Fake Review Detection on Yelp Dataset Using Classification Techniques in Machine Learning

Andre Sihombing and A.C.M. Fong
Department of Computer Science
Western Michigan University
Kalamazoo, MI 49008, USA
andre.sihombing@wmich.edu, acmfong@gmail.com

Abstract— **This paper provides a summary of our research, which aims to build a machine learning model that can detect whether the reviews on Yelp's dataset are true or fake. In particular, we applied and compared different classification techniques in machine learning to find out which one would give the best result. Brief descriptions for each of the classification techniques are provided to aid understanding of why some methods are better than others in some cases. The best result was achieved by using the XGBoost classification technique, with F-1 score reaching 0.99 in prediction.**

Keywords- machine learning; classification; fake reviews detection; online discussion forum

I. INTRODUCTION

With the growth of online information today, people tend to see reviews first for the places they want to visit, such as restaurants, hotels, or other businesses they need or before they go and buy some product. Yelp is a business directory service and crowd-sourced review forum that is usually used by people to post some review about their opinion for business. Statistics show that by the end of 2018, there have been more than 177 million reviews on the Yelp website. It is benefiting both consumers and businesses. For a business owner, they get free advertising from people who give a useful and positive review of their business. Unfortuantely, the problem arises when a small portion of irresponsible business owners try to boost up their market by hiring people to create some fake reviews about their business on Yelp website.

Yelp realizes this potential threat will create misleading information for their users. To overcome this problem, Yelp has already provided reviews policy for business owners. Other than that, Yelp has also implemented a recommended software system that aims to automatically filter all reviews have been determined to be problematic. In order to keep their content helpful and reliable, Yelp tries not to highlight reviews written by users that they do not know much about or reviews that may be biased because they were solicited from family, friends, or favored customers. The reviews are evaluated based on quality, reliability, and user activity[1]. Currently, about 75 percent of all reviews on Yelp website is recommended.

However, no system or method can be truly foolproof. In an attempt to improve the accuracy of identifying fake reviews, machine learning can be very useful. In particular, machine learning classification techniques can learning from data and then be applied to separate truthful reviews from fake ones.

The rest of this paper is organized as follows. Section II reviews relevant literature that sets the scene and forms the foundation of our research. In particular, it surveys four popular machine learning classification approaches. Section III explains our method. Section IV presents preliminary results of our method. Finally, Section V concludes the paper.

II. LITERATURE REVIEW

In this research, we have chosen to investigate the use of four well-known machine learning classification techniques. They are: Logistic Regression, Gaussian Naïve Bayes, Support Vector Machine, and XGBoost. A machine learning classification technique itself is used to group data based on their common characteristics. The grouping process is based on learning algorithm [2]. The learning process later will be resulting in a classification model. All the classification techniques in this research use an algorithm which relies on supervised learning. So, the dataset used in this research already includes descriptions of classes, which in this case are filtered review and non-filtered review.

A. Logistic Regression

There are fundamentally three kinds of classifiers, namely generative, probabilistic, and discriminative classifier [3]. Logistic regression is a discriminative classifier and belongs to the groups of regression methods. Logistic regression concerns describing the relationship between explanatory variables and discrete predictor [4], which is achieved by estimating probabilities using the underlying logistic function. Logistic regression assumes the explanatory variables are independent of each other, which can be an advantage or disadvantage depending on the dataset.

While in contrast, our next classifier, the Naïve Bayes assumes dependency exists between variables [3]. Logistic regression is ideally suited for binary classification problems

978-1-7281-5529-6/19/$31.00 ©2019 IEEE

meaning the predicted variable or the classes cannot be more than two. As our dataset only has two classes, so, this technique is suitable for our experiment.

Logistic regression will not perform correctly in a sketchy dataset, so, we must first identify all the essential independent variables and remove variable that unrelated and variable that are very similar to each other before we apply this classification technique.

B. Naïve Bayes

Naïve Bayes classifier is a probabilistic classifier. It is based on the Bayesian theorem and operates on conditional probability [5]. The Naïve Bayes calculates a set of probabilities by combinations of values in a given data set. Commonly used in text classification, for example, document classification and spam filtering, the Naïve Bayes classifier has fast decisions making process compared to other classifiers, and the Naïve Bayes often works well on even a small amount of training data.

In this experiment, we used the extent of the Naïve Bayes, which is Gaussian Naïve Bayes; this particular type of algorithm is specifically used when the variables or the features have continued values.

