerce and one after it. Nowadays, customers prefer buying most products or services through e-commerce or online portals. These e-commerce or online portals have given rise to new techniques for marketing as well as influencing customer's decision i.e. reviews. Reviews refer to any view or opinion made about a product or service by an individual usually not associated with the business. The reviews that appear on the website are specifically referred to as user generated content (UGC) [2]. Reviews present a new way to learn about customer preferences, product quality as well as product's shortcomings. A review left online is a permanent record of that customer's experience; it can be found by anyone and reach a far wider audience than ever before. Today, almost every online portal enables posting reviews, images and expressing our own views about products or services in blogs or forums or dedicated review websites like Zomato, Yelp etc. This user generated content can be used to discover customers' preferences, the strengths and weaknesses of the product, study the market conditions, identify new product launch opportunities and strategize to win from competitors.

The easy possibility of monetization using the intelligence obtained from reviews has led to the problem of opinion spam or creation of fake reviews. Companies hire spammers to write undeserving positive reviews to promote their products or negative reviews to destroy the competitor's reputation. Unfortunately, driven by the desire for profit or publicity, fraudsters have produced deceptive (spam) reviews [1]. There are various reasons that motivate people to write a review, like the desire to affect a change in the business, product or service or anger at poor product / service or delight at a great product / service or when a product / service is not as expected. The reason could also be an inherent desire to help the public, for instance if the customer is an expert in the product and one would want to share the expertise. Before making any decision about the product, one always first checks the reviews about the product or restaurants or services etc. [3]. Positive opinions can result in significant financial gains and/or fame for organizations and individuals. This provides a good incentive for creation of review/opinion spam. Fake reviews can be written by a shop retailer, business personnel, or individuals who maintain their online identity. As the reviews have become an important decision-making factor, some business hire experts to write spam review with the objective/ intention to promote their image or damage the competitor's reputation. There can be two types of fake review written for this purpose either forged positive review or undeserving negative review to encourage/discourage the customers from purchasing the product.

In this paper, fake review detection has been considered as binary classification problem with the two classes being: fake and genuine. This paper focuses on detecting fake reviews from a set of product reviews by simulating spam reviews that incorporates various types of opinion spam review features and building a training set and then classifying it using Naïve Bayes Classification and ensemble classification model like random forest to test the accuracy of the model. Various features have been considered while classifying fake reviews. However, the author's introduced two more features:

i. Using terms or bag of words as features for classification of reviews as either fake or genuine.
ii. The impact on the classification model considering the user review frequency on the same product

Classifying with these features, improved the accuracy by 26% for Naïve Bayes classifier. The F-Score has taken a leap by 23% for Naïve Bayes and 1% for Random Forest classifier.

The remainder of this paper is structured as follows: the next section discusses the work done in the fake review detection domain. Section 3 gives an idea about the cleaning and pre-processing done prior to classification. Section 4 presents the proposed technique for identifying spam reviews. Section 5 gives a brief overview about the dataset, the experiment carried out and the analysis of the results. Finally, Section 6 presents the conclusion derived from the results and the future work that needs to be done. Section 7 mentions the references used in this paper.

2. PAST OR RELATED WORK

Review content, emotional diversity and user behavior analysis as well as training models for classification has been used by many researchers for detecting fake reviews. The review spam problem was first formulated by Jindal and Liu in the context of product reviews [4] [5]. Most of the previous work on fake review detection can be categorized into 3 main detection techniques: Review content-based detection, Deviations among the rating-based detection and Review content along with user behavior-based detection.

Review Content-Based Detection: This category of technique is focused on identifying/classifying fake reviews purely based on the opinion/view expressed in an opinion. In [6], the authors have approached the problem of fake review detection as a binary classification problem and using different approaches for representing the content and using two popular classification techniques: LS-SVM and Naïve Bayes classifier, they classified the reviews. One thing to note in [6] is that they assume independence among the different features which may not always be the case.

Deviations among Rating-Based Detection: This category of technique involves identifying fake reviews by calculating rating based on the review content and comparing it with the given rating to find deviations. It can also encompass classifying reviews as fake if their ratings are outliers. Authors in [7] have incorporated the deviations between ratings and reviews by calculating the rating of the review based on the content and then finding difference from the rating given by the reviewer. Based on a selected threshold value the authors determined whether the given review is fake or not. In [7], user behavior is not taken into consideration or there have been no attempts to identify spammers or group spammers.

