ISSN (e): 2250 – 3005 || Volume, 05 || Issue, 05 || May – 2015 ||
International Journal of Computational Engineering Research (IJCER)

Fraud Detection in Online Reviews using Machine Learning Techniques

Kolli Shivagangadhar, Sagar H, Sohan Sathyan, Vanipriya C.H
Department of Information Science and Engineering
Sir M. Visvesvaraya Institute of Technology, Bengaluru

Abstract:
Online reviews play a very important role in today's e-commerce for decision-making. Large part of the population i.e. customers read reviews of products or stores before making the decision of what or from where to buy and whether to buy or not. As writing fake/fraudulent reviews comes with monetary gain, there has been a huge increase in deceptive opinion spam on online review websites. Basically fake review or fraudulent review or opinion spam is an untruthful review. Positive reviews of a target object may attract more customers and increase sales; negative review of a target object may lead to lesser demand and decrease in sales. These fake/fraudulent reviews are deliberately written to trick potential customers in order to promote/hype them or defame their reputations. Our work is aimed at identifying whether a review is fake or truthful one. Naïve Bayes Classifier, Logistic regression and Support Vector Machines are the classifiers used in our work.

Keywords: *Logistic Regression, Naïve Bayes Classifier (NBC), n-gram, Opinion Spam, Review Length, Supervised Learning, Support Vector Machine (SVM).*

I. Introduction

In the present scenario, customers are more dependent on making decisions to buy products either on e-commerce sites or offline retail stores. Since these reviews are game changers for success or failure in sales of a product, reviews are being manipulated for positive or negative opinions. Manipulated reviews can also be referred to as fake/fraudulent reviews or opinion spam or untruthful reviews. In today's digital world deceptive opinion spam has become a threat to both customers and companies. Distinguishing these fake reviews is an important and difficult task. These deceptive reviewers are often paid to write these reviews. As a result, it is a herculean task for an ordinary customer to differentiate fraudulent reviews from genuine ones, by looking at each review.

There have been serious allegations about multi-national companies that are indulging in defaming competitor's products in the same sector. A recent investigation conducted by Taiwan's Fair Trade Commission revealed that Samsung's Taiwan unit called Open tide had hired people to write online reviews against HTC and recommending Samsung smartphones. The people who wrote the reviews, foregrounded what they outlined as flaws in the HTC gadgets and restrained any negative features about Samsung products [12]. Recently e-commerce giant amazon.com had admitted that it had fake reviews on its site and sued three websites accusing them of providing fake reviews [13], stipulating that they stop the practice. Fakespot.com has taken a lead in detecting fake reviews of products listed on amazon.com and its subsidiary ecommerce sites by providing percentage of fake reviews and grade. Reviews and ratings can directly influence customer purchase decisions. They are substantial to the success of businesses. While positive reviews with good ratings can provide financial improvements, negative reviews can harm the reputation and cause economic loss. Fake reviews and ratings can defile a business. It can affect how others view or purchase a product or service. So it is critical to determine fake/ fraudulent reviews.

Traditional methods of data analysis have long been used to detect fake/fraudulent reviews. Early data analysis techniques were oriented toward extracting quantitative and statistical data characteristics. Some of these techniques facilitate useful data interpretations and can help to get better insights into the process behind data. To go beyond a traditional system, a data analysis system has to be equipped with considerable amount of background data, and be able to perform reasoning tasks involving that data. In effort to meet this goal researchers have turned to the fields of machine learning and artificial intelligence. A review can be classified as either fake or genuine either by using supervised and/or unsupervised learning techniques. These methods seek reviewer's profile, review data and activity of the reviewer on the Internet mostly using cookies by generating user profiles. Using either supervised or unsupervised method gives us only an indication of fraud probability.

No stand alone statistical analysis can assure that a particular review is fraudulent one. It can only indicate that this review is more likely to be suspicious. Detection and filtering of genuine reviews is an interesting problem for the researchers and e-commerce sites. One such review site that filters fake reviews is yelp.com. The filter used in yelp.com to hide fake reviews from public is a trade secret. In this work we try to analyze Yelp Academic Challenge Dataset [4] and determine whether a review is genuine or fake.

II. Related Work

A number of studies have been conducted which focused on spam detection in e-mail and on the web, however, only recently have any studies been conducted on opinion spam. Jindal and Liu (2008)[5] have worked on "Opinion Spam and Analysis" and have found that opinion spam is widespread and different in nature from either e-mail or Web spam. They have classified spam reviews into 3 types: Type 1, Type 2 and Type 3. Here Type 1 spam reviews are untruthful opinions that try to mislead readers or opinion mining systems by giving untruthful reviews to some target objects for their own gains. Type 2 spam reviews are brand only reviews, those that comment only on the brand and not on the products. Type 3 spam reviews are not actually reviews, they are mainly either advertisements or irrelevant reviews which do not contain any opinions about the target object or brand. Although humans detect this kind of opinion spam they need to be filtered, as it is a nuisance for the end user. Their investigation was based on 5.8 million reviews and 2.14 million reviewers (members who wrote at least one review) crawled from amazon.com and they have discovered that spam activities are widespread. They have regarded spam detection as a classification problem with two classes, spam and non-spam. And have built machine-learning models to classify a review as either spam or non-spam. They have detected type 2 and type 3 spam reviews by using supervised learning with manually labeled training examples and found that the highly effective model is logistic regression model. However, to detect type 1 opinion spam, they would have had to manually label training examples. Thus they had to use duplicate spam reviews as positive training examples and other reviews as negative examples to build a model.

