A NOVEL SUPERVISED AND SEMI-SUPERVISED LEARNING BASED FAKE ONLINE REVIEWS DETECTION SYSTEM

O.V Chakradhar Reddy*, S. Ramesh Babu**
*P.G. Scholar (M. Tech), Dept. Of Computer Science Engineering, Srinivasa Institute of Technology and Science, KADAPA - 516002. Email Id: reddychakri1266@gmail.com
**HOD, Dept. Of Computer Science Engineering, Srinivasa Institute of Technology and Science, KADAPA - 516002. Email Id: babu.ramesh19@gmail.com

Abstract—Online reviews have great impact on today’s business and commerce. Decision making for purchase of online products mostly depends on reviews given by the users. Hence, opportunistic individuals or groups try to manipulate product reviews for their own interests. This paper introduces some semi-supervised and supervised text mining models to detect fake online reviews as well as compares the efficiency of both techniques on dataset containing hotel reviews.

1. INTRODUCTION

Technologies are changing rapidly. Old technologies are continuously being replaced by new and sophisticated ones. These new technologies are enabling people to have their work done efficiently. Such an evolution of technology is online marketplace. We can shop and make reservation using online websites. Almost, every one of us checks out reviews before purchasing some products or services. Hence, online reviews have become a great source of reputation for the companies. Also, they have large impact on advertisement and promotion of products and services. With the spread of online marketplace, fake online reviews are becoming great matter of concern. People can make false reviews for promotion of their own products that harms the actual users. Also, competitive companies can try to damage each other’s reputation by providing fake negative reviews.

Researchers have been studying about many approaches for detection of these fake online reviews. Some approaches are review content based and some are based on behavior of the user who is posting reviews. Content based study focuses on what is written on the review that is the text of the review where user behavior based method focuses on country, ip-address, number of posts of the reviewer etc. Most of the proposed approaches are supervised classification models. Few researchers, also have worked with semi-supervised models. Semi-supervised methods are being introduced for lack of reliable labeling of the reviews.

In this paper, we make some classification approaches for detecting fake online reviews, some of which are semi supervised and supervised and others are supervised. For semi-supervised learning, we use Expectation-maximization algorithm. Statistical Naive Bayes classifier and Support Vector Machines(SVM) are used as classifiers in our research work to improve the performance of classification. We have mainly focused on the content of the review based approaches. As feature we have used word frequency count, sentiment polarity and length of review.

In the present scenario, customers are more dependent on making decisions to buy products either on ecommerce 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. Recently ecommerce giant amazon.com had admitted that it had fake reviews on its site and sued three websites accusing them of providing fake reviews, 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.

2. LITERATURE REVIEW

In [4], Jindal, et al. claimed they are the first to attempt to study review spam and spam detection. They collected 2.14 million reviews from Amazon for their research work. They found a large number of duplicate and near-duplicate reviews written by the same reviewers on different products or by different reviewers on the same products or different products. They proposed to perform spam detection based on duplicate finding and classification. They used logistic regression to learn a predictive model. Using 10-fold cross-validation on the data they got average area under the ROC curve (AUC) value of 78%.

In [6], Ott, et al. showed that psychological studies of deception and genre identification are both out-performed at statistically significant levels by n-gram based text categorization techniques. Notably, a combined classifier with both n-gram and psychological deception features achieves nearly 90% accuracy.

In [9], Ott, et al. worked on negative deceptive opinion spam which usually are reviews that aim at degrading other company’s reputations. They found that standard n-gram text categorization techniques can detect negative deceptive opinion spam with performance far surpassing that of human judges.

In [10], Sandulescu, et al. used one time reviewers such as a reviewer who leaves only one review. They exploited the singleton reviewers review. They tackled the problem of detect- ing fake reviews written by the same person using multiple names, posting each review under a different name. They propose two methods to detect similar reviews and show the results gener- ally outperform the vectorial similarity measures used in previous work. Their proposed methods are the semantic similarity between words to the review level and based on topic modeling and exploit the similarity of the reviews topic distributions using two models: bag-of-words (a simplifying representation used in natural language processing and information retrieval) and bag-of-opinion (a simplifying representation used in natural language processing and opinion mining) phrases.