Review Content along with User Behavior-Based Detection: This category of techniques encompasses various features related to review data along with features corresponding to user behavior. Some example for user behavior include user frequency, time density [8], store density [8] etc. Jindal and Liu categorized opinion spam into three categories [4] [5]:

- Type 1 (False Opinions): These are either forged positive or undeserving negative reviews.
- Type 2 (Reviews on Brand): These encompasses the opinion expressed on brand rather than the product itself.
- Type 3 (Non- reviews): This type of review does not really express any view or opinion. They basically only influence the overall rating of the product.

From these categories, Type 2 and Type 3 categories are identified by training the model using supervised techniques. Type 1 were identified using duplicates. In [4] [5], Jindal and Liu expressed that outlier reviews must be considered as suspicious, but an expert spammer may give rating in alignment with the average rating of the product and the content may be different. However, they have not considered incorporating deviations between rating and the review content to classify outlier reviews as fake reviews. Authors in [8] have considered a varied number of features associated with the reviews, like reviewer, user related behavior, features to identify expert spammers and identified emotions expressed in the reviews, similarity between the reviews, category, time and the store density for identifying the fake reviews. One good assumption in [8] is that they have identified that spammers may target the same product category which makes it easier to identify different spammers. Along with it one needs to consider if reviews are similar for different products that vary only in limited features, then those may not be spam. To detect spammers and groups of spammers, authors in [9] have further considered the user related features and the sequence as well as frequency with which a user posts reviews. Along with this, similarity among review from same products and different products are also taken into consideration. A step further is taken by authors in [10] by using natural language processing for topic detection for a given review. The authors have taken into consideration that a spammer will not share his/her experience with minute details rather the expression will be more of a generalization. Authors in [11] have taken it a step ahead to identify the features related to the product, to mine the opinion a reviewer has on the different features of the product. The authors then compared it with what other reviews specified and the deviations among them were used to tag the outliers as ‘suspicious’. Further on, they have considered 3 different factors to identify fake reviews: the trustworthiness of the reviewer, reliability of the product and the honesty of the review.

In this paper, the different features identified thus far, have been used. Additionally, in this work, the authors have also considered the frequency of review by a user for similar product. To incorporate in the training model for classification, variety of synthetic fake reviews were simulated to ensure that all possible types of fake reviews are covered to achieve higher accuracy. The proposed technique falls into Review Content along with User Behavior-Based Detection method.

3. PRE-PROCESSING

Table 1 contains the column headers and sample data of the original reviews extracted using import.io. To get accurate results, it was imperative to get precise and formatted data for classification. Thus, cleaning and pre-processing was an important step.

The dataset was cleaned to get the rating and the influence in number format from text (see table 1). The created date field was appropriately formatted to convert it from text to date type. Similarly, other fields like Verified Purchase were also formatted as desired. All white spaces and empty values were replaced with nulls for text columns and 0 for quantity fields. After the initial cleaning, the next step was to replace abbreviations like ‘Sr. to senior’, remove punctuations and convert the numbers to text. Further the stop words were removed.

Stop word removal in natural language processing is the process of removing words that would not contribute to classification, it is the process of removing most commonly used words from the text that even search engines do not use for indexing like ‘the, a’ etc. Stemming is the process of reducing the word to its base or root or stem word. For example: ‘loud’, ‘loudly’, ‘loudness’ can be reduced to make one base word ‘loud’. Once pre-processing is completed, the data is then ready for classification.

Table 1: Sample Data

Column Headers	**Sample Data**
Seller Website	Amazon
Product Name	Sennheiser CX 180 Street II In-Ear Headphone
Ratings	5.0 out of 5 stars
Review Title	... head phones around 2 years still no damage very good quality product with good sound i love this product
Reviewer Id	Amazon Customer
Review Date	on 30 June 2016
Review Content	I am using this head phones around 2 years still no damage very good quality product with good sound i love this product...;
Influence	One person found this helpful.
Verified or Not	Verified Purchase

4. PROPOSED TECHNIQUE

As already mentioned, the fake reviews can be categorized mainly into 3 categories as given by Jindal and Liu in [1]. The technique proposed involves simulating fake reviews encompassing several different features along with considering the user frequency in posting reviews for the same product. In this paper, the duplicates from the same user on the same product have not been eliminated. Jindal and Liu considered them as mistake, but the authors believe that if those reviews are posted at different time then those should be considered as suspicious. Also, if the user is posting reviews frequently on the same product and they have deviations then those are considered fake as well. Different Features can be of two types *Data Centric Features* and *Review Centric Features.* They are defined as follows:

4.1 Data Centric Features

It encompasses features related to the review content like emotion expressed, opinion orientation, rating on the content etc.