In the paper "Finding Deceptive Opinion Spam by Any Stretch of the Imagination" by Ott, et al. 2011[10], they have given focus to the deceptive opinion spam i.e. the fictitious opinions which are deliberately written to sound authentic so as to deceive the user. The user cannot easily identify this kind of opinion spam. They have mined all 5-star truthful reviews for 20 most famous hotels in Chicago area from trip advisor and deceptive opinions were gathered for the same hotels using amazon mechanical trunk (AMT). They first asked human judges to evaluate the review and then they have automated the task for the same set of reviews, and they found that automated classifiers outperform humans for each metric. The task was viewed as standard text categorization task, psycholinguistic deceptive detection and genre identification. The performance from each approach was compared and they found that the psycholinguistic deceptive detection and genre identification approach was outperformed by n-gram based text categorization, but a combined classifier of n-gram and psychological deception features achieved nearly 90% cross-validated accuracy. Finally they came into a conclusion that detecting deceptive opinions is well beyond the capabilities of humans. Since then, various dimensions have been explored: detecting individual (Lim et al., 2010)[6] and group spammers (Mukherjee et al., 2012)[7], time-series (Xie et al., 2012)[8] and distributional analysis (Feng et al., 2012a)[9].

Yoo and Gretzel (2009)[15] gather 40 truthful and 42 deceptive hotel reviews and, using a standard statistical test, they have manually compared the psychologically relevant linguistic differences between them. In (Mukherjee, et al., 2013)[11], authors have briefly analyzed "What yelp filter might be doing?" by working with different combination of linguistic features like unigram, bigram, distribution of parts of speech tags and yielding detection accuracy. Authors have found that a combination of linguistic and behavioral features comparatively yielded more accuracy.

III. Dataset

The Yelp Challenge Dataset includes hotel and restaurant data from Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison, Edinburgh, Karlsruhe, Montreal and Waterloo. It consists of

- 61,000 businesses
- 481,000 business attributes
- 31,617 check-in sets
- 366,000 users
- 2.9 million social edges
- 500,000 tips
- 1.6 million reviews

The yelp academic challenge dataset that we have used in our work consists of 50075 reviews that are genuine. Fake reviews are crawled from yelp.com not recommended review section. These reviews are put under not recommended review section because these are classified as fake/ suspicious reviews. A sophisticated algorithm is used in yelp to filter these types of deceptive reviews.

IV. Architecture/ Implementation Model

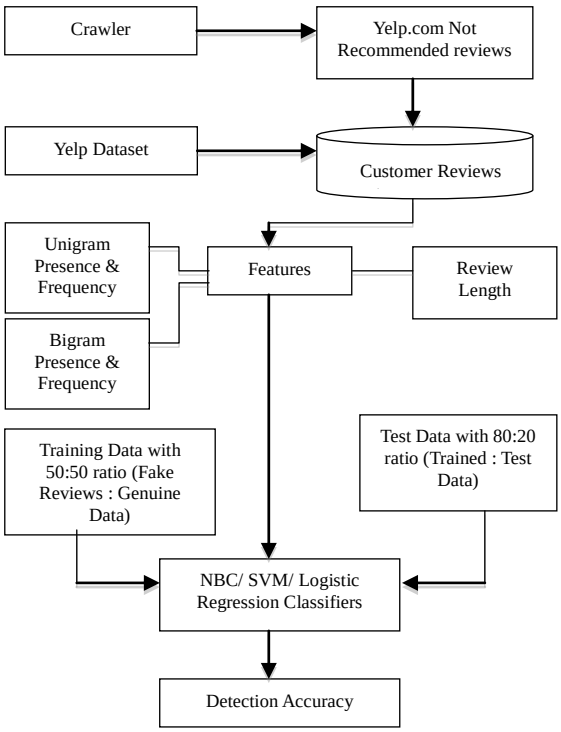

Fig 4.1 Architecture/ Implementation Model

The following are the steps involved in implementation of the model:
Step 1: Not-recommended reviews are extracted from yelp.com using crawlers. Text pre-processing is done to remove all the unwanted characters and find the reviews only. We consider these extracted reviews as suspicious or fake reviews. Number of fake/ suspicious reviews extracted are 8135.

Step 2: Genuine/ Truthful reviews are taken from Yelp academic challenge dataset. Since these reviews are cleaned, preprocessing is not required. Number of genuine reviews in the dataset considered for our work are 50075.

Step 3: We have used unigram presence, unigram frequency, bigram presence, bigram frequency and review length as features for our model. All these features are briefly explained in section 5 i.e. Feature construction.