In [5], Li, et al. manually labeled nearly 6000 reviews. They collected a dataset from the Epinions website. They employed ten college students for tagging all the reviews. Students were first instructed to read books and articles about how spam review looks like then they were asked to label those reviews. They first used supervised learning algorithm and analyze the effectiveness of different features in review spam identification. They also used a two-view semi-supervised methodology to exploit a large amount of unlabeled data. The experiment results show that two-view co-training algorithms can achieve better results than the single-view algorithm.

In [11], Luca, et al. worked on restaurant reviews that are identified by Yelp’s filtering algorithm as suspicious, or fake. They found that nearly one out of five reviews is marked as fake by Yelp’s Algorithm. These reviews tend to be more extreme than other reviews and are written by reviewers with less established reputations. Moreover, their finding suggests that economic in- centives factor heavily into the decision to commit fraud. Organizations are more likely to game the system when they are facing increased competition and when they have poor or less estab- lished reputations.

In [12], Wahyuni, et al. aimed to detect fake reviews for a product by using the text and rating property from a review. Their proposed system measures the honesty value of a review, the trustiness value of the reviewer and the reliability value of a product.

In [13], Jindal, et al. deal with identifying unusual review patterns which can represent suspicious behaviors of reviewers. They formulate the problem as finding unexpected rules. They analyzed an Amazon.com review dataset and found many unexpected rules and rule groups which indicate spam activities.

In [14], Lim, et al. recognize spammers based on behaviors of reviewers that deviate from usual practice. These reviewers are highly suspicious of review manipulation. Their research sug- gests that one should focus on detecting spammers based on their spamming behaviors, instead of identifying spam reviews. Their proposed review spammer detecting approach is user-centric, and user behavior-driven. They claimed their proposed methods generally outperform the base-line method based on helpfulness votes.

In [15], Mukherjee et al. dove down to Yelp’s secret filtering algorithm. They put a few existing research methods to the test and evaluated performance on the real-life Yelp data. They found the behavioral features perform very well, but the linguistic features are not as effective. Their analysis and experimental results shows that Yelp’s filtering is reasonable and its filtering algorithm seems to be correlated with abnormal spamming behaviors.

In [16], Li, et al. claimed they are the first one to present a large-scale analysis of re- staurant reviews. They were able to collect a large amount of data from Dianping which is a Chinese group buying website for locally found
food delivery services, consumer products and retail services. Dianping helped them to get user reviews about restaurants and, users IP addresses and profiles. They used a method called Positive-Unlabeled Learning. They used temporal and spatial features at various levels (reviews, users, IPs) for supervised opinion spam detection.

In [17], Xie, et al. developed a model for singleton spam review detection problem based on the observation that the arrival pattern of singleton review tends to be bursty and temporally correlated to the rating.

In [18], Li, et al. worked on detecting spamming network using reviewer posting frequency within short periods of times and also considered other users posting frequency within that short period of time for the same products. They primarily tried to find out individual spammers and spammer groups.

In [19], KC, et al. worked on the temporal dynamics of opinion spamming. They looked to find out if there are any specific spamming policies that spammers employ. They used a large set of reviews from Yelp restaurants and its filtered reviews to characterize the way opinion spamming operates in a commercial setting. Using time-series analysis, they found that there exist three dominant spamming policies: early, mid and late across the various restaurant. Their analysis showed that the deception rating time-series for each restaurant had statistically significant correlations with the dynamics of truthful rating time-series indicating that spam injection may potentially be coordinated by the restaurants/spammers to counter the effect of unfavorable rating over time.

In [20], Shebuti and Akoglu proposed a framework named Speagle that exploits both re- lational data (user-review-product graph) and metadata (behavioral and text data) collectively to detect suspicious users and reviews, as well as products targeted by spam. Their main contribu- tion is to employ a review-network-based classification task which accepts prior knowledge on the class distribution of the nodes, estimated from metadata. Their proposed framework works in an unsupervised fashion, but can easily leverage labels.

