Reliable Reviews Recommendation

Jiaqi Chen
Statistics Department
Columbia University
jc4260@columbia.edu

Chen Qian
Statistics Department
Columbia University
cq2171@columbia.edu

Tianhe Shen
Electrical Engineering Department
Columbia University
ts2957@columbia.edu

Abstract—**The objective of this project is to recommend reliable reviews to users on the Yelp Website. The approach combines the fake review filtration with classification prediction on review-usefulness. The algorithms mentioned in this paper include similarity analysis, logistic regression, Naïve Bayes, random forest, support vector machine. Finally, we reached a reasonable dataset of useful reviews might be recommended to users.**

Keywords--Similarity; Weight Coefficient; Classification; Predict usefulness;

I. INTRODUCTION

E-commerce nowadays is blooming in various businesses, like Amazon in shopping, Yelp in dining, and TripAdvisor in travelling. As the value of social networking grows, a new wave of opportunities has emerged by integrating social media and commerce. Organizations are learning how to embrace social media technologies, create niche communities and combine e-commerce to fully monetize their online initiatives. Customers now don't shy away from expressing how much they want, love or dislike a product or service online. Further, ratings and reviews from other customers have become the most common method for e-commerce users to know more about the products or the merchants. People make decisions based on these ratings and reviews. However, since the potential interest that the ratings and reviews could bring to the vendor and the anonymity of the ratings and reviews, there are some fake and spamming ratings and reviews occur in the e-commerce. Some businesses employ people to write fake reviews about themselves or about a competing product or business to drive the sales towards themselves and gain an unfair advantage by perpetrating false information and deliberately misleading the consumers. Even though e-commerce websites, such as Amazon and Yelp, have their own algorithm to detect some spamming reviews, it is still a challenge to provide the reliable and useful ratings and reviews to users.

In this paper, we proposed a mechanism to filter out the potential unreliable ratings and reviews, and further, create recommendation models to provide the most reliable and useful ratings and reviews to users. We work on the Yelp dataset and focus on the dining business. We study and conclude the unreliable and spamming ratings and reviews behavior patterns, and then, implement two algorithms to detect these potential unreliable reviews. Once filtering out these potential unreliable reviews, some classification models could be able to be established and recommend the relatively useful reviews in the dataset to users.

The rest of paper is organized as follow: Chapter II gives background and lists the related works. Chapter III describes our system design and Chapter IV shows algorithm and tools than we have used. Software package will be described in Chapter V. Experiment results shows in Chapter VI. Finally, Chapter VII will conclude the paper.

II. RELATED WORKS

The challenge of fake and spamming reviews was first studied in 2008 [1] and the interest in the study has peaked ever since. Several different problems related to opinion spamming have been studied ranging from (but not limited to) individual opinion spammers to spamming groups [2]. There have been studies in which the timeline of the reviews has been analyzed and inferences drawn [3]. Up to now, there have been two main successful approaches in the past to detect the fake reviews. The first method is using Amazon Mechanical Turk to collect fake reviews. The second way is to segregate the fake reviewers into identified groups.

Ott et. al [4] used genre identification and this method is easy to implement. Nevertheless, this technique cannot mimic the actual climate of opinion spamming. Incentive-based fake reviews are different from crowd-sourced fake reviews, where random people merely write easy-to-spot fake reviews as a result. Second method, which is proposed in [2], particularly showed that the spammer groups can be far more damaging to a business than an individual spammer, and thus can take a quasi control over the sentiment towards a particular product. It shows that is difficult to use both the content based detection and segregation and behavior based detection because any member can choose not to behave abnormally any time. It also indicates that the reviewer groups have a few traits in common like the commonality of the product and the time window in which it is reviewed.

III. SYSTEM OVERVIEW

A flowchart of how our system operates is provided as following Figure 1. Firstly, we prepare out raw Yelp dataset into csv. Then we implement two algorithms on the prepared dataset using R. The algorithms we applied here will be shown in the Chapter IV. Then, we filter out the potential useless reviews. Next, we create classification

models to find out the useful reviews that we can recommend to users. The output useful and reliable reviews will be shown in the Chapter VI.

