Yelp Data Analysis

Bharat Mukheja(bkmukhej)
Mitkumar Pandya (mhpandya)
Rachit Thirani (rthiran)

1 Introduction

Yelp is a website which publishes crowd sourced reviews about local businesses (Restaurants, Hotels, Department Stores, Night life, Home-Local Services, Cafes, Automotive, etc.). It provides opportunity to the business owners to improve their services and users to choose best business amongst available.

Currently Yelp just lists all user reviews and provide average ratings for particular business. It is very difficult for business owners to go through list of reviews and carve out important information out of it. Also it does not provide advanced analytics to business owners to grow their business and improve their services. Our solution will focus on providing advance analytics from yelp data to help existing business owners, future business owners and users.

2 Overview

Here are the main business problems we are targeting in our project.

What does the users review say about a business, specifically how are user sentiments divided in terms of reviews for a particular business. In terms of business how business owner can evaluate their performance based on user reviews over a period of time.

In addition to reviews we want to analyze the Check-In data to understand the headcount patterns at particular business. In addition to providing the most favorable times and headcounts at those times for the business, it would allow us to supplement such data to the sentiment analysis to form a better recommendation system.

What does the user's review history and his social network tell about him and his behavior and how trust worthy are his reviews. In terms of business how can we detect fake users.

3 Dataset description

Yelp dataset is subdivided into following categories and is available in json format.

Business: A local body listed on yelp like Restaurants, Hotels, Department Stores, Night life, Home-Local Services, Cafes, Automotive etc.

User: A person who has registered on yelp who is writing reviews about different business after vising them or a person who is using yelp reviews to choose business.

Review: It is text written by user after vising business about the over-all experience. I is also a numeric representation (out of 5) to compare it with other business.

Checkin: The checkin information for a particular business which is categorized on hourly basis over a week of time.

Tips: Tips is text written by user to identify specialties of a particular business kind of do's and dont's.

4 Business Problems

4.1 User Review Analysis

Our first business problem is related to user's review data. We aim to analyse from the business perspective what does users say about their product or service etc. and from the user's perspective what other users have to say about that particular business so as to try that business or avoid it.

4.1.1 Review sentiments

The first step in analyzing review data we first classified sentiments of the user reviews in positive and negative sentiments. For this task Yelp has provided User Reviews data in the form of json object. In this task to reduce the computational complexity we have considered data of only Phoenix (AZ) area and 'Restaurants', 'Hotels' and 'Night Life' as business categories. So we first extracted all the business from the above mentioned categories which are located in Phoenix(AZ) and all the user reviews related to this businesses.

4.1.2 Logistic regression model for sentiment classification

For the classification of review data we used gensim Doc2Vec technique with Logistic Regression model. We extracted review text from yelp reviews data and converted the text to Doc2Vec format which consists of every separated word from the text lines along with train and test labels e.g. TRAIN_POS_i or TEST_NEG_i etc. where i is the line number. Next, we trained a Logistic Regression classifier on IMDB movie reviews data and Twitter data. Using the trained model we classified our yelp reviews data in positive and negative reviews. Since the training and testing data are from completely different medium, we expect the accuracy of around 60-80% in the classification.
Here is an example of classified positive and negative sentiments.

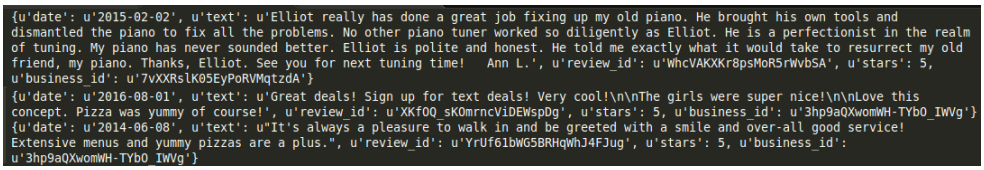


Figure 1: Reviews with positive sentiments.

We then evaluated business review counts, business sentiments (positive or negative) and business stars data as shown in the diagram below.

