Prediction of Review Sentiment and Detection of Fake Reviews in Social Media

Amani Karumanchi **Lixin Fu** **Jing Deng**
Department of Computer Science
University of North Carolina at Greensboro
Greensboro, NC 27403, U.S.A.
{a_karuma, lfu, jing.deng}@uncg.edu

Abstract - *For the past dozen years, social media has been widely used and has slowly become major sources of various information. Some social media sites specialize in people's reviews and making recommendations. Some reviews are positive while others negative. Either way, they have a significant impact on users' decisions of whether purchasing the products or services or not. However, fake reviews and fake users can greatly twist the true opinions that are out there. The detection of fake reviews is critical for any progress toward making great usage of the web. In this work, we analyze the statistics of a Yelp Challenge dataset and propose a simple-but-powerful new detection algorithm based on locations of the reviewers and businesses. We also investigate the prediction of the sentiments of the reviews. We compare the accuracy and false alarm rates of several prediction algorithms. While the results are less than optimum, but there are hopes for strong performance with minor revisions.*

Keywords: Detection, fake reviews, sentiment analysis, social media, data mining

1 Introduction

According to recent studies, more than 60% of the world population uses Internet these days compared to 1% in 1995. On social media platforms users can create public profiles and interact with other people. They share ideas, activities, events and interests within their individual networks. Most sites support the maintenance of pre-existing social networks, but others help strangers connect based on shared interests, political views, online reviews or activities which may have great impact on people's purchasing decisions. A significant 90% of marketers find that social media is important to their businesses. Positive reviews can render significant financial gains and fame for businesses. On the flip side, negative ones can also equally dent businesses. This is called “Social Media Marketing”.

However, fake reviews by illegitimate users can cheat people to make wrong decisions. Handling of large amounts of data is also a big challenge. So, the detection of fake news is one of important research issues. In this work, we propose a simple-but-powerful new detection algorithm based on locations of the reviewers and businesses.

Yelp is a website which publishes crowd-sourced reviews about local businesses and provides opportunity to business owners to improve their services and helps the users choose best businesses amongst available. However, it is quite unlikely for the business owners to go through all the user reviews and make important decisions for the improvement of their businesses. We need automatic data analytic models to predict the sentiments of the reviews. We describe and evaluate our experiments on yelp datasets as a case study.

The rest of the paper is organized as follows. Section 2 discusses some related work. Section 3 describes the Yelp datasets and our detection strategy. Section 4 includes our main results and we conclude in Section 5.

2 Related Work

A number of research used yelp datasets on research in review sentiment and fake detection. In [1], supervised learning is based on pseudo fake reviews rather than fake reviews filtered by a commercial Web site. Here supervised learning is done using linguistic n-gram features that performed extremely well (attaining around 90% accuracy) in detecting crowd sourced fake reviews generated using Amazon Mechanical Turk (AMT). Theoretical analysis was proposed to uncover the precise psycholinguistic difference between AMT reviews and Yelp reviews (crowd sourced vs. commercial fake reviews). Using linguistic features, they found that in the AMT data, word distributions of fake and non-fake reviews are very different, which explains the high (90%) detection accuracy using n-grams but for the Yelp data word distributions in fake and non-fake reviews are quite similar, which explains why the method is less effective on Yelp’s real-life data.

Wang et. al [2] proposed a review spam detection method using tensor decomposition. There were several limitations in the above research work based on the datasets from other researched sites. They found that the real users tend to review the restaurants nearby, but the spammers are not restricted to the geographical location, i.e. they may come from anywhere. However, it was not true for the Yelp datasets. Here they found that 72% of the Yelp’s review spam is posted from areas in and around restaurants compared to 64% of the authentic reviews. It addresses the spam detection issue with a new perspective.

They have found that the review sentiment is one of the main factors affecting review ratings and hence need to be further investigated for accurate prediction.

In [3], Luca found that consumers did not use all available information and were more responsive to quality changes that are more visible and consumers responded more strongly when a rating contained more information. Using a data set consisting of reviews from the website Yelp.com and revenue data from the Washington State Department of Revenue, he presented that the revenue generated from Yelp ratings had increased as consumer response was larger when ratings contained more information. The above analysis was also confirmed by using RD-Methodology where restaurants might submit false reviews. McClary density test was implemented to rule out the possibility that RD-Methodology was biasing the results. Also impact of Yelp on revenues for chain restaurants was examined. It was also investigated whether the observed response to Yelp was consistent with Bayesian Hypothesis. The author also performed business rating prediction based on sentiment analysis. He also compared the strengths and weaknesses of different sentiment analysis models. A similar work of predicting star rating based on sentiment analysis of business review data appeared in [4]. Most of these studies were focused on the star rating prediction.