- Reviews on Brand only (F1): For example: For a laptop, a person may give review as “HP is the best brand”. These do not reflect or express a person’s opinion on the product but more on the brand. This also introduces bias in the review or opinion.
- Reviews on competitor (F2): These types of reviews basically compare the product with competitors’ or praise the competitors’. Again, these do not give an honest view but an attempt by spammers to damage the reputation of their competitors’ product or boost their own reputation.
- Review on seller website (F3): These types of review contain a view or opinion about the ecommerce website rather than the product itself. Spammers may use these types to affect the overall rating of the product. For example, “Amazon delivered the product in one day. Amazon is simply best”
- Non-reviews (F4): These are basically factual statements or out of context conversation that have no relation with the product. Again, these are targeted to affect the summary of rating for a review.
- Reviews deviating with the given rating and calculated rating based on content (F5): In [3] the authors used this as selecting feature for detecting fake reviews. An experienced spammer may write a review differently and give the rating in alignment with other users. So, it is essentially important to detect the deviations among rating and calculated rating based on the review content.
- Outlier Reviews (F6): The review that is extremely positive or extremely negative as compared to the average rating must be deemed suspicious.
- Term Frequency (F10): The various terms or words are also considered as distinguishing features for classifying the reviews. As the whole text may consider a lot number of terms, we consider only those document terms that at least occurs in 5 different reviews, to be included as a classification variable.

4.2 Reviewer Centric Features

It involves features related with user behavior. For example, the frequency with which the user posts reviews, whether it’s a duplicate from another users’ review or his/her own duplicate etc.

- Duplicate Reviews from Different User-ids (F7): Usually spammers will use the written reviews from other users, to reduce efforts. Some spammer groups may write duplicate reviews from different user ids. So, these reviews should also be detected to identify spammers
- Duplicate Reviews from Same User-ids (F8): Jindal and Liu in [1] eliminated them stating that it is possible the user may mistakenly post multiple reviews by pressing an ‘enter’ number of times. But if the same review is posted by same user at different time then the user and the review must be termed suspicious.
- Different Reviews from same user-ids (F9): It is unusual human behavior to keep posting reviews on similar products repeatedly. So, if there is a scenario where the user posts review on the same product and if the reviews are not in alignment with the previously posted reviews by the same users then these reviews and the user must be considered as spam.

5. EXPERIMENTATION AND RESULTS

5.1 Process and Dataset

Amazon Reviews for Sennheiser CX 180 Headphone were scraped in February 2018. A total of 21760 reviews were scrapped using import.io. The review consists of seller website, product name, rating, reviewer id, review topic, review content, review date, influence (how many people found it helpful) and whether the purchase is verified or not. These are some of the features that were selected; more features can be extracted if desired from import.io. Import.io is GUI based web tool to extract data from web pages. It allows first 500 URLS to be extracted for free. It basically gives the user the structure of the desired web page, which can then be used to select the elements that is desired.

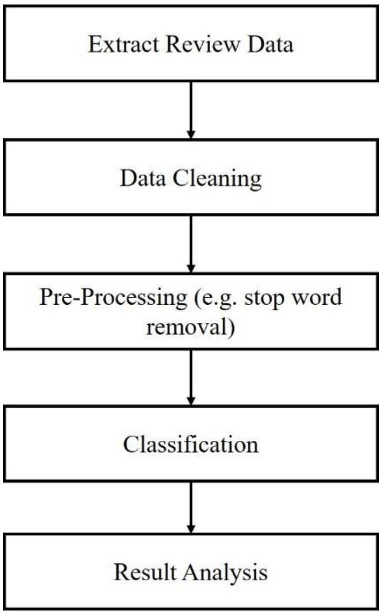

Figure 1: Steps for Fake Review Detection

5.2 Results & Evaluation Measures

Naïve Bayes classification is based on the famous Bayes theorem, and it assumes independence among its features. In simple terms, a Naive Bayes classifier assumes that the presence of a feature in a class is unrelated to the presence of any other feature [12]. Random Forest is a supervised learning algorithm. it builds, is an ensemble of Decision Trees, most of the time trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. [13]