C. Support Vector Machine

Support vector machine is originally developed to solve binary classification problems (although multiclass extensions have been proposed) and have been used successfully across a wide range of machine learning applications [6]. The support vector machine method classifies with the help of a linear or non-linear function. The support vector machine method is based on the estimation of the most appropriate function for separating data [7]. Imagine we have two separate classes in our classification model, the Support vector machine method objective is to find a particular linear boundary to separate the classes.

In the process, there is a possibility that we have more than one linear boundary that can separate the classes, the support vector machine methods then will choose a line that maximizes the margin between classes, and thus the maximum error tolerance is determined. The vector points that the margin line touches are known as support vectors, and that is where the name support vector machines come from.

Upon identification of training data and the margin line, test data is classified based on their places in reference to the margin. Recently, many classification algorithms have been proposed, but support vector machine is still one of the most widely and most popular used classifiers [8].

D. XGBoost

XGBoost stands for extreme gradient boosting created by Tianqi Chen [9]. It is currently one of the algorithms that have recently been widely used for applied machine learning. XGBoost is an implementation of gradient boosted decision trees that designed for improving speed and performance [9]. It has proven to push the limits of classification performance. The following are characteristics of XGBoost as presented by its creator.

- Push the limit of computation resources to solve one problem
- Gradient tree boosting
- Automatic handle missing value
- Interactive Feature analysis
- Extendible system for more functionalities
- Deployment on the Cloud

In this research, we investigate and compare these four machine learning classification approaches for separating fake reviews from true ones.

III. METHOD

A. Data Description

For this research, we used the dataset from Prof. Rayana from Stony Brook University; she used this data for her research before [10]. The reason why we use this dataset is because this dataset is already labeled, so every review is labeled either is it filtered or not filtered. The term filtered here means there is an indication that the reviews is created by some fake reviewer.

The dataset contains reviews of restaurants and hotels in Chicago. We used only restaurant reviews in our experiments. There are 61541 reviews from 33502 reviewers. The reviews are collected from the period October 12, 2004 to October 8, 2012. There are 53400 nonfiltered reviews and 8141 filtered review. The content of the dataset is as follows.

Review metadata:

- date (date)
- review ID (text)
- reviewer ID (text)
- business ID (text)
- label (Y/N)
- useful (integer)
- funny (integer)
- cool (integer)
- stars (integer)
- review (text)

B. Procedure

As shown in Fig. 1, we divided the workflow into three parts: data preprocessing, feature engineering, and classification process using machine learning algorithm. In data preprocessing part, we figured out how to handle the unbalanced data problem. In the feature engineering step, we tried to explore the data in order to extract more complex features using statistical method. The last thing in the

classification part, we used four machine learning algorithm such as logistic regression, support vector machine, Gaussian Naïve Bayes, and XGBoost. The three parts are described further below.

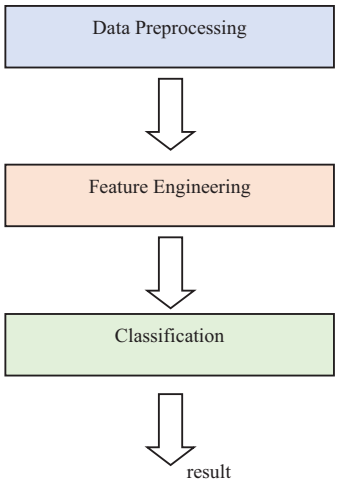

Figure 1. Overall workflow.

Data Processing: According to the dataset, the ratio of filtered reviews and non-filtered reviews is approximately 1:6, which is very unbalanced for the classification. Therefore, we try to apply two methods to deal with this problem. First is over-sampling, increases the weight of the minority class by replicating the minority class data. In this case, which is to add more copies of filtered reviews, so we copy the filtered reviews three-time, therefore, the ratio decreasing to approximately 1:3. The second method is under-sampling method; the basic idea in this method is to remove some non-filtered reviews from the training data. After we remove the non-reviews reviews, the ratio was decreasing to approximately 1:3. The result show oversampling method gives more good result than under-sampling method. It is reasonable because oversampling method keeps all the information in the training dataset. While in under-sampling method, we lost much information.

Feature Engineering: Before doing feature engineering, we do some statistical analysis on the dataset. We found that filtered review tends to give more extreme ratings such as 1 or 5 (see Figure 2) and also mostly filtered review is shorter review than non-filtered review, even this is not too obvious, but we can use this as additional features.

Besides the basic features we have in the data set such as useful, funny cool and star rating, we tried to extract some other complex features in order to give more characterization for the machine learning classification in training process. We analyzed the business background behind the fake reviews and extracted the possible features which might indicate the signs of suspicious or malicious reviews.