{u'date': u'2014-09-08', u'text': u"I didn't care for this pizza. The sauce is gross and the crust taste like Bisquik. It's also kind of on the expensive side for having to take it home and bake it yourself. I would rather pay the same price for an already cooked pizza. The cinnamon dessert was good though. I will not be going back. This pizza hurt my stomach. I will stay with RKidds down the street.", u'review_id': u'QU-16IzRnTR72_WpN9B3IQ', u'stars': 1, u'business_id': u'3hp9aQXwomWH-TYbO_IWVg'}
{u'date': u'2012-03-25', u'text': u'Im not sure what the hell people are smoking before they come here to give this place anything more than a 0 star... I literally hated this place. I usually go to the one in Tempe, which is very nice.. Okay so this place blows... the staff is running around like animals not tending the guest and its just chaotic... so we start and the lane isnt work, its not starting the balls dont come back when we throw them and the gates are randomly coming down before we throw the ball, or before the ball hits the pins? WTF? The food is expensive... as expected since the others are.. but seriously we paid for unlimited bowling for 2 hours and stayed for like one hour. Also all the balls... well unless you have little kids thumbs, or giant hands, they wont fit the balls.. which in my opinion is like the # 1 thing when bowling... having a ball that fits... oh well. The lanes are oiled to hell and you feel like youll slip every time you go down to bowl.. thats awesome.. pay 15.00 to fit my fingers into smaller holes, break my neck while bowling, have the ball get blocked by a gate, and not get served? wow.. i know ill never go back to this one at least.', u'review_id': u'gPranflF6kO_L08iQWpKkw', u'stars': 1, u'business_id': u'6pXwPMO871ir8ZnmyBbLEg'}

Figure 2: Reviews with negative sentiments.

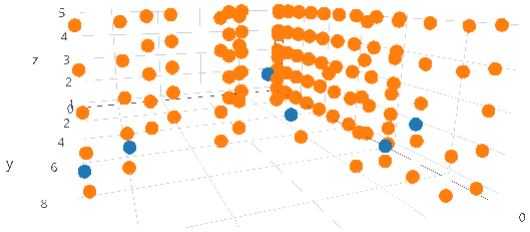


Figure 3: Business sentiments vs stars vs review counts.

4.2 Timeseries analysis on sentiments

Since positive and negative sentiments only are not sufficient for a business to evaluate their performance over time, we further tried to analyse the trends in positive and negative sentiments over time.
For this task we applied timeseries analysis for a particular business id "OgJ0KxwJcJ9R5bUK0ixCbg". First we extracted review date and review counts (positive counts and negative counts) from the reviews json based on the date of reviews. For simplicity we replaced dd part from yyyy-mm-dd as 01 so as to make month and year wise data. Below is the plot of year wise business sentiments (positive or negative) counts.

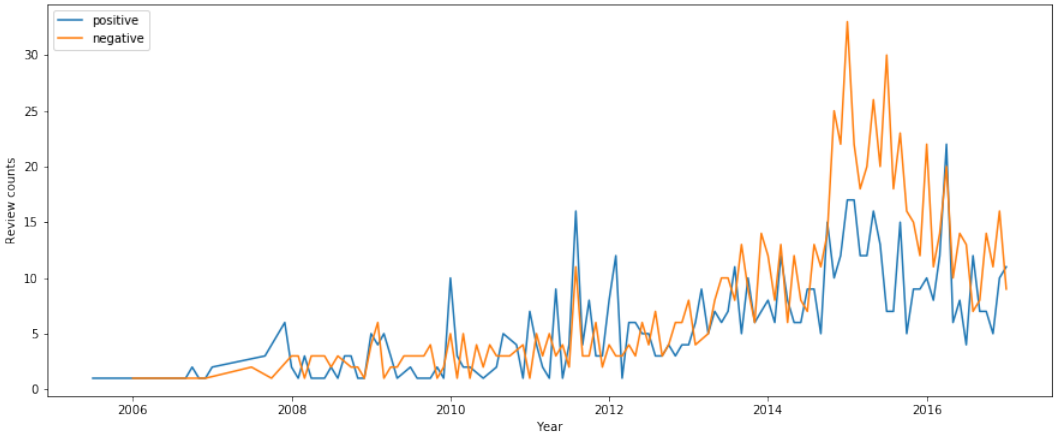


Figure 4: Positive and negative sentiments over time.

