Analyzing User Reviews on Yelp Dataset

Cavin Baptist Dsouza
Department of Computer Science,
Indiana University Bloomington,
Bloomington, Indiana
dsouzac@indiana.edu

Dwayne Dexter Dsouza
Department of Computer Science,
Indiana University Bloomington,
Bloomington, Indiana
dsouzad@indiana.edu

Melita Dsouza
Department of Computer Science,
Indiana University Bloomington,
Bloomington, Indiana
dsouzam@indiana.edu

ABSTRACT

Yelp consists of crowd-sourced reviews about local businesses. Different businesses publish their listings which allows users to rate them and write reviews based on their experience. With Yelp having more than a billion reviews in raw format, it becomes difficult for businesses to find relevant information from such reviews and improve their business. In recent times, user reviews have played a crucial role in understanding the parameters that affects the performance of a business in the market. In our work, we perform sentiment analysis using the AFINN word-list based approach and Liu Bing's Sentiment Lexicons, and compare their performances based on the accuracy in predicting the star ratings of user reviews. We perform POS tagging to get features and trends during Christmas season. We intend to use the yelp star rating for a city-restaurant name combination and thereby predict the accuracy of AFINN model in classifying reviews as positive or negative. We perform logistic regression and anova testing to depict the significance and priority of certain predictors over other variables.

KEYWORDS

yelp reviews, sentiment analysis, word cloud, restaurants, lexicons, AFINN, prediction.

1. INTRODUCTION

In recent times, reviews and star ratings have played an important role in impacting a user's decision on having a meal and shopping for accessories. Yelp contains millions of reviews given by users in raw format. Businesses today can leverage this raw data and find out their market value by analyzing their key features. The Yelp dataset consists of 4.1M reviews and 947K tips by 1M users for 144K businesses. With the help of this dataset, our research mainly focuses on finding interesting seasonal trends for businesses. We predict the star ratings of reviews with three classification models, namely, Naive Bayes, Support Vector Machines based on review text alone. We perform logistic regression testing with cross validation using parameters such as number of positive words and negative words in a review along with review length.

2. DATASET

The Yelp dataset is freely available on the Yelp website [1] and requires registration and acceptance to Yelp's terms of use. This dataset is unrestricted for use as long as it is used for academic purposes only. The data collection includes reviews that were recommended solely by Yelp. The textual dataset available online is in json (JavaScript Object Notation) format and can be readily converted into csv format using a python script. The file is dependent on the 'simplejson' package which is available freely. We have first computed a list of restaurant related business id's so as to filter out reviews related to 'Restaurant' category only. We remove unwanted elements such as URL's, abbreviations, double spaces and non-ASCII characters using a python script to make the data accessible to visualizations.

yelp_academic_dataset_business.json

```
{
    "business_id":"encrypted business id",
    "name":"business name",
    "neighborhood":"hood name",
    "address":"full address",
    "city":"city",
    "state":"state -- if applicable --",
    "postal code":"postal code",
    "latitude":latitude,
    "longitude":longitude,
    "stars":star rating, rounded to half-stars,
    "review_count":number of reviews,
    "is_open":0/1 (closed/open),
    "attributes":["an array of strings: each array element
  is an attribute"],
    "categories":["an array of strings of business
```

```
categories"],
   "hours":["an array of strings of business hours"],
   "type": "business"
}
```

yelp_academic_dataset_review.json

```
{
   "review_id":"encrypted review id",
   "user_id":"encrypted user id",
   "business_id":"encrypted business id",
   "stars":star rating, rounded to half-stars,
   "date":"date formatted like 2009-12-19",
   "text":"review text",
   "useful":number of useful votes received,
   "funny":number of funny votes received,
   "cool": number of cool review votes received,
   "type": "review"
}
```

3. RELATED WORK

There has been, unsurprisingly quite some research in the area of sentiment analysis. We attempt to create an SVM model using a linear kernel, similar to one mentioned in the report by Yanrong Li and Yuhao Liu [2] in order to predict the star ratings of user reviews. Our prediction uses review text as the only predictor variable. This is then created into a document term matrix which is fed as input to the SVM algorithm. We also use the sentiment analysis described in Hu and Liu [3], that uses sentiment lexicons consisting of positive and negative words.

4. PROPOSED WORK

We use the AFINN word-list based approach to perform sentiment analysis on the review texts and we predict the accuracy of the model as a measure of its ability to correctly classify a review rating. We analyze the ambience and cuisine of restaurants during Christmas season by using POS tag and observing the most frequently words in the list of nouns. We generate word cloud for one-star and five-star reviews in order to visualize the common words used in those type of scenarios. We use logistic regression with cross validation [4] on parameters such as review length, positive words and negative words in order to predict star rating.

5. DATA EXPLORATION

After the data is processed and saved in csv files we use statistical methods and visualization techniques to analyze the data.

What are the most frequently used words in a 1 star and 5-star review?

