Black Box of Businesses

Prediction of star ratings on Yelp Dataset*

Sai Teja Chava
University of Southern California
Los Angeles, CA
schava@usc.edu

Soumith BSV
University of Southern California
Los Angeles, CA
bodappat@usc.edu

Deepthi Devaraj
University of Southern California
Los Angeles, CA
ddevaraj@usc.edu

ABSTRACT

Yelp provides a platform for users to post reviews and images on businesses and products. These reviews and images are an invaluable source of information for users to choose among numerous available options as well as for businesses to improve their business. Reviews have also significantly influenced the user shopping experience. Usually an on-line review consists of free-form text, image, certain business attributes and a star rating out of 5. Business rating is shown to have the highest influence in user's decision of choosing a business. Using the Yelp Dataset [1], our goal is to identify the features that influences the rating of a business. In this paper, we have used various machine learning models to analyze business attributes, reviews and different deep learning architectures to label images and predict the business rating.

KEYWORDS

Yelp 2018 Dataset, Data Mining, Natural Language Processing, SVM, Logistic Regression, Linear Regression, Random Forest, Deep Learning, Computer Vision

1 INTRODUCTION

More than ever before, businesses are emerging across the world expeditiously. This has opened up a lot of options for users to choose from. With advent of technology, websites like Yelp provide users a platform to post their reviews and images related to a business or product.

With millions of people across the globe visiting multiple businesses and posting reviews and ratings, they have become a deciding factor for the user in choosing a business. There is a need for businesses to know the key factors affecting their growth. Businesses have to be on top of their game to attract customers. Though businesses are able to get the value out of data to improve their business, the process to estimate the attributes that have contributed to the rating is quite hard.

Some of the previous works on yelp dataset include, predicting a business star rating using review text alone [7] by generating unigrams, bigrams and semantic indexing. Most of the other rating prediction tasks is by using a subset of business attributes [3] or using latent factor analysis of features [10, 12]. These works haven't exploited the data completely.

Though there is significant amount of work done in predicting the ratings of a business, not much research has been done on analyzing the impact of a combination of business features, reviews, tips and images affecting the rating of a business. With the growth of social media platforms like Facebook, Snapchat, Instagram in recent years, images have become one of the primary form of expression. Much of the work done on images was in a well-defined context having good meta information [6] and less work has been done on images with limited meta-data. Considering this, we have explored the characteristics of images and it's significance on ratings of a business.

*Github Link: https://github.com/ddevaraj/black-box-of-businesses

2 DATASET

For this project, we used the Yelp Dataset [1], which contains information on businesses, users, reviews, tips, check-ins and images. The business data has about 188,592 unique business IDs and 5,996,997 reviews with each business having 1 to 2,000 reviews. In addition to this, there are around 0.3 million images and 1,185,347 tips given by users to the businesses.

To identify the key factors that contribute to the rating of the business, we did some exploratory data analysis on business and review dataset. One of things found was words such as 'good', 'food', 'place' and 'great' have the highest frequency amongst all. The features considered in predicting the rating are the attributes in the business dataset, text from the review dataset, text from tip dataset and images. We have also made use of the check-in count for each business from the checkin dataset. Image dataset consists of meta information like resolution and some text information like caption describing the image. Captions were used to generate the ground truth using sentiment analysis for training various deep learning models.

To evaluate the model, we split the data into 80% training and 20% test. The accuracy of the model was evaluated by comparing the predicted rating with the actual rating for a given business.

3 METHODS

3.1 Baseline

To establish a baseline, we evaluated a simple model - Linear Regression that correlates the review count of a business to it's star rating.

$$BusinessRating = \theta_0 + \theta_1 * ReviewCount \quad (1)$$

This model gave an RMSE of about 1.99 on our test data set. We evaluated all our proposed solutions against this RMSE value.

3.2 Feature Engineering on Business Attributes

The Yelp Business dataset contains 52 attributes and we have used most of them in our models. We dropped the features that are specific to an individual business. In this section we discuss a few important features available in our business dataset and how they are encoded in the proposed Machine Learning models.

3.2.1 Location. To analyze the location data, we considered the latitude and longitude values of individual businesses and performed a k-means(k=15) clustering on them. We encoded all the businesses to specific clusters from 0-14.

3.2.2 Hours of operation. We one hot encoded the hours of operation from Monday to Sunday. For example, Monday is encoded as [1000000].

3.2.3 State. We encoded the states to a number between 0-69, where numbers represent the fifty states in United States of America and states of few other different countries.

3.2.4 Noise Level. We encoded the noise level for businesses based on the follow mapping: {quiet:0, average:1, loud:2, veryLoud:3}.

3.2.5 Parking. We assigned a number in the range of 0-6 for each business which corresponds to the different types of parking available. For example, if a business had two different types of parking(say Garage and Vallet parking), the number assigned was 2.