Step 4: Training data obtained in the previous steps is used to train the Naïve Bayes Classifier, Support Vector Machines and Logistic Regression classifiers. Since the review dataset is not balanced, we consider only 8000 genuine or truthful reviews and 8000 fake/ suspicious reviews. This training data has a ratio of 50:50 i.e. it contains 50% of fake reviews and 50% of truthful reviews.

Step 5: Once the Naïve Bayes Classifier (NBC), Support Vector Machines (SVM) and Logistic Regression classifiers are trained separately for unigram, bigram and review length, it is now used to generate the detection accuracy. Now we input the test data. This test data has 80% of trained data and 20% of test data.

Step 6: Here our trained Naïve Bayes Classifier (NBC), Support Vector Machines (SVM) and Logistic Regression classifiers provide both test presence accuracy and test frequency accuracy.

V. Feature Construction

Each of the features discussed below are only for reviews of product/business

5.1 Review length (RL)

Review length is the average number of words present in a review [11]. Usually the length of fake review will be on the lesser side because of the following reasons

- Reviewer will not be having much knowledge about the product/business.
- Reviewer tries to achieve the objective with as few words as possible.

5.2 n-gram

An n-gram [2] is a contiguous sequence of n items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. These n-gram's typically are collected from a text or speech corpus. In this project we use unigram and bigram as important features for detection of fake reviews. Unigram is an n-gram of size 1 and Bigram is an n-gram of size 2.

5.2.1. Unigram Frequency: Unigram frequency is a feature that deals with number of times each word unigram has occurred in a particular review.

5.2.2. Unigram Presence: Unigram presence is a feature that mainly finds out if a particular word unigram is present in a review.

5.2.3. Bigram Frequency: Bigram frequency is a feature that deals with number of times each word bigram has occurred in a particular review.

5.2.4. Bigram Presence: Bigram presence is a feature that mainly finds out if a particular word bigram is present in a review.

VI. Results

Since the detection accuracy percentage varies with different sets of test reviews, we have used 5-fold cross validation technique by considering folds of trained dataset and test dataset in the ratio of 80:20. Test frequency accuracy obtained for unigram presence, unigram frequency, bigram presence, bigram frequency and review lengths are tabulated in table 6.1.1

Table 6.1.1 Detection Accuracy for Various Features

Classifiers/ Features	Logistic Regression	Support Vector Machines (SVM)	Naïve Bayes Classifier (NBC)
Unigram Presence	50.6	50.15	49.45
Unigram Frequency	49.7	48.675	49.95
Bigram Presence	50.4	49.95	50.025
Bigram Frequency	50.325	50.075	50.175
Review Length	50	50	

VII. Conclusion

Determining and classifying a review into a fake or truthful one is an important and challenging problem. In this paper, we have used linguistic features like unigram presence, unigram frequency, bigram presence, bigram frequency and review length to build a model and find fake reviews. After implementing the above model we have come to the conclusion that, detecting fake reviews requires both linguistic features and behavioral features.

VIII. Further Work

This paper concentrated much on how to detect fake reviews using supervised learning with linguistic features only. The same model can also be implemented with a combination of behavioral and linguistic features by using supervised, unsupervised or semi-supervised learning techniques.

IX. Acknowledgement

We sincerely thank Mrs. Premadurga Kolli for her expert advice and consistent support in guiding us through out the project.

References

[1] Wikipedia- Supervised Learning http://en.wikipedia.org/wiki/Supervised_learning
[2] Wikipedia- n-gram http://en.wikipedia.org/wiki/N-gram
[3] Wikipedia- SVM (Support Vector Machine) http://en.wikipedia.org/wiki/Support_vector_machine
[4] Yelp Challenge Dataset http://www.yelp.com/dataset_challenge
[5] "Opinion Spam and Analysis" by Nitin Jindal and Bing Liu. ACM-2008.
[6] Lim, E., Nguyen, V., Jindal, N., Liu, B. Lauw, H. 2010. Detecting product review spammers using rating behavior. CIKM.
[7] Mukherjee, A., Liu, B. and Glance, N. 2012. Spotting fake reviewer groups in consumer reviews. WWW.
[8] Xie, S., Wang, G., Lin, S., and Yu, P.S. 2012. Review spam detection via temporal pattern discovery. KDD.
[9] Distributional Footprints of Deceptive Product Reviews by Feng, S., Xing, L., Gogar, A., and Choi, Y. 2012a. ICWSM.
[10] Ott, Myle, et al. "Finding deceptive opinion spam by any stretch of the imagination." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies- Volume 1. Association for Computational Linguistics, 2011
[11] Mukherjee, et al. "What Yelp Fake Review Filter Might Be Doing?" ICWSM. 2013.
[12] "Samsung probed in Taiwan over fake web reviews" –BBC News http://www.bbc.com/news/technology-22166606
[13] "Amazon's Had Enough of Fake Reviews on Its Site, Files Lawsuit"- Yahoo Tech News https://www.yahoo.com/tech/amazons-had-enough-of-fake-reviews-on-its-site-116028350069.html
[14] "Comparison of deceptive and truthful reviews" by Yoo and Gretzel (2009).