Next, we tried to decompose the timeseries data so as to check trend and seasonality in the review sentiments over time. The plots below shows the trend in positive and negative sentiment counts over time.

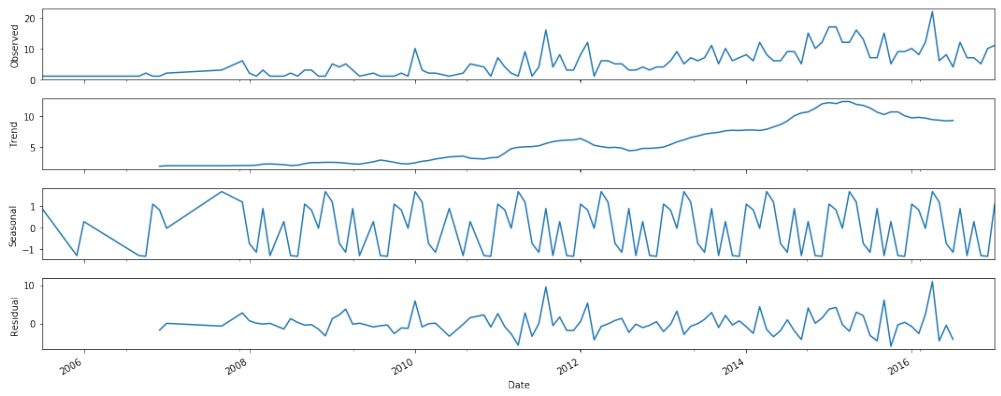


Figure 5: Positive reviews trend over time.

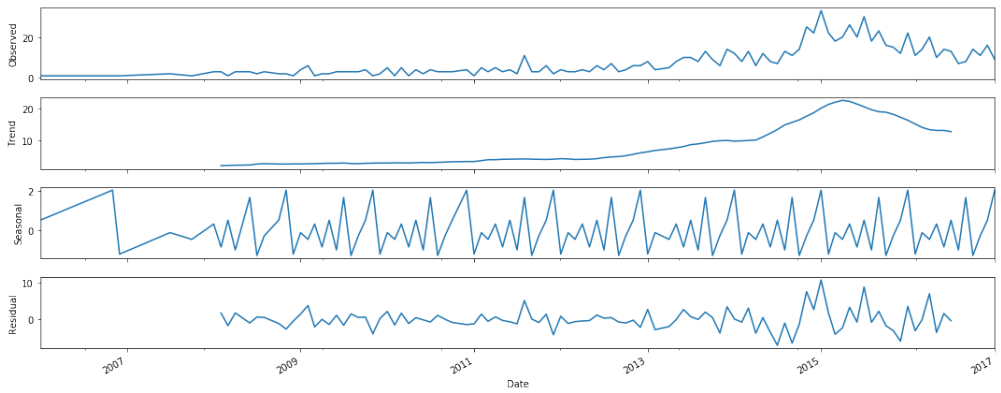


Figure 6: Negative reviews trend over time.

As we can see the trend in positive and negative reviews over time is increasing for this particular business and based on this information business company can evaluate their performance and improve over time. In this example we demonstrated how a business can not only from the review sentiments but also from the review sentiments trend and seasonality identify the user's opinion for their business and can evaluate the performance of their business and identify how they can improve their performance for the coming period. From the user's perspective this information shows how consistent or good is a business's performance over time.

4.3 Check-In Analysis

Our second business problem is related to user's checkin at various businesses. We aim to analyze from the timing perspective what the attendance at a business say reflect about the mass favorite and what business is opted by the users at a given time and for a particular category, so as to realize the crowd favorites. As the first step in analyzing checkin data aggregated all the checkins in a single continuous series. Yelp has provided the checkin data in the form of a json object. In this task to reduce the computational complexity, we have considered data of only Pheonix(AZ) area and 'Restaurants', 'Hotels' and 'Nightlife' as business categories. As done in the review analysis, we extracted all the business for Pheonix from the above categories and all the checkins related to these businesses. The hourly collection of checkins made it possible to accumulate all the checkins within a span

of week with each hour representing a data point. This provided us with 24*7 = 168 observation points.
For the stationarity analysis, we used Augmented Dickey Fuller Test and Rolling Statistics plot, for trend and seasonality analysis, we used decomposition plots and for fitting a prediction model, we used an ARMA model. Since, just knowing a single business's headcount isn't sufficient, we went ahead and created an api which gives user the most happening(or solitary) place in a given area, based on choice. We are also working on combining the review, trend and desirability of a place to augment the recommendations.