3.2.6 Review Count, Checkin Count. We assigned a total count of reviews and checkins for each individual businesses as separate features.

3.2.7 Ambience. We encoded the ambience of a business in the range of 0-9 based on the following mapping: {romantic:1, intimate:2, classy:3, hipster:4, touristy:5, trend:6, upscale:7, casual:8, divey:9}.

3.2.8 City. We encoded the cities in the range of 0 to length of unique cities , where each number corresponds to one of the different cities present in our data set.

3.2.9 BusinessAcceptsBitcoin, BusinessAcceptsCreditCards, GoodForKids etc. There are 15 of them which are binary and were encoded as 0 or 1. This encoding was obtained either directly from the data or converted to the appropriate format.

3.2.10 WiFi, Smoking, PriceRange etc. There are 7 of them which have more than two unique values and were assigned an appropriate value mapping for each business.

3.3 Feature Engineering on Review Text

Previous work done on predicting stars only from reviews[2] showed that reviews have a significant contribution in business growth. Reviews are a free-form text consisting excessively of capital letters, punctuation marks and other features. They can be exploited to analyze the user's opinion, sentiment and emotion. Mining these characteristics from the reviews may have an impact in the rating of a business. We approached this by introducing the below features in our prediction models.

3.3.1 Sentiment Analysis. VADER Sentiment was used to perform sentiment analysis on the reviews. On running VADER sentiment, each review produces an intensity value that determines the sentiment of the text. Polarity of the text was assigned to be either positive or negative. Neutral polarities were dropped. Since each business had multiple reviews, we grouped the review sentiments based on the business ID and determined the count of positive and negative reviews. This acts as the sentiment score feature in our model.

3.3.2 Other Features. Along with the sentiment, reviews can also determine the interest and emotion of the user. User reviews can have varying text lengths, repetitive words and punctuation. These characteristics can be extracted and added as features to our model. The features were determined after removal of stop words and punctuations excluding question marks and exclamation marks. Stemming of the words was performed using NLTK's Porter Stemmer. The features extracted after pre-processing the reviews are as follows:

(1) Average text length associated to each business ID
(2) Average number of '?' and '!' marks used in the review
(3) A TF-IDF vector of the number of occurrences of 500 and 200 most common unigrams
(4) A TF-IDF vector of the number of occurrences of 500 and 200 most common bigrams

We also extracted the positive count, negative count, average text length and average count of '?' and '!' in the text from the tip dataset. These were added to the values identified from reviews for each business.

3.4 Prediction using Machine Learning Models

Our aim is to identify the features affecting the rating of a business. We tried regression and classification models on the various combinations of features and compared them using Root Mean Square Error.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2} \quad (2)$$

The first approach we took was to reduce the feature set size from 52 by dropping columns based on correlation analysis[1]. The resulting feature set was then grouped based on similar features, for example, bike parking was combined with business parking. Latent variables were found on the grouped feature set by performing Principal Component Analysis [8]. By doing this, we reduced the feature set size to 10. We then trained a Linear Regression model with the above feature set. Training on this feature set was not efficient, as it performed worse than the baseline.

The second approach we took was to train the model using all the business related attributes after dropping columns based on correlation analysis[1] and by not performing any latent variable analysis. Linear Regression for our problem is in the form:

$$Rating = \theta_0 + \theta_1 * IsOpen + \theta_2 * Location + ... + \theta_N feature\_N \quad (3)$$

The performance of Linear Regression using this approach was better than the model trained using latent variables. Logistic Regression was also trained, using the same subset of features used in Linear Regression. We used the 'lbfgs' optimizer for this model and found that it's performance was not as good as Linear Regression. We then extended our work to other models like Support Vector Machines and RandomForest Regressor.

The above models though performed better than baseline, could be improved by adding additional information extracted from reviews. The earlier models were retrained by adding information such as unigrams, bigrams, sentiment score obtained from section 3.3 to see it's impact on the business rating. We initially trained the

[1]The correlation value was chosen to be 0.3

models with 200 unigrams, followed by 500 unigrams and found that the performance of the model trained on 200 unigrams was better.

We also incorporated the count of positive and negative photos for each business as features obtained in section 3.6 and trained our machine learning models to find it's significance. We also tried some simple neural networks with minimal tuning to see how the networks performed compared to other classic machine learning techniques.

3.5 Supervised Image Classification

The goal is to classify images as positive or negative. A positive image is one which has enough information describing it and is visually appealing. Due to a large amount of data, we cannot manually annotate all images as positive or negative. We need to train a classifier to achieve this task so that new images can be assigned the right label.

3.5.1 Generating Ground Truth using VADER Sentiment. In order to train a classifier, ground truth has to be generated since it is not associated with the dataset. One of the ways to go about generating the ground truth is to look at the meta information associated with the data. The meta information associated with the images include photo_id, business_id, caption and label. Photo_id or business_id cannot be used to generate the ground truth as they uniquely identify an image and the business it belongs to. They do not provide any information about the image itself.

