Inferring Future Success of a Business Based on Internal and External Factors

Mohammad Adil Harika Harini Sabella Sugeerth Murugesan

ABSTRACT

In order to predict if a particular business will do well in a city/town, we must uncover the properties of the business location and the tastes of the residents. For example, a not-so-good Indian restaurant in very popular area is more likely to prosper than a very good restaurant in a relatively unpopular place. One way to measure the success of a business is to find out the number of positive reviews a business will get based on the reviews it currently has. While taking into account the location of the business, we have developed techniques and visualizations that can be used to derive the future success factor of a business. We also came up with an algorithm that lets us detect fishy users who were paid by Yelp to write reviews and by eliminating reviews written by these users before calculating the success factor for a business, we can paint a very accurate picture of the future of a given business

1. INTRODUCTION

Yelp started off in 2005 as a platform for users to rate and review businesses in their city. Yelp has made its dataset for Phoenix, Arizona public. The Yelp dataset contains 335,000 reviews texts written by at least 70,000 users. We used the dataset to build models to infer the future success factor of a particular business. The review dataset of Yelp provides useful information about the users, businesses, and their ratings. While being in the market for few years, there is no system in Yelp to forecast trends for a particular business. Interesting questions arise: How much of an influence does the location have when considering the success rate of businesses? What business model should one adopt so that their business gets maximum profit in the next few years?

In this work, we describe techniques to forecast/predict the future success of a particular business. The success of a business depends on many outside factors not pertinent to the just the business itself (those factors that are outside the control of the business)–some of these factors include year, the current economy, and the customer base. Another such factor which influences the success of a business is the success of other businesses within the same community. Our hypothesis is that the success of a particular business depends not only on the quality of the products the business offers, the satisfaction of its employees and customers, and the profit it generates, but also on the success of the neighboring businesses.

Our idea which involves integrating the success of neighboring businesses and removing reviews from fishy users before calculating the success factor, is novel and different from approaches used before which were based simply on just the internal factors of the business at hand–these approaches did not consider any external factors. In order to make a sound determination of the future success of a business, such external factors must be considered. In our approach, we take into account the success of neighboring businesses in order to come up with a success factor for the business in question.

2. BACKGROUND

The paper "Inferring Future Business Attention" mentions that there are two categories of features: metadata (longitude/latitude) about business and description of reviews (number of stars/number of reviews) [1]. By performing what the authors call "sentiment analysis" on the reviews, they determine the features of a business and the positivity or negativity associated with these features. Sentiment analysis involves "[mining] the most frequently occurring keywords among all restaurant reviews and [using] the counts of the sentiments for each of these keywords as features" [1]. Each feature is associated with an adjective and the adjective is determined to be positive or negative. The features which are extracted from this process are used to determine the positivity or negativity of a review using PCA. Principal component analysis (PCA) is a tool that reduces dimensionality of a feature space–in other words, they reduce the number of features in the working set and predict future business using these new features. Three ways to collapse the data to extract out certain features:

- Univariate Feature Analysis – look at the prediction quality (meaning how much would including this feature help us in predicting the total number of future reviews) of each feature individually; use the best n of these in the prediction model
- Greedy Feature Removal – Involves "[removing] the worst performing feature until a certain performance

is achieved or until there are a specified number of remaining features left" [1].

- "A naive approach to feature selection uses an exhaustive search over all possible subsets of features to pick the subset with the smallest test error" [1].

Their work involved running a number of experiments to predict the future number of reviews using simple reviews (using features about the business metadata) first and then using review text features. By collapsing the methods as mentioned above, they were able to choose a few of these different features (to eliminate what they call noise) and build a regression model (which relates the current user reviews to future number of reviews) to predict the number of future reviews using regression model called Support Vector Regression (SVR).

We have built upon the algorithms and methods suggested in the "Inferring Future Business Attention" paper and come up with a way in which we can account for the external factors (particularly location) that affect the success of a business.