In terms of length of reviews [10], [11] People tend to choose not to spend much time in non-realistic things like making up reviews. Therefore, fake reviews are generally shorter than the true one. This feature is a number of words after we do preprocess (tokenizing the words and removing stop words and punctuations) in the review.

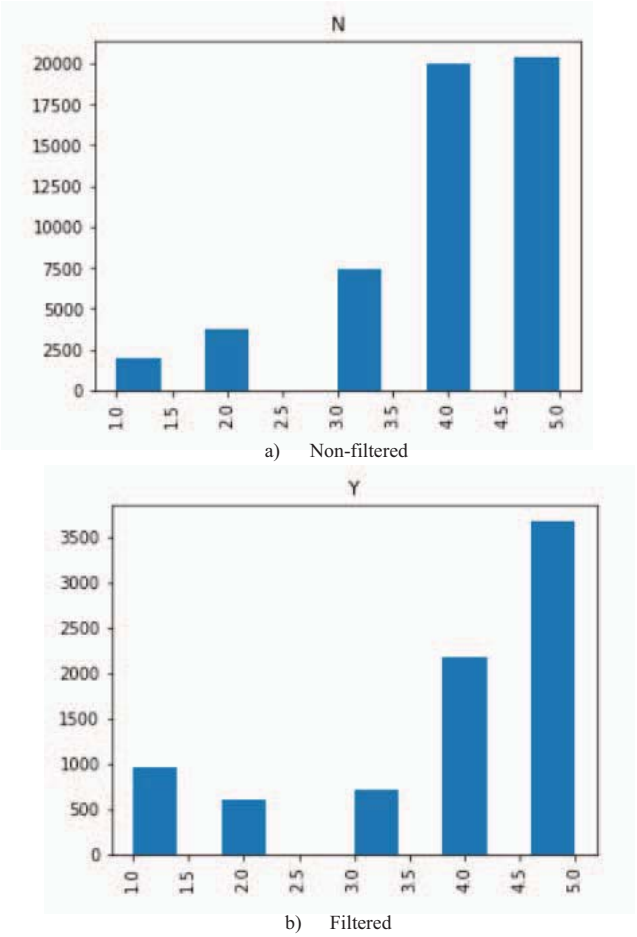

Figure 2. Rating distributions for a) non-filtered and b) filtered review.

In terms of similarity of reviews [10], [11], since the process of doing several fake reviews needs to be done in a short time, mostly fake reviewers duplicate their previous fake review to create a new fake review. To extract this feature we do the following. First, for all unique reviewers, the corresponding reviews (after tokenizing and preprocessing) were extracted. Cosine similarity for each pair of these reviews was calculated, and then the maximum of these numbers was taken as the feature for the reviewer. For all the unique reviewers, we calculated this value and fed this value to the review, which was reviewed by the corresponding reviewers.

Another consideration is rating deviation from the mean rating reviews [10], [11]. Rating reviews that divergence from others are suspicious, so this is the deviation of the rating by the corresponding reviewer from the average rating of other reviewers for a business. The mean rating was calculated for each business (excluding the rating by the reviewer in hand), and then the absolute deviation was extracted.

Another useful feature is maximum review numbers per day of reviewer [10], [11]. Usually, a reviewer will make one

or two reviews per day for the business they just visited, on the other hand, fake reviewer generally writes many reviews in a short time, so we can say it is suspicious when a reviewer can make more than three reviews per day.

We calculated the max number of reviews in a day for the corresponding reviewer of the review. These values were then normalized by the maximum number of reviews by any reviewer in the dataset. For all the unique reviewers, we calculated this value and fed this value to the review, which was reviewed by the corresponding reviewers.

Finally, extreme rating ratio of the reviewer [10], [11] is also an interesting feature. Fake reviewer will always give either (1 or 5) star to convince people of their opinions, according to this, I calculated the extreme rate (1 star or 5 stars) ratio for every reviewer and used the ratio as one feature of every review. For all unique reviewers, the ratio of extreme rating (1 or 5) was calculated by dividing the number of extreme ratings by the reviewer by the total number of reviews by the reviewer. For all the unique reviewers, we calculated this value and fed this value to the review, which was reviewed by the corresponding reviewers.

IV. Evaluation

To train our model, we used all the basic feature and complex features such as useful, funny cool rating, length of reviewer reviews, the similarity of reviewer reviews, extreme rating ratio of the reviewer, rating deviation from the mean rating reviews, max review numbers per day of the reviewer. Based on all the extracted features, we implement four machine learning algorithm, logistic regression, support vector machine, Gaussian Naive Bayes, and XGBoost. The results are shown in Tables I - IV below.