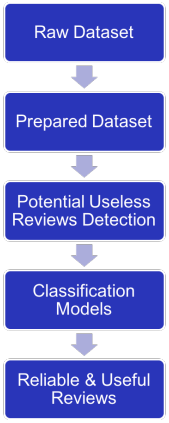

Figure 1. System flowchart

IV. ALGORITHM

A. Fake Review Filter

From the previous studies of the fake reviews and the behavior patterns of those unreliable ratings and reviews, we find that those kind of reviews share some extend of similarity. Generally speaking, they normally give more eccentric ratings to the merchants compared with most of other reviewers. That is because that the purpose of fake reviews is to drive the sales towards specific merchants and gain an unfair advantage, either through giving top ratings and good reviews to these specific merchants themselves, or by making extremely low ratings and bad reviews to the competitors of these merchants. Besides, the useless and unreliable reviews always have a very low review votes, which are the "Funny", "Useful", and "Cool" on Yelp. It may be because other users who have read these eccentric reviews don't think this review or rating is helpful. Further, we also find that the eccentric reviews sometimes have particularly higher review votes than normal reviews. This is common sense when people find someone says the truth form the flooding good reviews. However, those fake and spamming reviews are eccentric and own low review votes, since people don't think their "eccentric" is true and useful. Thus, through these two behavior patterns, we implement the similarity algorithm and weight coefficient function to detect the potential unreliable and useless ratings and reviews.

Similarity Algorithm

We use the similarity algorithm to describe the eccentric behavior we mentioned above. Let $r_1, r_2, \ldots, r_n$ be the ratings from reviewer n, ranged from 1 to 5. The number of total reviewers is N. The similarity score of each reviewer is,

$$R_{score} = 1 - \frac{\sum_{i=0,k=0,i\neq k}^{N}(r_i - r_k)/5}{N}$$

The similarity score thus is ranged from 0 to 1. The lower the score is means the more eccentric the rating and the review is.

Weight coefficient algorithm

After we successfully describe the eccentric behavior of the fake and spamming reviews, we use the weight coefficient algorithm to measure the opinions on the ratings and reviews from other users. Suppose f is the number of "Funny" review votes the specific rating and review owns. Similarly, u is the "Useful" review votes number and c is the "Cool" review votes number. Now, we further assume the weight coefficient for each kind of review vote is A_f, A_u, and A_c, respectively. Thus, we can calculate the other users' agreements score as following.

$$V_{score} = (A_f * f + A_u * u + A_c * c) / (f+u+c)$$

This agreement score shows the other users opinion on the specific rating and review. The lower the score is represents the less agreement from other users. We will show how to set the reasonable weight coefficient for each kind of review votes in Chapter VI.

Once we calculate the two score, we combine these two measurements through setting threshold for both of the scores. And then, we filter out the ratings and reviews which have both the low similarity score and agreement score.

B. Predicting Models

In our research, we want to recommend the useful reviews to users. Previously we have already removed the fake reviews from our dataset, the next thing we want to do is to predict whether the review is useful or not by variety of classification models.

Feature Selection

- *Useful reviews manually label*

In our dataset there is no label that indicates whether the review is useful or not, therefore we decide to manually label the dataset by using the following method:

--Firstly we added up each reviews FUC rating.

--Secondly we grouped the reviews by years in order to remove the influence of an increasing in FUC with time.

--Thirdly we ranked the total FUC rating in each year group and label the top 20% total FUC as useful.

- *Independent variable selections*

In the yelp dataset, we choose review length, similarity between reviews for same restaurant, user's existence length, user's useful votes, user's count of reviews, user's fans, and user's average stars as our main attributes to construct the model. The following table explains the features in details.

Independent variable	*Description*	*Data from*

Review length	The number of characters in the review	Review dataset
Similarity	A coefficient measures the similarity between this review and others	Calculation
User's votes	Number of useful votes voted by a user.	User dataset
User's existence length	Length of user's membership by month.	User dataset
User's count of reviews	The number of reviews the users has written	User dataset
User's fans	The number f fans that the user have	User dataset
User's average stars	the average star that the user gives	User dataset

Classification Model

- *Random forest*

Random forest is a widely used model for classification and regression. Growing trees and making them voting for the most popular class results in significant improvements in classification accuracy. Often random splits are generated by using 'Bagging' method and govern the trees to grow in the ensembles.