On the other hand, labels cannot be used since it only tells us to which category {food, drink, outside, inside} the image falls into. This leaves us behind with the caption associated with the image. Sentiment analysis(VADER Sentiment) was applied to the caption associated with the image and based on the intensity value of the sentimental analysis, image was labeled positive or negative. On applying VADER Sentiment, 31,674 images were assigned a positive label and 5,101 images were assigned a negative label. The rest 244,217 images were assigned a neutral label. In order to avoid any kind of bias in the training data, only a subset of images i.e 5,101 negative images and 5,101 positive images were used in training.

3.5.2 Training a Neural Network. Deep learning in recent times has shown that much higher accuracies can be obtained for classification tasks that were not possible earlier with traditional Computer Vision techniques. In order to train a deep learning architecture you need huge amounts of data, typically in the order of hundreds of thousands. Huge computational resources like multiple GPUs are also needed. Since our task was to implement a simple binary classifier, fewer number of images sufficed. Data augmentation techniques were applied to be robust enough to the variations in data. Different classic networks like LeNet [11], MiniVGGNet[2], ResNet [4], MiniGoogLeNet[3], DeeperGoogLeNet [2] were tried.

3.5.3 Transfer Learning. Instead of trying to learn the weights from scratch, a pre-trained network like VGG16 [5] trained on a huge dataset like ImageNet has shown to generalize well to other datasets and can be used as feature extractors. We used VGG16 as a feature extractor and trained Logistic regression, SVM classifier on top of that.

3.5.4 Generating Ground Truth by manual annotation. We took a subset of images i.e 2,000 and manually annotated them as positive and negative. Number of positive images were 1,715 and number of negative images were 285 after the annotation process.

3.5.5 Training a Neural Network on Manually Annotated Data. The results obtained by generating ground truth using VADER Sentiment were not good enough. The same networks used in section 3.5.2 were trained with the manually annotated data. In order to remove the bias in the training data, only 285 images were taken from the positive set out of 1,715 along with 285 negative images.

3.6 Semi-Supervised Image Classification

The results obtained using supervised classification techniques were not good enough, hence semi-supervised techniques were explored. In order to train a semi-supervised classifier, you still need ground truth for at least few set of images. The images annotated in the section 3.5.4 were used here.

3.6.1 Trained an SVM(Support Vector Machines). SVM is a linear classifier commonly used in Computer Vision Community for the classification task. It is shown in literature to obtain best results. We trained an SVM on the conv features corresponding to the 285 images chosen from each of the positively and negatively annotated images. The trained model was used to predict the labels for all the remaining images in the dataset(280,422). We added 50 images to the training set from each of the positively and negatively labeled images that were predicted with highest confidence. The SVM was then retrained on this new set of training data(335 positive and 335 negative). The process of adding images to the training dataset and retraining the classifier was repeated until the accuracy on the validation dataset[4] was unchanged. Once the training was done, images in the test dataset[5] were classified.

4 RESULTS AND DISCUSSIONS

Here we outline key differences in our model approaches and showcase the results.

4.1 Feature based Rating Prediction

From Table 1, we observed that on increasing the feature set size by adding more business related features such as tips, reviews etc, the models consistently performed better. Previous work in [7] showed that regression models performed better than classification models considering reviews alone. We observed the same phenomenon after considering additional business features, reviews and tips data. Linear Regression and RandomForest Regressor(RFR) have performed the best among all the models with latter producing the highest RMSE of 0.50. The best RFR model consisted of business features from section 3.2, features from section 3.3 excluding top 500 unigrams. All the models have performed better than the baseline and the best model showed an improvement of 1.488. This also outperformed the best performance RMSE of 0.84 using RandomForest

[2]Smaller version of VGGNet with a fewer number of conv layers
[3]Simplified version of Inception module has been used

[4]Validation dataset was created by taking 252 positive images and 42 negative images that were manually annotated
[5]Test dataset was created in the similar fashion as validation dataset

Table 1: Prediction Results

Model	Set I[1]	Set II[2]	Set III[3]
Linear Regression	0.99	0.92	0.59
Logistic Regression	2.67	1.54	1.02
RandomForest Regression	0.94	0.62	0.50
SVM	2.50	1.54	1.12
Neural Networks	1.98	1.66	1.27

The numbers in the table denote the RMSE value
[1] Feature set consists of only the latent variables
[2] Feature set consists of all the business attribute
[3] Feature set consists of all business attributes, features extracted from reviews, tips and checkin dataset

Regression as reported in [9], which predicts ratings using reviews only. This is likely due to the fact that we included more business related features and tip data which hasn't been exploited by other works.