In this paper, the classification is evaluated by the measure of accuracy, positive predictive value or precision, negative predictive value, recall or sensitivity and specificity all of which are calculated from the confusion matrix as given by table 2.

i. Accuracy: the proportion of the total number of predictions that were correct. [14]

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

ii. Precision or Positive Predictive Value: the proportion of positive cases that were correctly identified. [14]

$$Precision = \frac{TP}{TP + FP}$$

iii. Negative Predictive Value: the proportion of negative cases that were correctly identified. [14]

$$Negative\,Predictive\,Value = \frac{TN}{TN + FN}$$

iv. Recall or Sensitivity: the proportion of actual positive cases which are correctly identified. [14]

$$Recall = \frac{TP}{TP + FN}$$

v. Specificity: the proportion of actual negative cases which are correctly identified. [14]

$$Specificity = \frac{TN}{TN + FP}$$

vi. F – Measure: It is also called the F Score or the F1 score. Put another way, the F- Measure is the balance between the precision and the recall. [15]

$$F - Measure = 2 \times \frac{Precison \times Recall}{Precision + Recall}$$

Table 2: Confusion Matrix

	Reference	
Prediction	Fake	Genuine
Fake	TP	FP
Genuine	FN	TN

In the experiment, the classification model is trained by dividing the data into 70:30 ratios. 70% is used as a training set while rest 30% as testing set. Naïve Bayes and Random Forest Classification is done twice by considering different number of features for the model both the time. First by taking into considerations the features F1 through F6 and later all feature through F1 – F10 are considered and then classified using Naïve Bayes algorithm and Random Forest technique.

In this research, each word in a review is represented as a term or feature and each review as vector of features or bag of words. This bag of words tokens is then represented by document term matrix (DTM). The rows of the DTM corresponds to reviews in the collection, columns correspond to terms, and its elements are the term frequencies. Not all the term features are considered for classification but only those which appear in at least 5 reviews. Apart from this all the other features that have been defined are used to classify whether a given review is genuine or fake.

5.3 Result Analysis

The accuracy when considering features F1 to F10 is observed to be 98% for Naïve Bayes and 99% for Random Forest. From table 3, it can be clearly concluded that the Random Forest classifier performs better than the Naïve Bayes classifier when only features F1 to F6 are into consideration. When features F1 to F10 are examined, both the Random Forest and Naïve Bayes classifier are at par in terms of accuracy, but when looked at the f-measure, it is noticed that random forest has a gain of approximately 43%. Random forest is more of a balanced classifier.

Table 3: Results

Classification Technique	**Features**	**Accuracy (%)**	**Precision (%)**	**Negative Predictive Value (%)**	**Recall (%)**	**Specificity (%)**	**F- Measure (%)**
Naïve Bayes Classification	F1 to F6	72.08	30.22	79.09	19.50	87.12	23.70
	F1 to F10	98.15	44.62	99.13	48.74	98.97	46.59
Random Forest Classification	F1 to F6	99.45	99.27	99.51	98.38	99.78	98.82
	F1 to F10	99.55	99.94	99.44	98.12	99.98	99.02

From table 3, it can be clearly concluded that the Naïve Bayes models are more accurate in classifying the genuine reviews rather than fake reviews as negative prediction value and specificity is far greater than the positive prediction value and recall. This model can be better used to determine genuine reviews and can give the consumer the real opinion about the product by summarizing the classified genuine reviews. In case of Random Forest classification, the model performs well in classifying both genuine as well as fake reviews. The Naïve Bayes model using features F1 to F10 has improved significantly as the f-measure has increased up to approximately 12% and it also improves the precision value by 15%. From table 3, it can be can also be seen that considering all the features, gives a very good accuracy. All the features use the bag of word representation along with the emotional diversity in the reviews.