Research in [5-7] handles the detection methods of fake reviews or spams. A collective opinion approach is used in [8-9]. Work in [10-11] handles opinion spam and analysis. Fan and Khademi [12] used a combination of three feature generation methods as well as four machine-learning models to find the best prediction of star ratings for the businesses. You et al. [16] analyzed sentiment using images instead of textual contents. Nakov et al. investigated sentiment analysis on Twitter [17]. Based on a neural network method, Tang et al. [18] gave a model for predicting review ratings.

3 Yelp Dataset and Detection Strategy

The yelp academic challenge datasets are provided by Yelp.com. Yelp connects people with local businesses and the dataset provides data about users' experiences at each business via reviews, tips, check-in and business attributes during a period between September 2017 and December 2017. The scope of local businesses in the chosen dataset is mostly in Canada, U.S.A. and some parts in Germany and U.K. Yelp provides a way for users to explore, rate and review the businesses they visit. Businesses can highlight their products and services that will attract users to them and finally rate the business. Yelp dataset contains a vast variety of businesses, like restaurants, bars, cafes, spa, hotels and so on. Users having accounts can also add their friends to yelp. Users can give a star rating from 1 to 5 for a business and can also write a text review which clarifies the rating. These ratings are very useful for other users who are exploring local business and will help them in finding which one would be the best for them. These features of yelp make it a highly popular website. Each business has an overall rating, an average of the star ratings for all the reviews that the business has reviewed. Users can also vote for reviews written by other users.

3.1 Structure of the Yelp dataset

The Yelp dataset is a single zip-compressed file, composed of five compressed json files. Every file contains a 'type' field, which implies whether it is a business, a user, a review, a check-in or a tip. The fields are separated by comma. The size of business file: 126MB, review file: 3.55GB, user file: 1.46GB, tip file: 176MB, check in file: 57.3MB. Yelp dataset contains 156639 businesses and 1183362 users with 4713214 reviews. The dataset includes businesses in different countries. Following is a description of all the fields in each of the data types.

Business: For each business, there is a business id and name, address including latitudes and longitudes of that location. It also has the number of reviews that have been written for the business, and an average star rating across all the reviews. Yelp also stores other information as attributes such as hours, parking, wheelchair accessibility, ambience etc.

Review: Each review consists of business_id, review_id, and user_id, review date, a star rating and review text, possibly justifying the star rating. Each review can also get votes from other users. They can vote if a particular review is useful.

User: Yelp contains a strong user network, and stores information about them: their name, user_id, average_stars, vote counts, the number of reviews they have written, how long they have been using Yelp for, their friends those who uses Yelp, user's fans count.

Check in: This gives an aggregated view of all the check-ins for a business for every hour of the day, for every day of the week, and gives a great idea about what are the busiest times for the businesses.

Tip: Tips stores random comments that users leave about a business. They are different from reviews, in that they do not have a star rating and are just quick indications for others. Information includes business-id, date, likes, text, and user_id.

The Yelp data is first passed through pre-processing, sorted, and elimination of unnecessary fields. The following fields are maintained:

Review Related Features: Even without the content of review text, we can still find other helpful information out of review data.

- Reviews Identity (Review_id)
- Length of review text
- Review stars
- Date of the review: to transform the date into a feature, we count the days from a specific date.

User Related Features: We assume the popularity of a user is helpful for prediction. Following are few attributes that can describe user's popularity.

- User Identity (User_id)
- Average star given by the user
- Review count: number of review from the user
- Number of fans of the user
- Number of friends of the user

Business Related Features: Attributes of the targeted business can influence the results.

- Business Identity (Business_id)
- Review count
- Stars
- Latitude and longitude
- City
- Postal

Detection Strategy: The detection of fake reviews in Yelp is a rather challenging task, due to its natural complexity and diversity. However, when a reviewer writes a review to the restaurants in multiple countries on the same day, it is more likely that these reviews are fabrication unless such reviews specifically mentioned international trips. Even though it is possible for the reviewer to have travelled through these countries and summarized/reviewed different business at a later time, the accuracy and its timing are still questionable.

Using Python, we generated the list of fake users based on the latitude, longitude, postal, and dates of the reviews written for the particular businesses by a particular user.