4.3.1 Data Collection and Parsing

Data is available as a day-time count json file, and really needs to be in a time series format for any tsa analysis. For this, we first aggregated the series for all nightlife avenues within pheonix into a single business.
Note: We'll show below the plots and analysis only for Nightlife category, but similar has been done for others. As expected the checkins showed peaks at midnight because thats the most favorable time for nightlife, and Weekend had higher crowd.

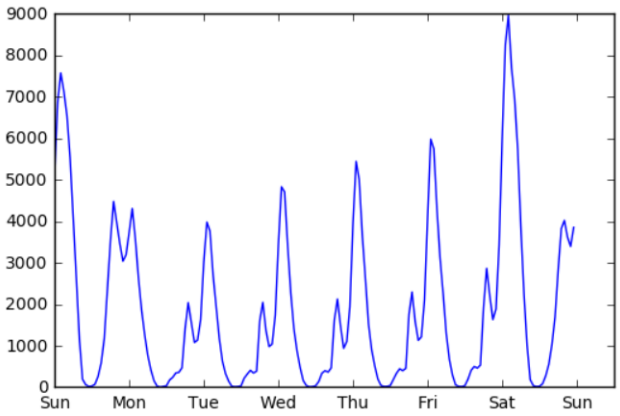


Figure 7: Checkins for Pheonix Nightlife Avenues

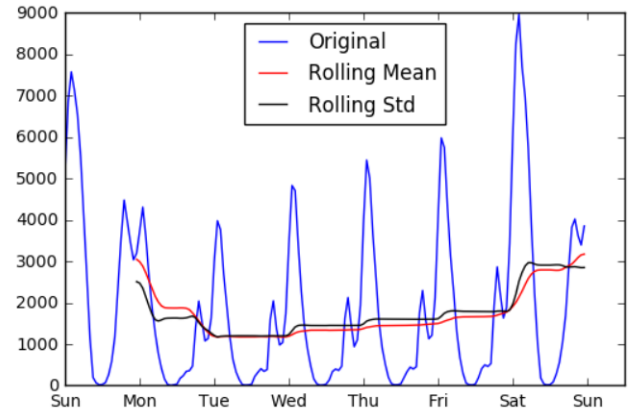
4.3.2 Stationarity Test

The graph above shows the number of checkins per hour starting from Sunday 00:00 AM and ending at Sat 23:00, making a total of 168(i.e. 24*7 observations.). The data is mostly periodic. This shows a good possibility of seasonality in the series. We'll take a step-by-step approach towards analyzing the series. We first check the series for stationarity. A time series is expected to be stationary for time series models to work. Also a non-stationary time series has more chances of showing an unexpected behavior in the future. More research has been done on stationary series which makes it the more favored option as well. Fortunately, there are methods to convert a non-stationary time series to a stationary one. A stationary time series has the following three properties - Constant mean Constant Variance A time independent autocovariance

The formal method to check stationarity are -
1. Rolling Statistics plot
2. Dickey-Fuller Test

We'll do both of them below.

The rolling mean and std are mostly constant but they increase for Saturday and Sunday with a maximum on Sunday 00:00 am(i.e. weekend night). This shows a lot of checkins




```
Dickey Fuller Test Results:
Test Statistic                  -5.112506
p-value                          0.000013
#Lags Used                       9.000000
Number of Observations Used    158.000000
Critical Value (5%)             -2.880013
Critical Value (1%)             -3.472431
Critical Value (10%)            -2.576619
dtype: float64
```

Figure 8: Rolling Mean and Std for checkin data

during the weekend midnight. As such no trend or seasonality can be noticed from the visual plot, but nevertheless the trend discovery would be made in the following section. Also, the test statistic is way less than critical values meaning the series is stationary.