6. CONCLUSION AND FUTURE WORK

The reviews play an important role in how the end users view a product. It is human nature to judge a metaphorical book by its reviews. No matter how good a product is or how good a user’s experience with it is, they tend to believe in the word of mouth (WoM). Thus, fake reviews are a hidden threat to e-commerce businesses. It is unethical but widely practiced. The advancement in technology has brought with it tools to deal with said fake reviews and the subsequent fallout. Using these tools, e-commerce sites can curb this malpractice and bring integrity to the e-commerce business.

This paper proposes two new types of features: user review frequency on the same product and term frequency feature. Experiments show that the model and technique defined are accurate in classifying fake and genuine reviews. Experimental results lead to the conclusion that Random Forest performs better than Naïve Bayes and can be used to detect the genuine as well as the fake reviews. One thing to note here is if the interest lies in getting to surface the true opinion about the product Naïve Bayes algorithm must be preferred, as the results depicts its partiality in detecting genuine reviews well. Considering the frequency of various terms as well as reviewer centric features increases the accuracy and it also improves the precision value significantly as compared to the model which focused only on the data centric features in case of Naïve Bayes classifier. The short coming of the proposed approach is that it is a time-consuming process to manually label fake reviews and that process needs to be automated with the use of supervised techniques.

Future work includes collecting review data from different review websites, computer aided generation of fake reviews while incorporating the different features and context aware classification to avoid misclassification of fake reviews. The future model will also include more reviewer related characteristics as it also helps to identify spammers and the sequence with which a reviewer posts review. Context aware classification is necessary as it helps to identify sarcasm and other human emotions that is missed in the given model.

7. REFERENCES

[1] Atefeh Heydari, Mohammad ali Tavakoli, Naomie Salim, and Zahra Heydari. 2015. Detection of review spam. Expert Syst. Appl. 42, 7 (May 2015), 3634-3642. DOI: https://doi.org/10.1016/j.eswa.2014.12.029

[2] 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), Tokyo, 2017, pp. 658-666. doi: 10.1109/DSAA.2017.51

[3] Nitin Jindal and Bing Liu. 2007. Review spam detection. In Proceedings of the 16th international conference on World Wide Web (WWW '07). ACM, New York, NY, USA, 1189-1190. DOI: https://doi.org/10.1145/1242572.1242759

[4] N. Jindal and B. Liu, "Analyzing and Detecting Review Spam," Seventh IEEE International Conference on Data Mining (ICDM 2007), Omaha, NE, 2007, pp. 547-552. doi: 10.1109/ICDM.2007.68

[5] N. Jindal, B. Liu. "Opinion spam and analysis." International Conference on Web Search and Data Mining ACM, 2008, pp. 219--230.

[6] R. Patel and P. Thakkar, "Opinion Spam Detection Using Feature Selection," 2014 International Conference on Computational Intelligence and Communication Networks, Bhopal, 2014, pp. 560-564. doi: 10.1109/CICN.2014.127

[7] S. P. Algur and J. G. Biradar, "Review spamicity based on rank and content of the review," 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Davangere, 2015,pp.140-145. doi: 10.1109/ICATCCT.2015.7456871

[8] 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), Beijing, 2016, pp. 317-320. doi: 10.1109/ICISCE.2016.77

[9] Y. Lin, T. Zhu, H. Wu, J. Zhang, X. Wang and A. Zhou, "Towards online anti-opinion spam: Spotting fake reviews from the review sequence," 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), Beijing, 2014, pp. 261-264. doi: 10.1109/ASONAM.2014.6921594

[10] D. Runa, X. Zhang and Y. Zhai, "Try to Find Fake Reviews with Semantic and Relational Discovery," 2017 13th International Conference on Semantics, Knowledge and Grids (SKG), Beijing, 2017, pp. 234-239. doi: 10.1109/SKG.2017.00048

[11] Wahyuni, Eka & Djunaidy, Arif. (2016). Fake Review Detection From a Product Review Using Modified Method of Iterative Computation Framework. MATEC Web of Conferences. 58. 03003. 10.1051/matecconf/20165803003.

[12] 6 Easy Steps to Learn Naive Bayes Algorithm (2018, June 6) [Online] https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[13] The Random Forest Algorithm (2018, June 6) [Online] https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

[14] Model Evaluation – Classification (2018, June 5) [Online] http://www.saedsayad.com/model_evaluation_c.htm

[15] Classification Accuracy is Not Enough: More Performance Measures You Can Use (2018, June 5) [Online] https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/