TABLE I. Logistic Regression.

Report for Logistic Regression (Over Sampling: 53400:24421)				
	Precision	Recall	F1-score	Support
N/True	0.83	0.85	0.84	15969
Y/Fake	0.65	0.63	0.64	7378
Avg/total	0.78	0.78	0.78	23347

TABLE II. SVM.

Report for Support Vector Machine (Over Sampling: 53400:24421)				
	Precision	Recall	F1-score	Support
N/True	0.83	0.84	0.83	15969
Y/Fake	0.63	0.62	0.63	7378
Avg/total	0.73	0.77	0.77	23347

TABLE III. Gaussian Naïve Bayes.

Report for Gaussian Naïve Bayes (Over Sampling: 53400:24421)				
	Precision	Recall	F1-score	Support
N/True	1.00	0.49	0.66	15969
Y/Fake	0.48	1.00	0.64	7378
Avg/total	0.74	0.74	0.65	23347

TABLE IV. XGBoost.

Report for XGBoost (Over Sampling: 53400:24421)				
	Precision	Recall	F1-score	Support
N/True	1.00	0.98	0.99	16007
Y/Fake	0.96	1.00	0.98	7340
Avg/total	0.99	0.99	0.99	23347

The result of machine learning prediction using XGBoost classification algorithm achieved 0.99 in F-1 score average in prediction while using the other algorithms, the maximum F-1 score is 0.78 on average. Another finding is YELP filter consider behavioral features than Linguistic feature as we find that entirely same review by the same reviewers was not considered as a fake review.

V. Conclusion

This paper has reviewed four popular machine learning classification methods for finding fake Yelp reviews. Reviews rates such as useful, cool and funny only acquired by non-filtered review mean soon after the reviews get filtered by Yelp, the review will be hidden so it cannot be rated by others.

The experiment results showed a very high score in prediction, when using XGBoost. There are still many features that we cannot implement because of the limitation in the dataset such as user trust factor based on user friendship, and also user Profile (join date, photo, etc.).

Imbalance on the dataset needs to handle because imbalance dataset gives poor result in our experiment. While running the experiment, we found that SVM took the longest time to train the model, and Gaussian Naïve Bayes gave the lowest score on average.

In our opinion, we cannot say that reviews got filtered by YELP recommendation system is 100% fake, because there are still other factors that may lead machine learning into false prediction. Other techniques that are potentially reliable and can be used for filtering review is using verified buyer method as some crowdsource web have been used.

References

[1] "Recommended Reviews | Support Center | Yelp." [Online]. Available: https://www.yelp-support.com/Recommended_Reviews?l=en_US. [Accessed: 07-Aug-2019].

[2] H. I. Bülbül and Ö. Ünsal, “Comparison of classification techniques used in machine learning as applied on vocational guidance data,” Proc. - 10th Int. Conf. Mach. Learn. Appl. ICMLA 2011, vol. 2, pp. 298–301, 2011.

[3] A. Prabhat and V. Khullar, “Sentiment classification on big data using Naïve Bayes and logistic regression,” 2017 Int. Conf. Comput. Commun. Informatics, ICCCI 2017, 2017.

[4] J. R. Brzezinski and G. J. Knafl, “Logistic regression modeling for context-based classification,” pp. 755–759, 2008.

[5] B. Çiğşar and D. Ünal, “Comparison of Data Mining Classification Algorithms Determining the Default Risk,” Sci. Program., vol. 2019, 2019.

[6] S. Edition, A First Course in Machine Learning, Second Edition. 2018.

[7] Y. Ozkan, Data Mining Methods. Istanbul: Papatya Publications, 2008.

[8] E. Elmurngi and A. Gherbi, “An empirical study on detecting fake reviews using machine learning techniques,” 7th Int. Conf. Innov. Comput. Technol. INTECH 2017, no. June 2018, pp. 107–114, 2017.

[9] “A Gentle Introduction to XGBoost for Applied Machine Learning.” [Online]. Available: https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/. [Accessed: 07-Aug-2019].

[10] S. Rayana and L. Akoglu, “Collective Opinion Spam Detection : Bridging Review Networks and Metadata,” Kdd, pp. 985–994, 2015.

[11] A. Mukherjee, V. Venkataraman, … B. L.-S. International A., and undefined 2013, “What yelp fake review filter might be doing?,” Aaai.Org, pp. 409–418, 2011.