4.3.3 Decomposition

By doing decomposition, trend and seasonality of a series are modeled separately and the residuals are shown. It gives us a clear picture of seasonality.

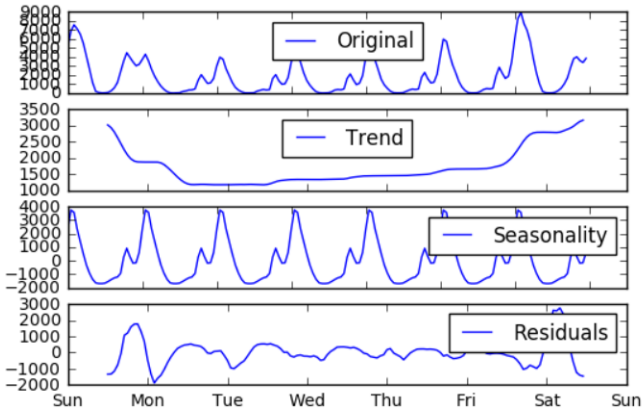


Figure 9: Checkins over time

The trend is linearly increasing but there is a fixed seasonal component present. When detrended and deseasonalised residuals are checked for their autoregressive component and moving average, the below acf and pacf plots are recovered.

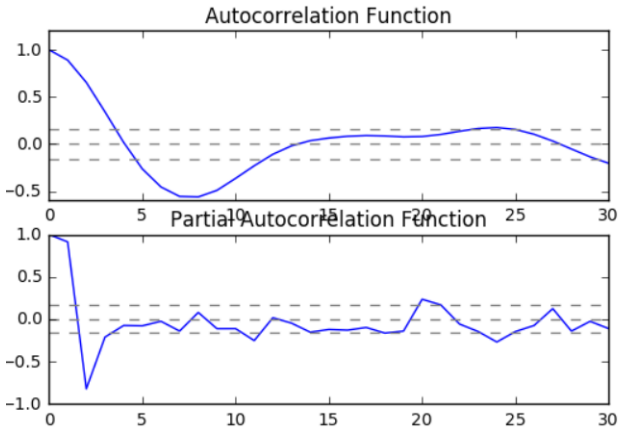


Figure 10: Classification and Regression Tree Plot.

4.3.4 Model

Looking at above graph, the p and q value for ARIMA are looking as 4 and 2.

$$AR(p) : (1 - \phi_1 L - \phi_2 L^2 - \phi_3 L^3 - \phi_4 L^4) x_t = \epsilon_t$$

$$MA(q) : x_t = (1 + \theta_1 L + \theta_2 L^2) \epsilon_t$$

$$ARMA : \phi(L) x_t = \theta(L) \epsilon_t$$

where L is the lag polynomial.

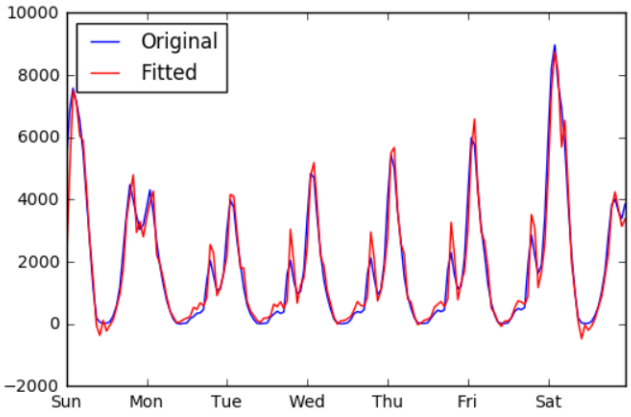


Figure 11: AR-Model

Finally we plot the forecast for the next 32 hours. This shows that the fore-casted checkins for day beyond Sunday are showing a curvical graph. This demonstration shows how a user, not only from the reviews, but also from previous headcounts analyze the possible crowd at an avenue and identify desired location. From a business perspective this shows how much gathering they can expect on a given time in future.