4 Analysis Results

We present our analysis results in this section.

4.1 Yelp Dataset Statistics

Table 1, Statistics of Yelp Challenge dataset.

Number of Reviewers	1,183,362
Number of Businesses	156,639
Total Reviews	236,890,795
Average number of Reviews per Reviewer	200.18
Average number of Reviews per business	1,512.33
Average Review Rating	3.65

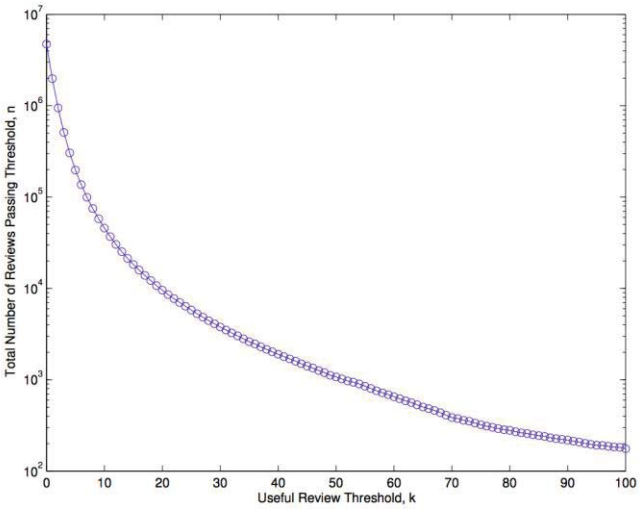

Figure 1, Total number of reviews with at least k "useful" flags.

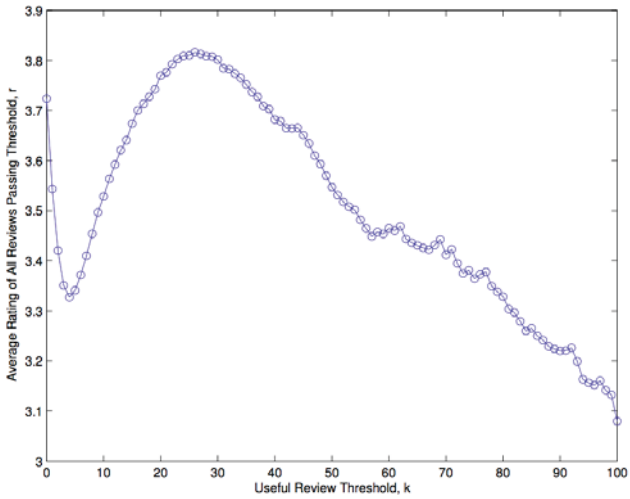

Figure 2, Average rating of reviews with at least k "useful" flags.

Statistics of the Yelp challenge data are given in Table 1. There are a total number of 236,890,795 reviews made available in the dataset. These were submitted by 1,183,362 reviewers toward 156,639 businesses. On an average, each reviewer submitted 200.18 reviews and each business received 1,512.33 reviews, both of which looked surprisingly high from a first glance.

All reviews should not be treated equally. Some are rated as useful, as indicated by the large number of "useful" flags (similar to the "likes" on Facebook.) When we remove those reviews with lower number of "useful" flags, the reviews have different statistics. Figure 1 shows the total number of reviews among all of the 236,890,795 reviews with at least k number of "useful" flags. The number of such reviews decreases with k, naturally. For instance, when we require reviews to receive at least 5 and 10 "useful" flags as compared to 0, the total number of available reviews drops from about 237 million to 200,000 and 40,000, respectively.

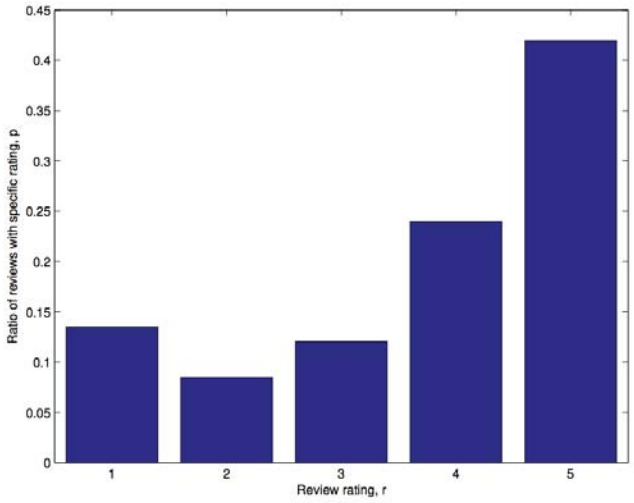

Figure 3, Ratio of reviews with specific rating (k=0).