Section II reviews presents past work; Section III contains information about our technical approach and results; Section IV presents the implications of our work; and Section V reviews future work.

3. TECHNICAL APPROACH AND RESULTS

3.1 Heatmap Visualization

In this section, we will be describing we methods that we have used to generate our initial results. We begin with a simple context visualization of the Yelp dataset in Arizona, Phoenix. We generated a heatmap visualization, which defines clusters of businesses that are located close to one another. The areas that are colored in red are the areas where there are a large number of businesses.

With the overall idea of where the businesses are located in Arizona, Phoenix, we used two techniques for data analysis and generation of a regression model–semantic and location analysis.

3.2 Semantic Analysis

"Semantic analysis is the process of relating syntactic structures, from the levels of phrases, clauses, sentences and paragraphs to the level of the writing as a whole, to their language-independent meanings" [6]. Semantic analysis is central to deriving the success factor for businesses. More specifically, the following are the steps of our approach:

Given a business, P, we would like to derive a value, S, that indicates P's current success; in order to derive the success factor:

- Given a business, P, derive a value, S, that indicates P's current success (based on qualities of the business itself as well as the influence of the surrounding businesses). Then, by using S, we derive a regression model to find out the future success factor of a particular business P.

```
Chicken: [1, 'best', 0, 0],
Thai Pan: [7, 'divine, only, new, dish, wonderful, authentic, best', 0, 0],
Thai curries: [1, 'Chinese', 0, 0],
place: [3, 'dark, good, great', 0, 0],
restaurant: [2, 'next, clean', 0, 0]
```

Figure 2: Snapshot 1

  - (Step 1) Derive a value q which indicates how positive P's reviews are
  - (Step 2) Derive values [b1 ,b2 ,..., bn] where b1...bn indicate how positive the reviews of the n neighboring businesses of P are
  - (Step 3) Integrate q and [b1 ,b2 ,..., bn] to derive a value S which is the success factor for P.

- Given a business, P, we would like to predict how likely it is that P would succeed in the future
  - (Step 4) Take the success factor derived earlier, build a regression model, and derive a future success factor value based on its location

To find out the number of positive reviews that a business, P, currently has, we first extract all the associated reviews for a particular business from the Yelp review dataset. Then, we tokenize these reviews into individual words. Using the Python Natural Language Toolkit, we assign each word to its part of speech thus when encountering a word that is a noun, we push it into the stack In order to build a noun phrase, if the immediately occurring words after we encounter a single noun are also nouns, we push them onto the stack. We keep the latest occurring noun/noun phrase in the stack. When we encounter a word that is an adjective, we associate the noun/noun phrase in the noun stack with this adjective. At the end of this process, we have a counter which indicates the number of times a noun/noun phrase is seen throughout the reviews for the business and the all associated adjectives for the noun/noun phrase. More concretely, here is a sample of the noun adjective dictionary: Noun: [# of occurrences, associated adjectives, # of (+) adjectives, # of negative (-) adjectives] (See Snaphot 1 in Figure 2 for an example).

Then, we built a list of adjectives that are generally considered to be positive and a list of adjectives that are considered to be negative This is done using a lexical database called Wordnet[5]. To build these lists, we start out with a small list of positive adjectives such as "great" or "wonderful" and recursively derive all the synonyms of these words to generate a somewhat complete list of positive and negative adjectives.

Using these positive and negative adjective lists, we can derive whether the given noun/noun phrase is associated with a positive adjective or a negative adjective. We then increment the appropriate positive and negative counter for each noun/noun phrase. For example, using the same example as above, we would now have the noun adjective dictionary depicted in Snapshot 2 (Figure 3)

Note that, for the noun "Thai pan" the total count of the positive and negative adjectives does not add up to 7 because the adjectives "new" "dish" and "only" are not particularly