4.4 Anomaly Detection

This business problem aims at finding not normal users accounts based on their behavior. Yelp offers features of social networking website and its uncommon to have no friends

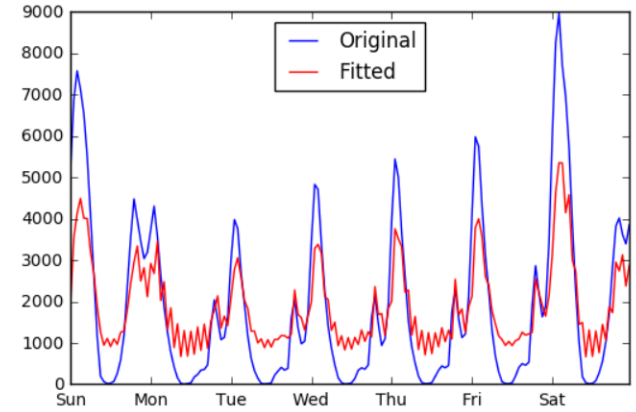


Figure 12: MA-Model

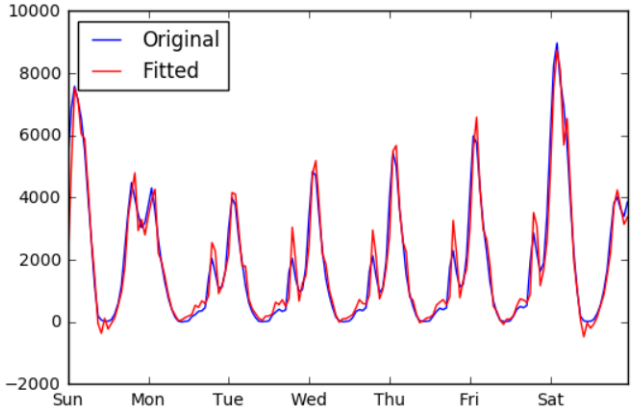


Figure 13: Combined Model

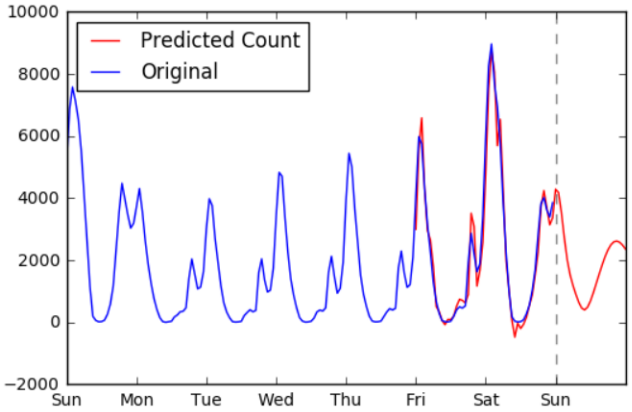


Figure 14: Forecast

unless its possibly a fake account.

the first step was to find out all the users who has no friends as they are potentially fake accounts, just created to promote the business. The result was quite surprising as 40% of

the users has no friends.

Based on these the findings, we used used different techniques to find out some more abnormal behaviors of the users.

In the second step we tried to find out how many users comment for businesses on different states and made a list of all the users who had review comments for more than three states as there are high chances they are just trying to promote different businesses. Out of the first 1000 users, only 3 users has comments for businesses spanning more than 3 states.

The third step was to find out how many users who had no friends had weird ratings. the measure which we used for weird rating was based on the calculating total rating which was less than or greater than the given stars to the business +/- the standard deviation of the ratings give to the business by different users. if more than 1/4 of total ratings given by the users were categorized as weird, the user was kept under the scanner for giving weird ratings.

Then we tried to find out the users whose average ratings were more more than 4.5 or less than 2.5, as its not possible that the user always get a good service or always get bad service at the business. The results were quite surprising as 63% of the first 1000 users who has no friends gave average ratings more than 4.5 or less than 2.5 , which forces us to keep such users in the scanner.

Finally we tried to check mow may such users comment on the same business again and again. surprisingly we didn't find any user who was rating the same business over and over.

After the analysis we kept all such users in the list scanner list of having a fake account, but as we know it is difficult to predict the userś truthfulness, but its not bad to keep them under the lens and detect their behavior for a period of time to check out if their reviews are reliable or not as their behavior is not normal.