Algorithm:

For b= 1…B:

1. Draw a sample N_b of size n from training data.
2. Train a tree classifier f_b on N_b, where each split is computed as follows:
   - ✧ Select m axes in R_d at random.
   - ✧ Find the best split (j^*, t^*) from the subset of dimensions.
   - ✧ Split current node along axis j^* at t^*.

- *SVM(support vector machine) model*

A support vector machine constructs a hyperplane or set of hyper planes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the nearest training-data point of any class (so-called functional margin). Since in general the larger the margin, the lower the generalization error of the classifier.

- *Logistic Regression model*

Logistic regression is a regression model where the dependent variable (DV) is categorical. This article covers the case of binary dependent variables—that is, where it can take only two values, such as pass/fail, win/lose, alive/dead or healthy/diseased.

Algorithm:

$$\pi(x_i)|\, x_i = \frac{e^{\beta_0+\sum_{j=1}^{p}\beta_j x_{i,j}}}{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j x_{i,j}}}$$

$$y_i|\, \pi(x_i) \sim Bernoulli(\pi(x_i))$$

- *Naïve Bayes Model*

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

Algorithm:

$$\begin{aligned} &\mathrm{P}\,(\mathrm{Y}=\mathrm{y} \mid X_1 = x_1, \dots X_m = x_m|\mathrm{Y}=\mathrm{y}) \\ &\quad \propto \mathrm{P}(\mathrm{Y}=\mathrm{y})\ \mathrm{P}(X_1 = x_1, \dots X_m = x_m|\mathrm{Y}=\mathrm{y}) \\ &\quad \approx \mathrm{P}(\mathrm{Y}=\mathrm{y})\ \textstyle\prod_{i=1}^{m} \mathrm{P}(X_1 = x_i|\mathrm{Y}=\mathrm{y}) \end{aligned}$$

$$\begin{aligned} \widehat{y^{NB}} &= argmax_y\ \frac{P(Y=\tilde{y})\mathrm{P}(X_1 = x_{test}|\mathrm{Y}=\mathrm{y})}{\mathrm{P}(X_1 = x_{test})} \\ &= argmax_y P(Y=\tilde{y})\ \textstyle\prod_{i=1}^{m} \mathrm{P}(X_1 = x_{test,i}|\mathrm{Y}=\mathrm{y}) \end{aligned}$$

V. SOFTWARE PACKAGE DESCRIPTION

In our research, we mainly use R to process our algorithm. The packages we used are shown as followed:

1. JSON: Converts R object into JSON objects and vice-versa.
2. Random Forest: construct model based on random forest algorithm.
3. E1071: Functions for latent class analysis, short time Fourier transform, fuzzy clustering, support vector machines, shortest path computation, bagged clustering, naive Bayes classifier,
4. Glmnet: Extremely efficient procedures for fitting the entire lasso or elastic-net regularization path for linear regression, logistic and multinomial regression models, Poisson regression and the Cox model. Two recent additions are the multiple-response Gaussian, and the grouped multinomial. The algorithm uses cyclical coordinate descent in a path-wise fashion, as described in the paper linked to via the URL below.

VI. EXPERIMENT RESULTS

(1) Fake Review Selection

Compute the similarities among reviews which belongs to same restaurant and the score of each review.

After inspecting the results along with a consideration of reality, we choose the thresholds to be:

Similarity < 0.4 & Score < 0.1

In this way, we eliminate 160 fake reviews.

Here list 2 sample spam reviews from results:

1	Maybe I just don't like German food but was very unhappy with all the food I tasted. It could have been what I ordered? I was disappointed with the selection of food as well. I ordered a Chicken Salad Sandwich and it had way too much sauce on it. My husband let me try a bite of his Rueben which was really good. My kid's loved the authenticity of the restaurant which made it worth it to see them so happy. I wouldn't choose to go back but I am guessing it's probably b/c German food isn't my thing.
2	Yucky greasy food.

(2) Usefulness Prediction

Diagnosis Measurements

To measure how good our models are at predicting usefulness, we first introduce in the following three statistics that are widely used in classification diagnoses[5]:

--Sensitivity: the proportion of true uselessness's that are correctly identified by the model.

--Specificity: the proportion of true usefulness's that are correctly identified by the model.

--Balanced Accuracy: the average of the previous two statistics.