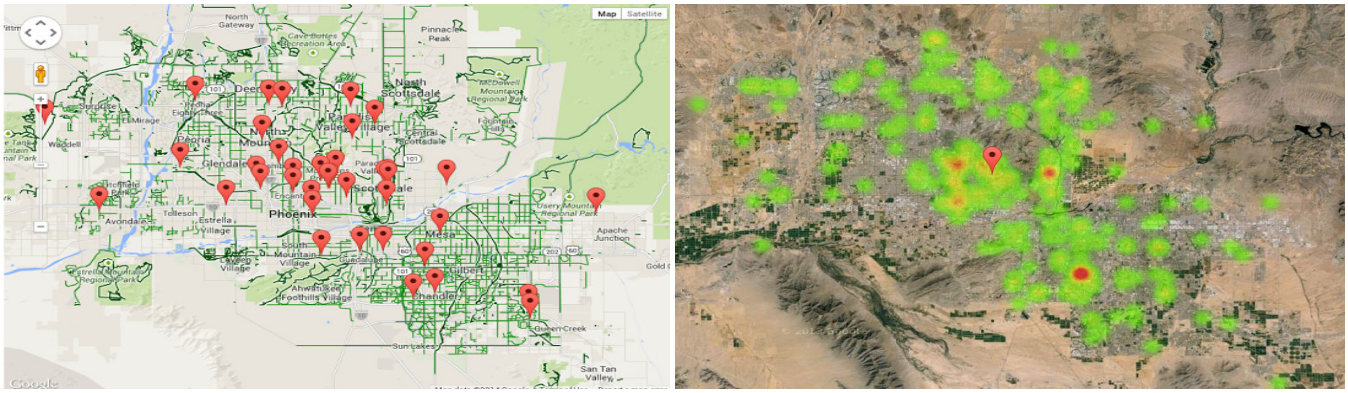

Figure 1: Heatmap and Geospatial Visualization

```
Chicken: [1, 'best', 1, 0],
Thai Pan: [7, 'divine, only, new, dish, wonderful, authentic, best', 4, 0],
Thai curries: [1, 'Chinese', 0, 0],
place: [3, 'dark, good, great', 2, 1],
restaurant: [2, 'next, clean', 1, 0]
```

Figure 3: Snapshot 2

positive or negative words. At this point, we have a list of all the good "features" (or nouns/noun phrases) associated with a particular business and all the negative features associated with the business. We sort the noun adjective dictionary by the number of occurrences of the noun and we pick the top fifteen most occurring noun/noun phrases (using the number of occurrences counter) and using the positive and negative counters, we make a determination of the success of the business–we say that a business is successful if we see that the number of positive noun/noun phrases associated with the top fifteen occurring noun phrases is greater than the number of negative noun/noun phrases. More specifically, the value $q = (positive/(positive + negative)) * 100$.

3.3 Location Analysis

Next, we derive values [b1 , b2 ,..., bn] where b1...bn indicate how positive the reviews of the n neighboring businesses of P are. To derive the neighboring businesses, we can calculate the distance between business P and bi using haversineâĂŹs formula for the great circle distance between two points, used widely for navigation. Then, we find the positivity of reviews bi for each neighboring business Pi. In order to integrate q and [b1, b2,...,bn] to derive a value S which is the success factor for P, we use a kernel smoothing function to calculate the success factor. A smoothing kernel defines a set of weights for each x :

$$f(x) = \sum_{i=1}^{n} W(x) * S(x)$$

The weights, Wi(x) are described by a density function with a scaling parameter that adjusts the size and form of weights near x. In order to compute the weight density, a gaussian function, on the distance from the business to be studied (P), is used. This means that businesses close to P have a higher impact on the success compared to businesses farther

