Finding Valuable Yelp Comments by Personality, Content, Geo, and Anomaly Analysis

Jay Koven  
NYU Polytechnic School of Engineering  
Brooklyn, NY 11201  
Email: jkoven@nyu.edu

Hossein Siadati  
NYU Polytechnic School of Engineering  
Brooklyn, NY 11201  
Email: hossein@nyu.edu

Ching-Yung Lin  
IBM Watson Research Center  
Yorktown Heights, NY USA  
Email: chingyung@us.ibm.com

Abstract—User reported experiences and opinions are used by peers to make decisions about where to go and what to buy. Unfortunately, not all users or opinions are honest. Many opinions are fabricated and may be submitted by automated systems or by people who are recruited by businesses and search engine optimizers to write good reviews. Such reviews and ratings are called spam reviews. These are misleading for users and troublesome for honest businesses. While most current efforts to tackle this problem are focused on spam review detection, in this paper we focused on detecting authentic and valuable reviews for a front-end application that reorders the reviews. In this manner, we have identified several features based upon the content of the reviews as well as identifying behavioral features of reviewers to pinpoint useful reviews with 80% accuracy.

Keywords—Valuable Review Detection, Machine Learning, Security, Usability

I. MOTIVATION

This demo paper presents a new technological advance in applying data mining techniques for detecting valuable reviews in review websites. Consumers are increasingly turning to online reviews as a guide to decide among numerous options available for dining, shopping, traveling, and other activities. At the same time businesses and service providers are finding it increasingly important to receive positive recommendations from users to present a positive image to new customers. Luca [1] showed that a one-star increase in Yelp review ratings increases revenue by 5-9 percent.

Unfortunately, a number of businesses pay search engine optimizers (SEOs) to improve their ratings by either writing fake reviews for them, or by recruiting others to do so. Businesses will ask users to like their page or write positive reviews in exchange of extra services such as WiFi connection or they will offer financial incentives in the form of coupons or discounts to incentivize legitimate users to write good reviews. Being aware of these misuses, review websites do their best to lower the effect of such reviews by detecting and discouraging them. Unfortunately, reviews of questionable quality or those that are outright fraudulent are still finding their way onto legitimate websites.

Users want to get a complete and correct understanding of a business by reading online reviews. They also desire to make a decision quickly by reading a limited number of reviews. However, due to the large numbers of reviews and the existence of spam reviews, efficient and accurate decision making can only happen if users have a way of distinguishing accurate and useful reviews. By reading the most valuable reviews which contain dense and important points, users can more reliably make decisions. In addition, the effect of spam reviews will be automatically canceled when they are not read by potential customers.

In this paper, we focus on detecting valuable opinions. We present this information to the users by developing an extension which would overlay an existing review website. The focus of our initial design is Yelp, which is one of the most used review websites. We provide a method to evaluate Yelp reviews and classify them into two categories, valuable and valueless, where valuable reviews represent reviews that consumers would find to be both useful and accurate and valueless reviews that might be either inaccurate or empty of useful content. We use both direct and indirect features in our classifier. The direct attributes are those that can be found directly from reviews such as a reviewers average star rating, the number of reviews and simple statistics of words which are used in the review. The indirect attributes are those that can be derived from analyzing information in the dataset. These include reviewer personality traits, interpersonal relationships and geographical relationships of the reviews and reviewers.

The rest of the paper is organized as follows: Chapter II gives background and lists the related works. Analysis on dataset and the features we extracted are discussed in III. Classifier and results are reported in IV.

II. RELATED WORKS

A. Content based detection

There are a number of works on how to detect spam opinions in general and Yelp in particular. One approach is to analyze the content of the reviews and detect the fake reviews based on computational linguistic analysis. Ott et. al [2] used genre identification, and POS (part of
speech) as linguistic cues to classify reviews. They also used the LIWC (Linguistic Inquiry and Word Count) to extract more features from the reviews to make a stronger classifier. The idea behind this approach is that fake reviews have content with characteristics which differentiate them from normal reviews. Some of these characteristics appear when the reviewers have not been in the places they are reviewing. A classifier on a corpus of fake reviews written by MTurkers has achieved accuracy of 90% when classifying fake reviews.

Even Though this approach is accurate in the case of distinguishing fake reviews written by MTurkers, it is only 68% accurate when applied to fake reviews in general. The reason is that these classifiers take advantage of linguistic cues that appeared in the reviews from people who had not been in the place they reviewed. While this covers some portion of fake reviews, it can not be used to detect all fake reviews [3]. Some fake reviews are modified versions of cloned reviews, maximum content similarity (MCS) [2] is used to find out reviews of this type that maybe automatically generated using a core review that has minor modifications or are created by a lazy reviewer using cut and paste to create new reviews.