We filtered the reviews for 1-star and 5-star ratings and using the nltk package we extracted only the adjectives as they make a significant prediction of the sentiment of a review. Fig 1.1 and fig 1.2 show the word clouds generated for 1-star and 5-star reviews respectively using Python word cloud generator script.

Fig 1.1 Wordcloud for 1-star review

Fig 1.2 Wordcloud for 5-star review

Do star ratings matter for the success of a restaurant?

We consider review counts and stars as the factors for a business's success. We check how these features correlate with each other. We plot a map where the data is filtered on longitude and latitude and the view is filtered on average review count.

Fig 1.3 Average review counts vs location

We plot a map where data is filtered on location and view is filtered on average star ratings.

Fig 1.4 Average stars vs location

From fig 1.3 and 1.4, we can say that the average star rating for every city is more or less similar but the average review count tells us that Las Vegas has a greater number of average review counts compared to all the other cities. Thus, average star rating does not seem to be a good predictor variable in determining success of a restaurant.

Analyzing Seasonal Trends

We have considered one of the most popular season, that is, the season of Christmas in order to predict certain trends during years 2011-2016. We plot a Year vs Review count bar chart for each year from 2011-2016. We display only the top 10 cities based on total review count.

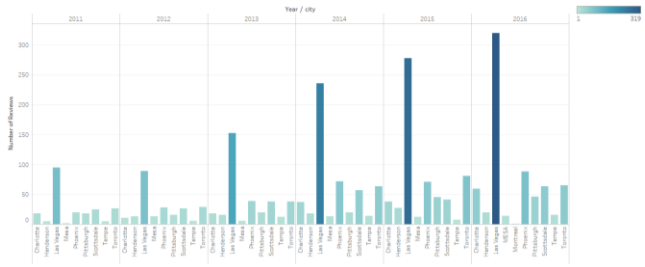

Fig 1.5 Christmas reviews during 2011-2016 for top 10 cities

From Fig 1.5 we can observe, for e.g. that reviews for restaurants in Las Vegas increased every year except for the year 2012. Also, restaurants in Montreal had no worthwhile mention in any review until the year 2016.

Most spoken about topics in Christmas reviews

We plot a word vs freq bar chart for each year to show most frequently used words in reviews during Christmas. The data is filtered on frequency which ranges from 80 to 904. The view is filtered on top 35 words.

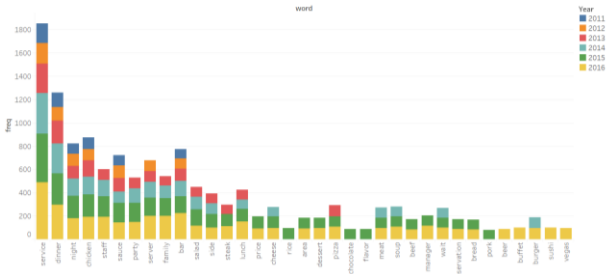

Fig 1.6 Frequently used words during Christmas 2011-2016

From Fig 1.6 we can see that the words {“service”, “dinner”, “night”, “chicken”, “bar”, “party” and “sauce”} are the most frequently used words in the years from 2011-16 during Christmas. In recent times (2015-16) it can be observed that the words “wait” and “reservation” are frequently used indicating a possible surcharge of customers visiting restaurants during Christmas. Also, Vegas is the most talked about city in the reviews.

6. SENTIMENT ANALYSIS

Sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.

We will be performing sentiment analysis using AFINN [5] word-list model to compute the overall score of the review. AFINN is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive). The words have been manually labeled by Finn Årup Nielsen in 2009-2011.

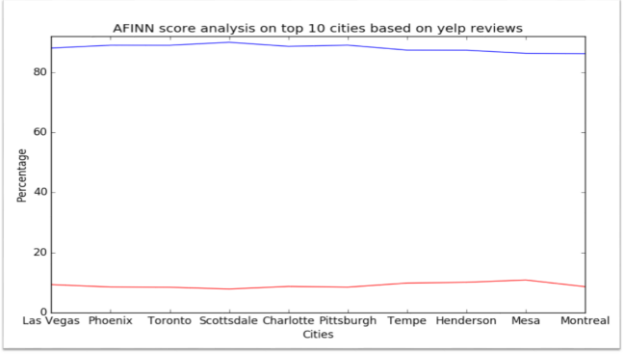

Fig 1.1 AFINN score analysis on top 10 cities based on yelp reviews

Fig 1.1 depicts the average AFINN score/review of the top 10 cities based on total number of reviews. As observed, the AFINN score/review remains fairly constant for both positive and negative score ratings. We test the accuracy of the AFINN model by predicting the sentiment of the review and assuming true sentiment class as discussed in 5(c). Overall accuracy was computed as approximately 70%. The other approach of sentiment analysis discussed is Liu Bing's lexicon method. It uses a dictionary of positive words and negative words and sums up the count of each in a review by comparing each word of the review text with the positive and negative list. The net sentiment of the review is decided by max number of words belonging to a particular list. Firstly, the review text is tokenized into a set of atomic elements, after which we eliminate the presence of stopwords that exhibit no sentiment value to the text. We use the popular Porter Stemming algorithm to stem words such as 'read, reader, reading' into 'read'.