Figure 2 is even more interesting, showing the average rating among those reviews with at least k "useful" flags. Note that these do not consider different businesses but just compute the overall average. As the value of k increases, the average rating first drops significantly from 3.72 to 3.32 then slowly rises to 3.81 before eventually decreasing from there. The two extremes require some explanation and even further investigation. It seems that the average rating of all reviews, i.e., without considering any "useful" flags, is rather arbitrary (or not very useful). The valley of the average rating occurs at k=4 and the average rating is 3.32, about 10.7% lower than the initial average of 3.72 (k=0 case). This could have been caused by the generous reviews/ratings submitted by both casual and more serious reviewers (and some of these did not receive enough "useful" flags). As the value of k further increases from 4 to 26, the average rating continues to rise, reaching as high as 3.81. This could have been caused by the increasing impact of useful reviews, or elimination of casual/generous reviews. In fact, some of those businesses might have been purged from the review system with such a strict review screen process (larger k).

We consider it difficult for any review to receive a large number of "useful" flags, e.g., k>26, unless it is truly useful and deemed helpful by other users. Therefore, this is an indication that users prefer lower-than-perfect review ratings because they point out issues, albeit sometimes minor ones, that users need to look out for.

In Figure 3, we show the bar-graph of ratios of review ratings among all reviews without any screening (k=0). It is interesting to see that there are more ratings of 1 and 5 compared to others, an indication of reviewers' tendency of going to extremes, either perfect score (5) or lowest score (1). The ratios for ratings of 2 and 3 are relatively rare, accounting for only 8% and 13%.

In order to investigate the impact of such a screening process on individual business' average rating, we calculated the mean and standard variations of individual business ratings with different k values (Table 2). Therefore, these results indicate the average rating and its overall range when we ignore those reviews without sufficient "useful" flags. For small k values, more businesses would have been rated higher and with smaller variations. As the screening threshold k increases toward k=5 (note the slight difference with the valley of k=4 in Figure 4), the average rating decreases with about 10% drop and the overall ratings are more diverse (standard variation changing from 0.97 for k=0 to 1.33 for k=5, a 40% increase). Fewer and fewer businesses have reviews with at least k "useful" flags when k is greater than 5 and their ratings rises with smaller variations. Therefore, k=5 seems to be a good threshold to screen reviews, below which they should be ignored.

In Table 2, the comparisons of average rating and standard variation of those reviews for k=0 and k=20 (two extremes in filtering reviews) are also interesting: both of these average ratings are higher than others that we investigated with smaller variation (low standard variation). This suggests that such reviews are rather inflated and should be treated with cautions.

Table 2, mean and standard variation of reviews with at least k "useful" flags.

"Useful" flag threshold, k	Total number of business with such reviews	Rating average	Rating standard variation
0	156,639	3.65	0.97
1	140,237	3.48	1.17
2	111,844	3.37	1.25
5	51,686	3.32	1.33
10	18,947	3.48	1.31
20	5,536	3.76	1.19

4.2 Performance of Detection Techniques

Sentiment analysis is a more interesting technique to study review accuracy. We used supervised learning as a tool to accomplish the task. To predict an unknown class from known attributes, we have chosen few useful algorithms for that space. It can be used to detect fake users by looking at it as the classification problem of separating users into two classes: fake and non-fake reviews. Supervised Algorithms like K-Nearest Neighbor (KNN), Logistic Regression, Decision Tree

Classifier, Linear Discriminate Analysis and Support Vector Machine are used on the extracted data.

The K-nearest neighbor uses the majority vote of the class labels of k nearest neighbors. Logistic Regression is a statistical model to estimate a binary dependent variable through a linear combination of independent variables. Linear discriminate analysis is another statistical method to separate two or more classes through a linear combination of features.

Table 3 shows the successful detection rates and false alarm rates for different algorithms, including LR, LDA, KNN, CART, and SVM. Among all of the algorithms that we investigated, SVM has the highest successful detection rate and lowest false alarm rate.

Table 3, Detection Rates and False Alarm Rates

	LR	LDA	KNN	CART	SVM
Detection Rate	0.798	0.800	0.771	0.670	0.806
False Alarm Rate	0.013	0.006	0.045	0.181	0.000

Finally, Table 4 lists the precision, recall, F1-score, and supports for SVM.