B. Behavioral detection

Even though content based approaches are useful in detecting some types of fake reviews, they are not very successful with real data [4]. Instead, features extracted based on behaviors of the users have shown to be more promising. One of the features used in Mukherjee et. al. is maximum number of Reviews (MNR) for each user based on the fact that a normal reviewer does not submit more than a few reviews in one day. Percentage of positive reviews (PR) is used to detect those false accounts which submit only positive reviews. While review deviation (RD) is used to identify reviews which deviate dramatically in comparison with other reviews for the same venue.

III. DESIGN

The architecture of our system is given in Figure 1. The analyzer receives raw data from Yelp and extracts direct and indirect information and features from them, which we will describe in chapter III-B. The classifier marks the reviews as valuable and valueless. The extension then injects HTML elements into the web page of the business and modifies the order of presentation of reviews to improve the visibility of valuable reviews to users.

A. Dataset and Classes

To develop our classifier we used the the Yelp challenge dataset which contains over 320,000 reviews from over 250,000 reviewers. We decided to use the “useful” field that is included with each review to define our classes of valuable and valueless. If the useful count for a review was 4 or more we classified the review as useful. This means that 4 different users who read the review clicked on the useful button for the review indicating that they found it useful.

B. Feature Extraction and Selection

In order to classify Yelp reviews into valuable and valueless classes, we needed a set of attributes which determine the quality of a review. These attributes fall into two categories, direct and indirect. The direct attributes are those that can be found directly from the dataset or with minor processing such as a reviewers average star rating, number of reviews and the number of words and sentences in a review. The indirect attributes are the ones that can be derived by further analysis of the information in the dataset. These include reviewer personality traits, interpersonal relationships and geographical relationships of the reviews. We used both direct and indirect features in our classification.

Direct attributes we have used in our classifier are a reviewer’s average star rating, which is the average of all the ratings of an individual reviewer’s reviews and the star rating of the review. We believed that the remaining information in the dataset would be best used to create the indirect attributes.

To create indirect attributes for our work we made the decision to focus on three areas: the topic and personality analysis, the geographical distribution of reviews by an individual reviewer, and the relationships between reviewers. The boundaries between these areas are not distinct and some of our attributes combined information from overlapping areas.

1) LDAT: LDA Topics: Topic modeling is a technique in Natural Language Processing which tries to detect topics related to a set of texts. In our case the texts are represented by the reviews. The extent and strength of a topic in a review represents the value of a topic in a review in relation to the value of the topic in the set of reviews.

We used LDA topic classification which is an unsupervised machine learning method which creates a set of topics in a dataset based upon the frequency of various words found in each document. The algorithm requires that a user specify the number of topics to create but the rest of the analysis is automatic. The result is a vector of topics comprised of
words that define those topics. A weight is then given to each of the words. For each review we computed topics associated with it. LDA generates a vector of weights of topics for each review. For example, the LDA vector for a review might consist of Topic 1 with a weight of 0.5, Topic 20 with a weight of 0.1, and Topic 38 with a weight of 0.05. We decided to use 50 topics for our analysis.

2) PSA: Personality Analysis: Personality of the reviewers is an indicator of the amount of the information they are willing to share with others, and hence, the usefulness of their reviews. For example, people who are more open are willing to share more details of their experience with others, making their reviews more useful. We have considered 5 major personality traits in our attribute list. These personalities include Extraversion, Emotional stability, Agreeableness, Conscientiousness and Openness to experience.

For the personality analysis we separated and consolidated the reviews by reviewer into individual files. The MIT Personality Recognizer Tool (PRT) [5] is used to analyze and create a personality profile for each reviewer. The PRT tool uses the LIWC[5] dictionary to identify various personality information of the reviewer based on the choice of words used in the reviews. It then builds a personality profile based on the big five personality traits [5].

C. RS: Rating Similarity

Birds of a feather find each-others ideas more useful. Similarly, the users who give the same or similar rating to a business find each-others’ reviews more acceptable and valuable. For each user we counted the number of people who gave the same or similar (±1) rating to the venue she/he reviews. This was an implicit measure of how useful the review of each user was. Based on this, for each user, we considered six new features. Five of these features indicate the number people who gave exactly the same star rating as the user being evaluated gave(1-star to 5-stars). The sixth feature considers the number of people who voted similarly to this user, using either the same rating, or one above or below.

1) RGC: Review Geographical Clusters: Legitimate users have patterns which unify them and distinguishable from made up accounts and fake review writers. For example, it is expected that false users often write reviews for different venues in scattered geographical areas, whereas authentic reviews are written by users focused in areas clustered around their workplace or home. One possible way of modeling this is to consider the geographical location of the business which the user reviewed as a blob on a map. These blobs form a map of points. From the map of blobs we extract the following features for each user; number of total blobs, number of clusters created by their blobs, and their minimum and maximum cluster sizes.