5 Conclusion

While doing the review sentiment analysis we found out that even though labeled and unlabeled data were completely from different domain, gensim doc2vec along with logistic regression model still classifies yelp review data correctly in positive and negative sentiments. The timeseries analysis on sentiment counts over a time helps us in identifying trend and seasonality in user reviews for a particular business.
These two approaches helps businesses evaluate their performance over a time as well as users also can get a grasp of the business performance before obtaining service from that business.
Since this is a novel technique that applying timeseries analysis on user sentiments, we conclude that our model can be used as a proof of concept in new research approaches in this field.

For forecasting headcount on a business avenue, we utilised the checkin data from yelp. By applying time series analysis and utilizing the ARMA model, we could make a prediction with almost perfect fit, even from checkin data of a week only. This is a novel approach to analysing headcount and could be used as a Proof of Concept in devising solution to a more practical problem of finding most happening places in a location.

There are a lot of factors involved in detecting not normal behavior of the user. By utilizing some of the techniques and introducing various parameters on the given data set, we could see that there are around 36% of the users whoś reviews need to kept under the

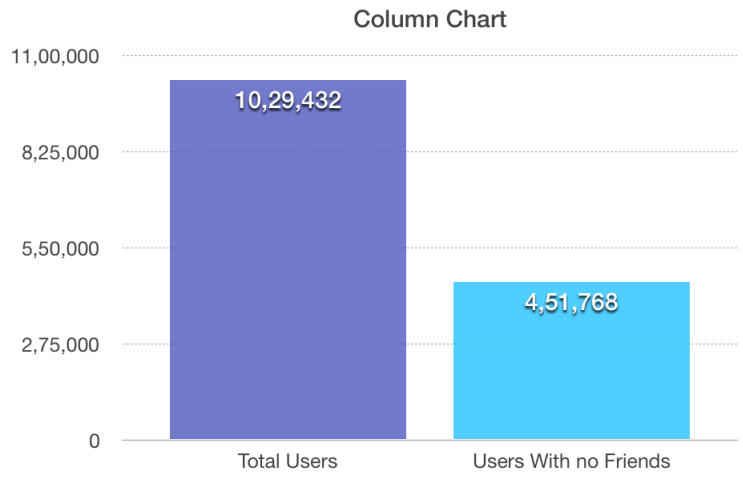


Figure 15: Users with no Friends vs Total Users

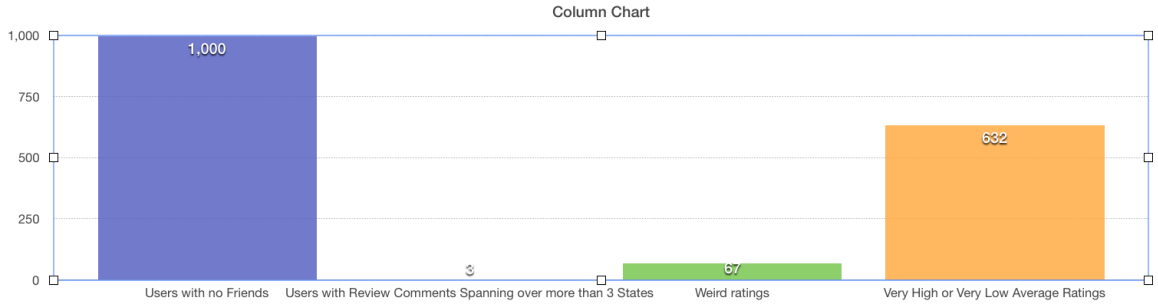


Figure 16: Behavioral Analysis

scanner and with the help of more information, we an safely categories these users as fake and their reviews as not trust worthy.

6 Github code repository

`https://github.ncsu.edu/mhpandya/BI-Capstone`

7 References

[1] GENSIM Doc2Vec Model `https://radimrehurek.com/gensim/models/doc2vec.html`

[2] StatsModels package `http://www.statsmodels.org/devel/generated/statsmodels.tsa.arima_model.ARMA.html`

[3] TSA forecasting
https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/

[4] Yelp Challenge Project Report, Tingting Zhang, Yi Pan, University of Washington
http://courses.cs.washington.edu/courses/cse544/13sp/final-projects/p09-tingtinz.pdf

[5] Class Notes