Table 4, SVM performance

	Precision	Recall	F1-Score	Support
Fake	0.80	0.84	0.82	155
Unfake	0.34	0.29	0.31	45
Avg/Total	0.70	0.71	0.71	200

Note that the results above should be taken with cautions: it is possible for a review to stay at one location and write a review for a business at another location, even out of country. It was found that 70% users who are writing reviews mentioned above, there is possibility that user might have mentioned the time they were present at that location as shown in Figure 4.

5 Concluding Remarks

In this work, data analysis is performed on the Yelp Challenge dataset. The data is processed and various spam detection patterns were researched and was implemented. From the experimental results, we find that different feature sets perform quite well for the extracted datasets. Furthermore, the support vector machine algorithm works the best and gives the best prediction accuracy. For instance, our analysis has identified an interesting threshold of k=5 as a strong indication of more accurate reviews and this can be used in practice.

This work certainly leaves a lot of future work to pursue: detailed analysis of review sentiments can be helpful to confirm rating accuracy; reviewer travel trajectory should be correlated with their review history; reviewer tendency should be taken into consideration when a business' overall rating instead of a simple average. We expect the addition of such techniques, while more complex, would improve detection accuracy.

Figure 4, a U.S.A. reviewer wrote a review on a business in the U.K.

References

[1] Arjun Mukherjee, Vivek Venkataraman, Bing Liu, Natalie Glance, "What Yelp Fake Review Filter Might Be Doing," ICWSM (2013), February 2017.

[2] Xuepeng Wang, Kang Liu, Shizhu He and Jun Zhao, "Learning to Represent Review with Tensor Decomposition for Spam Detection", EMNLP 2016. Association for Computational Linguistics (2016), March 2017.

[3] Michael Luca, "Reviews, Reputation, and Revenue: The Case of Yelp.com", Harvard Business School Working Paper, No. 12-016, March 2017.

[4] Luca, Michael, and Georgios Zervas, "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud", Management Science 62 (12), 3412-3427, February 2017.

[5] Arjun Mukherjee, Vivek Venkataraman, Bing Liu, Natalie Glance, "Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews", Technical Report UIC-CS-2013-03 (2013), April 2017.

[6] Huayi Li, Bing Liu, Arjun Mukherjee, and Jidong Shao, "Spotting Fake Reviews using Positive-Unlabeled Learning", Computación y Sistemas 18 (3), 467-475, May 2017.

[7] Geli Fei, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, Riddhiman Ghosh, "Exploiting Burstiness in Reviews for Review Spammer Detection", ICWSM 13, 175-184, August 2017.

[8] Shebuti Rayana and Leman Akoglu, "Collective Opinion Spam Detection: Bridging Review Networks and Metadata", KDD 2015, 985-994, August 2017.

[9] Huayi Li, Zhiyuan Chen, Bing Liu, Xiaokai Wei and Jidong Shao, "Spotting Fake Reviews via Collective Positive-Unlabeled Learning", ICDM, 2014, 899-904, November 2017.

[10] Huayi Li, Zhiyuan Chen, Arjun Mukherjee, Bing Liu and Jidong Shao, "Analyzing and Detecting Opinion

Spam on a Large-Scale Dataset via Temporal and Spatial Patterns", ICWSM, 634-637, December 2017.
[11] Nitin Jindal and Bing Liu, "Opinion Spam and Analysis", WSDM 2008, December 2017.
[12] Mingming Fan and Maryam Khademi, "Predicting a Business' Star in Yelp from Its Reviews Text Alone", arXiv preprint arXiv:1401.0864, December 2017.
[13] Andreea Salinca, "Business reviews classification using sentiment analysis", SYNASC 2015, December 2017.
[14] Yun Xu, Xinhui Wu, Qinxia Wang, "Sentiment Analysis of Yelp's Ratings Based on Text Reviews", 2014, December 2017.
[15] Yelp Dataset. http://www.yelp.com/dataset/challenge
[16] Q. You, J. Luo, H. Jin, J. Yang, "Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks," AAAI, 2015.
[17] Preslav Nakov, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Fabrizio Sebastiani. 2016a. SemEval-2016 task 4: Sentiment analysis in Twitter. In Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval '16, San Diego, California.
[18] D. Tang, B. Qin, T. Liu, Y. Yang, "User Modeling with Neural Network for Review Rating Prediction," IJCAI, 2015.