7. DATA MODELING

a) Naive Bayes:

A Naive Bayes classifier estimates the class-conditional probability by assuming that the attributes are conditionally independent, given a class label. To perform Naive Bayes classification on our dataset using Python, we use scikit-learn package. The multinomial Naive Bayes classifier is suitable for classification with discrete features (eg. word count for text classification). We transformed word frequency matrix to word occurrence matrix, i.e., a matrix of 0's and 1's, where 0 indicates absence of a term and 1 indicates its presence. In most real world data, there does exist correlation among the predictor variables and hence Naïve Bayes is not always the right choice of classification in such a scenario. We create a training dataset of 1,50,000 and a test dataset of 50,000 records. Table 1.1 depicts the confusion matrix after running the Naïve Bayes classifier.

	1	2	3	4	5
1	3925	390	352	680	358
2	1500	531	1156	1796	359
3	568	175	1121	5063	942
4	219	21	254	8564	5455
5	157	4	25	3501	12884

Table 1.1 Confusion matrix

The accuracy of the classifier is approximately 54.05%.

b) Support Vector Machine (SVM):

Support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. A SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. To perform

SVM classification on our dataset using Python, we use scikit-learn package. We run the SVM classifier for our training dataset of 1,50,000 samples and predict the star rating of our test dataset of 50,000 records. Table 1.2 depicts the confusion matrix after running the SVM classifier.

	1	2	3	4	5
1	6067	973	453	427	227
2	940	2827	987	973	253
3	613	593	4467	2893	520
4	333	240	2187	9867	1120
5	227	160	453	4520	7680

Table 1.2 Confusion matrix

The accuracy of the classifier is 61.81%.

c) Logistic Regression:

We implemented logistic regression with cross validation to create a model that predicts the sentiment (positive or negative) of a review using review length, positive words and negative words as features. We assume reviews with more than 3 stars to be regarded as positive and those less than 3 stars as negative as our true class of polarity. Accuracy of the model is approximately 60.47%.

d) ANOVA test:

We run an ANOVA test on the regression model: stars~ review_length+pos_words+neg_words. We see that review length, pos_words and neg_words have a significant impact on the star ratings. The F-value is very high for all three features and p-value is below 0.05 which indicates statistical significance. It is pretty straightforward that if the number of positive words for review length are greater the star rating will be higher and vice versa.

8. CONCLUSIONS

In our project, we have analyzed useful patterns and trends in businesses. We gathered a list of words which are most commonly related to 5 star and 1 star ratings. For restaurants, it is observed that average star ratings does not help determine its success accurately. Finding the seasonal trends for restaurants allows businesses to better prepare for special occasions.

Based on data modeling and analysis, we find that it is statistically significant to predict stars using review length, number of positive and negative reviews. Naive Bayes and SVM were able to predict the star ratings of the reviews with good accuracy.

9. FUTURE WORK

The analysis we performed considers only the restaurants category of the businesses and user review dataset.

Future work could include building a module by which all these trends and insights are communicated to businesses in real time so that businesses can get faster feedback. As of now the AFINN model works only on English and Danish languages due to which we had to restrict ourselves to reviews only consisting of those two languages. The sentiment analysis algorithm used in our project can't deal with sarcastic reviews. The detection of sarcasm can improve the model predicting sentiment polarity of a review.

ACKNOWLEDGMENTS

We would like to thank Prof. Vincent Malic and Ashley Dainas for their continued support and assistance throughout the length of the project.

TEAM CONTRIBUTIONS

Key accomplishments are listed here:

Cavin Dsouza

- Wrote Python Scripts for JSON to CSV generation, afinn accuracy, naïve bayes and svc classification, wordcloud generation, generating data for Christmas trends
- Equally contributed in editing the project report.

Melita Dsouza

- Wrote R Scripts for Logistic Regression and Anova Testing
- Used Tableau for visualizing and analyzing the data for the poster as well as the project report.
- Used SQL queries to merge multiple csv files in Tableau to generate Christmas trends.
- Equally contributed in editing the project report.

Dwayne Dsouza

- Researched on data pre-processing and sentiment analysis
- Equally contributed in editing the project report

REFERENCES

[1] https://www.yelp.com/dataset_challenge

[2] Yanrong Li, Yuhao Liu, Richard Chiou, Pradeep Kalipatnapu, "Prediction of Useful reviews in Yelp Dataset"

[3] Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews."; Proceedings of the ACM SIGKDD International Conference on Knowledge; Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle,;Washington, USA

[4]Yelp-review-analysis,@minimaxir, https://github.com/minimaxir/yelp-review-analysis

[5] Finn Arup Nielsen, "A new ANEW: evaluation of a word list for sentiment analysis in microblogs, Proceedings of the ESWC2011 Workshop, Volume 718