3. PROPOSED WORK

As briefly introduced in Section II, many and different are the features that have been considered so far in the review site context to identify fake reviews. In some cases, features belonging to different classes have been considered separately by distinct approaches. In other cases, the employed features constitute a subset of the entire set of features that could be taken into account; furthermore, new additional features can be proposed and analyzed to tackle open issues not yet considered, for example the detection of singleton fake reviews. For these reasons, in this section we provide a global overview of the various features that can be employed to detect fake reviews. Both significant features taken from the literature and new features proposed in this article are considered. Since the most effective approaches discussed in the literature are in general supervised and consider review- and reviewer-centric features, these two classes will be presented in the following sections. The choices behind the selection of the features belonging to the above mentioned classes will be detailed along each section. When the features are taken from the literature, they will be directly referred to the original paper where they have been initially proposed. The absence of the reference will denote those features that have been widely used by almost every proposed technique.

Finally, the presence of the label denoted by [new] will indicate a feature proposed for the first time in this article.

A. Review centric Features

The first class of features that have been considered, is constituted by those related to a review. They can be extracted both from the text constituting the review, i.e., textual features, and from meta-data connected to a review, i.e., metadata features. In every review site, the time information regarding the publication of the review, and the rating (within some numerical interval) about the reviewed business are metadata, are always provided. In addition, in relation to metadata features, those connected to the cardinality of the reviews written by a given user must be carefully studied. In fact, a large part of reviews are singletons, i.e., there is only one review written by a given reviewer in a certain period of time (this means that in the user account there is only one review at the time of the analysis). For this kind of reviews, specific features must be designed. In fact, as it will be illustrated in the following, many of the features that have been proposed in the literature are based on some statistics over several reviews written by the same reviewer. In the case of singletons, these features loose their relevance in assessing credibility.

Therefore, the definition of suitable features that are effective for detecting also singleton fake reviews becomes crucial.

1) Textual Features: as briefly illustrated in Section II, it is practically impossible to distinguish between fake and genuine reviews by only reading their content. The analysis provided by Mukherjee et al. in [19] has shown that the KL-divergence between the languages employed by spammers and non spammers in Yelp is very subtle. However, the good results obtained in [26] by using linguistic features on a domain specific dataset (i.e., a Yelp’s dataset containing only New York japanese restaurants), show that at least on a domain specific level, textual features can be useful. It is possible to use Natural Language Processing techniques to extract simple features from the text, and to use as features some statistics and some sentiment estimations connected to the use of the words.

• Text: several approaches employ as textual features both unigrams and bigrams extracted from the text of reviews, as illustrated in Section II.

• Text statistics: several statistics on the review content have been proposed as features by Li et al. in
[21]: – Number of words, i.e., the length of the review in terms of words; – Ratio of capital letters, i.e., the number of words containing capital letters with respect to the total number of words in the review; – Ratio of capital words, i.e., considering the words where all the letters are uppercase; – Ratio of first person pronouns, e.g., ‘I’, ‘mine’, ‘my’, etc.; – Ratio of ‘exclamation’ sentences, i.e., ending with the symbol ‘!’.

• Sentiment evaluations: – Subjectivity, i.e., a number representing the proportion of subjective words (expressing sentiment, judgment) as opposed to objective (descriptive) words.

2) Meta-data Features: these kinds of features are extracted from the meta-data connected to reviews, or they can be generated by reasoning on the reviews’ cardinality with respect to the reviewer and the entity reviewed.