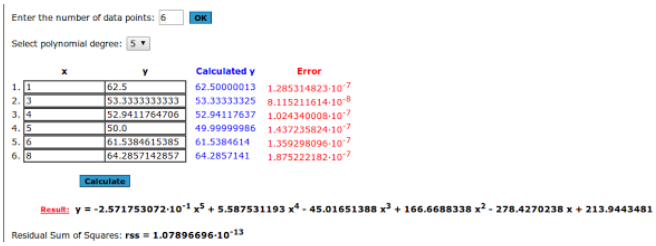

Figure 4: Regression model output

away. Also, by adjusting the parameters of the gaussian, we can change how quickly this effect declines with distance. In the final step we combine the sum of the contribution of neighbouring businesses with the success factor of the business in attention (shown in Figure 5)

$$S(x) = \alpha * S_i + \beta * \sum_{i=1}^{n} W(x) * S(x)$$

$$where\ W(x) = a * \exp(-(x-b)^2/(2*c))\ and\ \alpha = 1 - \beta$$

3.4 Regression Model

We used a polynomial regression model to derive the future success factor for a business. After taking a closer look at the Yelp dataset, we determined that there are reviews starting from approximately January 2007 to January 2014. We determined the success factor for each six month interval between this time period using the method described above. Then, using the data points, we came up with a polynomial regression fitting and using this function, we can predict the future success for a given business. For example, these are the points we derived for the Thai Pan Fresh Exotic Cuisine (here, when x = 1, we would be dealing with the time interval between January 2007 to July 2007 and the success factor during this interval is calculated to be 62.5; if x = 2, the time interval would be July 2007 to December 2007 and the success factor is 53.3 and so on).

This is the graph of the function above :

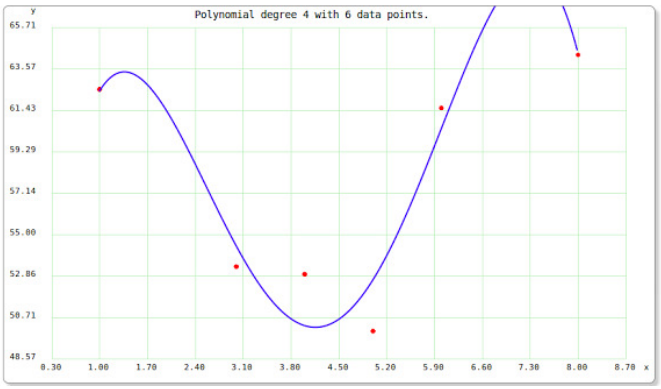

Figure 5: Graph of the Regression Model

3.5 Fishy Users

Recently, Yelp has been accused of taking bribes for assigning favorable reviews to certain business. In order to achieve this, Yelp (or other companies) have been hiring anonymous individuals who give highly favorable (or negative) reviews to a certain set of businesses. We can distinguish between these "fishy" users and normal users by noticing that such users are most likely to post a large number of reviews for many businesses in a short period of time. In contrast, a normal user would have a widespread timespan over they would post reviews. Another pattern for finding fishy users is to check for multiple users using similar "grammar" this would allow us to identify whether multiple accounts are being used by the same user to post reviews–which might indicate "fishy" behavior. Unfortunately, both these problems are quite challenging to solve so we will only address the first task. In order to detect users that post a large number of reviews in a small period of time, we looked at the frequency of reviews over time. If we find a single peak which falls rapidly, it means that the user has given a large number of reviews in a small amount of time which is a cause for suspicion as stated before.

The primitive method for finding extremes is to use the zero-derivative technique. However, noisy graphs have accidental zero-crossings of the first derivative and this produces false results. A low-pass filter can be used to smooth the graph, but that would result in a significant loss of data. Instead of trying to find the mathematical maxima, we turn to the concept of "peaks" and "valleys". A peak is simply the highest point with a range of lower points on both sides. We use a peak threshold to get rid of rapidly occurring small peaks due to noise. The graph in Figure 6 shows the peak detection algorithm working on a sample graph (generated using sines and normal random distribution)

4. CONCLUSION

In this paper, we have described an algorithm that can be used to predict future success of a business while taking into account internal and external factors. Our method for determining the future success factor for a particular business involves first splitting all of the businesses' review text according to the dates when they are posted and place them into appropriates "buckets" – where the buckets are six month time intervals between January 2007 and January 2014. Then, we parse the review text for each bucket of reviews and extract the number of positive nouns to derive a value q for the business and a list of values q1...qn for the neighboring businesses. Then, we combine q and q1...qn to come up with a success factor for the business. Using the success factors calculated for each bucket, we come up with a polynomial regression model and use that to predict the future success factor. Our method, which accurately predicts the success of a business can be applied to many other situations–for example, we could use the same general idea to predict the future success of a student and this would be invaluable information for colleges as they make their admissions decisions. In addition, we can apply the idea to calculating the success of scientists based on their collaboration network–how likely is it that a scientist will succeed based on the success of the scientists that they concur with?