2) RC: Review Co-authorships: Connection between reviewers reveal relative behavior of each reviewer. We define a yelp friendship relationship as the relation between two people who reviewed the same business. The number of places that two users write a review for determines the strength of their friendship (We can imagine that they have been together to these places). The number of friends for a user is computed as the sum of the number of friends in different venues. We consider number of friends, the maximum number of friends, and minimum number of friends as other features in our classifier.

3) UNQS: Uniqueness Score: We define a review uniqueness score as the combination of a number of reviews by a user multiplied by the inverse of the number of reviews written for that place. This number indicates the value of a review based upon the reviewer’s reputation and reputation of the venue. Reviews given by reputed reviewers written for small businesses with few reviews are of more importance then reviews written for developed businesses. We considered this uniqueness score as one feature for our clustering algorithm.

D. User Interface and Important Phrases

We locate important and informative phrases within reviews by using a combination of an named entity detection and a modified TF-IDF algorithm. Figure 2 shows a snapshot of how our extension adds information to reviews in Yelp. This information includes usefulness rate of the review, the personality of the reviewer, and information about the number of geo-blocks they visit. We also spot the most important phrases inside a review.

IV. Evaluation

A. Classifier

For the classification part of our research we chose to use WEKA, which is a Machine Learning workbench from the Machine Learning Group at the University of Waikato in New Zealand. WEKA incorporates a large number of known machine learning techniques and allows the efficient analysis of large amounts of data in a relatively short
time. We chose to evaluate the data using five different techniques. The techniques we chose were Decision Trees (J48 and Random Forest), Multi Level Perceptrons (MLP), Support Vector Machines (SVO in WEKA), Naive Bayes and Multiboost which is a variation on the AdaBoost algorithm.

B. Evaluation

We started our evaluations by running each of the machine learning techniques with their default settings on a test set of 100,000 reviews with the full attribute vectors. The initial results were promising. We achieved an accuracy rate ranging from the high 50% range to the low 70% range depending on the algorithm. We found that changing the various settings for the algorithms did not produce significantly better results than the default settings. Naive Bayes and SVO produced the least promising results and the J48 decision tree algorithm produced the best results. In addition to producing accurate results the J48 Algorithm produces a visual representation of the decision tree that shows decision attributes and (indirectly) the Information Gain (IG) produced by those attributes.

Our next step was to use the combination of the IG information derived from the J48 technique and the WEKA select attributes feature to start reducing the number of attributes in our dataset as a method of increasing accuracy. Through an iterative process the attribute vector for each review was reduced to the following nine attributes in addition to the review Class: extraversion, conscientiousness, openness to experience, number of Friends, maximum number of friends in venue, close stars, number of clusters, number of blobs, maximum cluster size.

Interestingly, we found that the LDA topic vectors and the personality traits could be substituted for one another to achieve the same accuracy. This should not have been unexpected since the LDA algorithm is attempting to find hidden information in the reviews, which can match the same combinations of words associated with the emotional analysis. We decided to use the emotions as the attributes because they seem more intuitive than the un-named LDA topics represented by a collection of words and weights. Several other attribute combinations also showed similar overlap. For our final results we chose the shortest list of attributes that produced the most accuracy.

At each stage of reducing the attribute list we reevaluated the various machine learning techniques listed above. The J48 algorithm consistently produced the most accuracy. In addition the decision tree produced by the algorithm gave us a road map for evaluating the relevance of the attributes left in the dataset.

C. Results

Using the reduced attribute list and the J48 algorithm we were able to obtain an accuracy of 79.2 percent using the Weka generated classifier. We believe that this accuracy is a good start that can be improved with some additional evaluation of attributes and the possible addition of a few new attributes not considered in this initial research.

In addition to the accuracy that we achieved we also were very successful in identifying bad reviews. The false positive rate for the bad reviews, represented by the false positive rate for the valueless class in Weka is .051 which means that we only classified 5% of the reviews that should have been classified as bad incorrectly. We believe that this is a result that validates our methods.

We were less successful in our recall rate of the good reviews. We were only able to identify about 33% of the reviews that were classified as good. A manual survey of the reviews classified as good showed good classification, again reinforcing the validity of our methods. However, addition analysis of the remaining reviews that were classed as successful by our ground truth choice is necessary to understand what additional features we should be evaluating in order to improve our results.

V. STATE OF THE PROJECT

In this work, we have given a framework for our Review Tester which recognizes valuable reviews. At the user interface level, it elevates useful reviews and helps users to get more insight by highlighting important phrases in the reviews. We have built the classifier and it works as described. We have designed the extension but it is not fully implemented. We plan to implement the UI using chrome extension. We have future plans to improve our classifier by adding additional Natural Language Processing features to better identify deceptive reviews. In addition we plan to do a usability analysis based on our new extension.

REFERENCES