• Basic features: – Rating, i.e., the rating attributed in the review to the entity, in the form of some numerical value belonging to a given interval (e.g., 1-5 ‘stars’); – Rating deviation [27], i.e., the deviation of the evaluation provided in the review with respect to the entity’s average rating; – Singleton [25], i.e., it indicates the fact that the review is the only one provided by a reviewer in a given period of time (e.g., a day). These basic features rely on some simple and intuitive heuristics. A fake review tends to contain a more ‘extreme’ rating with respect to genuine reviews, thus implying that the rating deviation from the entity’s average rating is higher; furthermore, a singleton review provided by a user could indicate that s/he is not particularly involved in the review site community, which constitutes a possible indication of unreliability.

• Burst features: it is said that reviews for an entity are ‘bursty’ when there is a sudden concentration of reviews in a time period. These review bursts can be either due to sudden popularity of the entities reviewed or to spam attacks. Since it has been proven that reviews in the same burst tend to have the same nature [28], it is possible to easily identify groups of fake reviews by analyzing the nature of the burst. Two burst detection studies have been described in [27], [28]. Taking inspiration from the just cited works, in this paper several features considering burstiness have been introduced. These features are related to the time window in which a review has been posted, relatively to a given reviewed entity. Basically, a review is more likely to be fake if it is posted on a day when the number of reviews is abnormally high, and when the average rating associated with an entity in a review (in a specific time window) varies significantly with respect to the entity’s average rating (in general it decreases, for example passing from 3.5/5 to 2/5). The patterns of externalities sketched concisely in Figure: 3.1 below:

![Figure: 3.1 Proposed Block Diagram](image)

4. RESULTS AND ANALYSIS

We have applied our experiments on a machine with Processor: Intel (R) Core (TM) i5-4200U and CPU - 1.6GHz, RAM: 6 GB, System type: 64 bit OS, x64- based processor, Hard Disk: 1 TB. We have used Linux(Ubuntu 16.04) as our operating system.

We have used Expectation maximization (EM) algorithm for semi-supervised classification. As classifier we have used Support Vector machines (SVM) and Naive Bayes classifier. We have divided our dataset into a train test ratio of 75:25 and 80:20 for each classification process. For semi-supervised classification with SVM classifier, we have found an accuracy of 81.34% for 80:20 split ratio and 80.47% for 75:25 split ratio with gamma equal 0.3 and 0.6 respectively. For semi-supervised classification with Naive Bayes classifier, we have got an accuracy of 85.21% and 84.87% respectively for split ratio of 80:20 and 75:25. Jiten et al. [8] using semi-supervised classification with EM and Positively Unlabeled learning respectively, got highest accuracy of 83.00% and 83.75% for train test ratio of 80:20. They have tried Logistic regression,
K-nearest neighbor, Stochastic Gradient Descent and Random Forest as classifier. We have also tried supervised classification techniques to find out performance of them for our dataset. We have used Naive Bayes and SVM classifiers.

For SVM classifier we have tuned gamma parameter keeping C parameter constant for having a better fit of the model. The results are shown in the following figure 4. For supervised classification with SVM classifier, we have found an accuracy of 82.28% for 80:20 split ratio and 82.04% for 75:25 split ratio with gamma equal 0.1 and 0.8 respectively. For supervised classification with Naive Bayes classifier we have got the highest accuracy of 86.32% and 86.21% respectively for split ratio of 80:20 and 75:25.

![Fig 4.1 Review Analysis of Proposed System](image)

5. CONCLUSION

We have shown several semi-supervised and supervised text mining techniques for detecting fake online reviews in this research. We have combined features from several research works to create a better feature set. Also we have tried some other classifier that were not used on the previous work. Thus, we have been able to increase the accuracy of previous semi-supervised techniques done by Jiten et al. We have also found out that supervised Naive Bayes classifier gives the highest accuracy. This ensures that our dataset is labeled well as we know semi-supervised model works well when reliable labeling is not available.

In our research work we have worked on just user reviews. In future, user behaviors can be combined with texts to construct a better model for classification. Advanced preprocessing tools for tokenization can be used to make the dataset more precise. Evaluation of the effectiveness of the proposed methodology can be done for a larger data set. This research work is being done only for English reviews. It can be done for Bangla and several other languages.

REFERENCES