To explain further, following table is an example of a confusion matrix of a two-class problem:

	Reference	
Predicted	Label 0	Label 1
Label 0	A	B
Lable 1	C	D

The sensitivity and specificity are defined as:

Sensitivity=A/(A+C), Specificity=D/(B+D)

Model Comparison

After implementing the four classification methods on the train set, we use the models to do predictions on the test set. Then the results are listed below:

Model Type	*Sensitivity*	*Specificity*	*Balanced Accuracy*
Logistic Regression	0.9743	0.1517	0.5630
Naïve Bayes	0.9534	0.1838	0.5686
Random Forest	0.9560	0.2479	0.6019
Support Vector Machine	0.9544	0.1987	0.5766

Results Explanations & Samples

In our cases, the sensitivity represents the success rate in detecting the useless reviews. On the other hand, the specificity represents the success rate in detecting the useful reviews.

Due to the prevalence of useless reviews being labeled in the data, the sensitivities are quite high among all the models. On the contrary, the specificities are all extremely low, less than 25%. It seems we lose a huge amount of useful reviews after classification prediction.

But the classification models are still meaningful. First, we successfully controlled the error rate of turning a useless review to be a useful one at a very low level, which will result in providing the users with lots of spam information if is high. Second, our final objective is to present the users only a few reviews that are most reliable and informative. In this sense, low specificities around 25% have already captured enough useful reviews for further research.

Here the table lists 5 sample reviews from results:

1	Good old-fashioned American diner. The ser**vice was a lot friendlier than "kiss my grits"** level, so kudos for that, but the kitchen had a hard time handling my request to put out my short stack a little while after my main meal. I ended up waiting ten more minutes after we had finished our other food.
2	The buffet is of a good size, and it does appear to be kept relatively fresh, but I'm going to have to say that the quality is... questionable. I'm not trying to say that the food is undercooked or anything along those lines. Just that when you bite into a piece of teriyaki chicken on a stick, it should generally taste like chicken (or at least a meat that tastes like chicken).
3	Great old school diner with some tasty comfort food.
4	Good for their simplicity, they have this cheapo little sandwich called a "skinny" that is just meat and bread. Of course, if you want anything else on it they start jacking up the price, even for mayo or veggies. Pretty good though for a quick snack.
5	Not bad Chinese food at all - we came here for Sunday dim sum which was pretty decent. Service was decent, dim sum was good but there were some dishes that I was a bit disappointed by, but the ones that were pretty satisfactory included the spareribs in black bean sauce, beef tripe, bean curd, and the shrimp shui mi.

VII. CONCLUSION

(1)Conclusion

In aims to recommending reliable reviews to the users, we first filtered potential fake reviews by using the approach of combining both similarity analysis and a reasonable score which roughly measuring readers' responses of a review. This step is very meaningful for users who may be mislead by the fake reviews.

After removing the fake reviews from the data, we then predicted the usefulness of reviews by setting up classification models including logistic regression, Naïve

Bayes, random forest and SVM using the attributes in the yelp datasets such as review's length, user's fans, the existence length of user's account, etc.The prediction results show that we successfully extract enough reliable reviews to be recommended to the users.

(2) Future work

In the future steps, we may improve the precision of fake reviews' detection. One likely approach is known as sentimental analysis.

Also, another meaningful work is to find out the keywords in the useful reviews. We have a direct intuition from the results of useful reviews after classification prediction, a good review contains much information in the form of keywords like famous dishes, service time, restaurant location, etc. This objective can be realized by the topic models such as LDA.

Besides, recommendation of the selected useful reviews could be implemented to improve the users experience while searching for helpful reviews. The recommendation can be designed according to the most concerned elements of each user such as taste, environment, service, and etc.

REFERENCES

[1] N. Jindal et al., Opinion Spam and Analysis, Proceedings of the International Conference on Web search and web data mining (WSDM)., 2008.

[2] A. Mukherjee et al., Spotting Fake Reviewer Groups in Consumer Reviews, WWW 2012, ACM 978-1-4503-1229-5/12/04.I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] S. Xie et al., Review Spam Detection via Time Series Pattern Discovery, Proceedings of the ACM conference on Knowledge Discovery and Data Mining (KDD), 2012

[4] M. Ott et al., Finding Deceptive Opinion Spam by Any Stretch of the Imagination, Association of Computational Linguistics (ACL), 2011.

[5] Altman, D.G., Bland, J.M. Diagnostic tests 1: sensitivity and specificity, British Medical Journal, vol 308, 1552,1994.