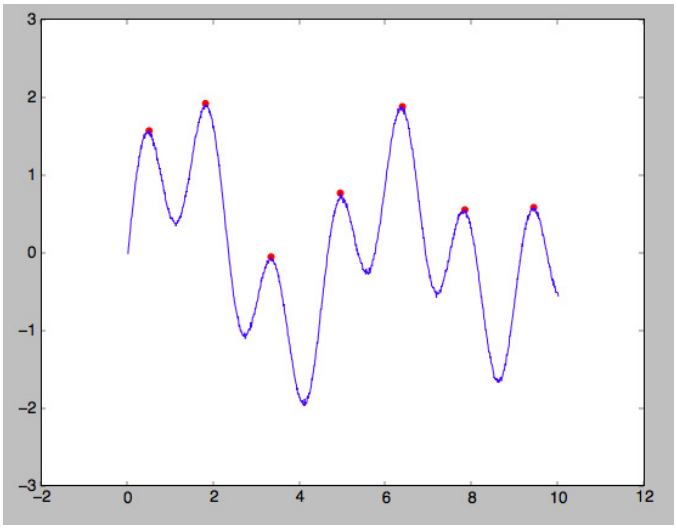

Figure 6: Peak Detection Algorithm

5. FUTURE WORK

Our work can be extended in many ways. Our algorithm for finding fishy users, which involves finding peaks (which represent high frequency of posting reviews), is insufficient to accurately determine fishy users. While our peak detection algorithm works quite well, it is not easily applicable to our case. It is not clear, for example, how the peak for one user measures against other users and what the general frequency for giving reviews is – it's possible that a user might not give reviews regularly but is overly active, posting many reviews during some weeks and not so many during others. In this case, the detected peak would be a false positive. Unfortunately, due to time restrictions, we couldn't get any significant results by applying this algorithm to the Yelp dataset.

Another interesting approach to finding such users is to consider the fact that most of these invalid reviewers will be posting a lot of favorable reviews but are not likely to have visited the businesses themselves. We can use the "check-in" data from Yelp dataset, which contains the timestamps for people visiting a certain business to determine if a reviewer has actually tested a business before reviewing. This com-

bined with an unusually high frequency of giving reviews would give much more accurate results with a minor chance of false positives.

We also did not get a chance to test the validity of our results; one way to test the accuracy of our method would be to derive a regression model that uses data between January 2007 and July 2013 and use this model to predict the success factor for the business in January 2014 and see how close or far this prediction is from the actual data. Also, categorizing businesses before deriving the success factor would improve the results of our method significantly. For instance, a hardware store would not have as much of an impact on the success of a Chinese restaurant as another Chinese restaurant in the vicinity would–so by categorizing the businesses and taking into account the influence of businesses of the same kind, we can come up with an even better estimate for the future of the business.

6. REFERENCES

[1] Semantic analysis. `http://en.wikipedia.org/wiki/Semantic_analysis_%28linguistics%29`.

[2] V. H. BHood and J. King. Inferring future business attention.

[3] R. B. G A Miller and C. Fellbaum. Introduction to wordnet: An on-line lexical database. 1990.

[4] M. Hu and B. Liu. Mining opinion features in customer reviews. *In Proceedings of the National Conference on ArtiïňĄcial Intelligence*, 2004.

[5] A. J. Smola and B. Scholkopf. A tutorial on support vector regression, statistics and computing.