4.2 Prediction for Images

From Table 2, we observed that no matter what architecture we chose the results were pretty much the same. ResNet Decay gave the highest accuracy of 61.88 % amongst all. The next highest accuracy of 61.62% was given by MiniGoogLeNet followed by Modified LeNet with data augmentation and mean subtraction.

The accuracies of the networks with manually annotated images were higher by 10%-20% when compared to ones using VADER Sentiment. Though the caption associated with the image was negative, the image itself was quite good and was labeled positive[6] during manual annotation. This phenomenon was observed with many images. This led to ambiguities in the training dataset when using VADER Sentiment as manually annotated positive images[6] were present in both positive set and negative set obtained using VADER. Thus networks had a hard time in learning this. Thus the assumption of VADER Sentiment in generating the ground truth for images did not work in our favor. To our knowledge VADER was the best open source tool available and using other paid state of art sentimental analysis techniques would have given us better results.

Once the images were annotated manually, the ambiguity in the data set was bare minimum and networks started performing better as you can see from the Table 2. Though the results were impressive, we did not proceed in that direction due to limited resources. Instead we trained an SVM classifier on top of the conv features given with the dataset corresponding to manually annotated images and obtained an accuracy of 75.17%.

4.3 Features and Image based Rating Prediction

We trained the two best models from section 4.1 by combining features from section 4.2. The interpretation of the feature weights and RMSE corresponding to the models is as follows:

[6] A positive image is one which has enough information describing it and is visually appealing

Table 2: Supervised Image Classification Results using Vader Sentiment and Manually Annotated Data

Model	Acc. Vader	Acc. Manual Annotation
LeNet[1]	61.00	80.42
LeNet[2]	60.14	-
Modified LeNet[3]	60.38	78.32
Modified LeNet[4]	61.31	-
MiniVGGNet[5]	60.88	-
MiniVGGNet[1]	60.46	-
MiniVGGNet[2]	60.88	-
Transfer Learning	59.00	73.00
MiniGoogLeNet[2]	61.62	79.72
DeeperGoogLeNet[1]	60.84	76.92
ResNet[1]	60.27	78.94
ResNet Decay[1]	61.88	-

[1]With Data Augmentation
[2]With Data Augmentation and Mean Subtraction
[3]Additional conv layers with more number of filters in each of them with Data Augmentation
[4]Additional conv layers with more number of filters in each of them with Data Augmentation and Mean Subtraction
[5]Without Data Augmentation

Linear Regression:

The RMSE for the model improved to 0.983 when the features from section 4.2 were included in training.

Intercept: The value of θ_0 (Equation 3) was about 3.2, which seems to be a good approximation of the rating when no features are available and is also close to average rating of the data which is 3.49.

The top 5 features that contributed the most for predicting the rating was calculated by sorting the coefficients obtained from the trained model. The features in order of the contribution are: PriceRange, IsOpen, Location, Reviews, BusinessAcceptsCreditCards.

RandomForest Regression:

The RMSE for the model improved to 0.497 when the features from section 4.2 were included in training. The top 5 features that contributed the most for predicting the rating was calculated by sorting the feature_importance values obtained for each feature from from the trained model. The features in order of the contribution are: Open24Hours, BusinessAcceptsCreditCards, Location, Parking, Ambience.

We performed correlation analysis on ratings and sentiment score obtained from images. It was observed that the correlation value is 0.14. From this, it can be concluded that since the correlation value is small, the improvement in RMSE using image features is also minute.

5 FUTURE WORK

We plan to improve our basic Neural Network and also explore Deep Neural Networks to find the attributes affecting the star prediction.

We also plan on identifying the strength of opinions and incorporate POS tagging for reviews. We also plan to use other state of art sentiment analysis techniques for reviews and images. Due to a limitation in computational resources and man power to label the data, only a subset of images were used to train deep learning models and training these models on bigger datasets would be one of our future works. Newer architectures like DenseNet, SqueezeNet, ResNeXT needs to be explored.

ACKNOWLEDGMENTS

We would like to thank Prof. Anoop Kumar and Prof. Rafael Ferreira da Silva for giving us an insight on Data Mining and providing their valuable feedback throughout this process.

REFERENCES

[1] 2018. *The Yelp Dataset Challenge.* https://www.yelp.com/dataset/challenge.
[2] Yangqing Jia Pierre Sermanet Scott Reed Dragomir Anguelov Dumitru Erhan Vincent Vanhoucke Andrew Rabinovich Christian Szegedy, Wei Liu. 2014. *Going Deeper with Convolutions.* https://arxiv.org/abs/1409.4842.
[3] Tejeswini Sundaram Jeyavaishnavi Muralikumar, Nivetha Thiruverahan. [n. d.]. *Business Star Prediction on Yelp Dataset.* https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a035.pdf.
[4] Shaoqing Ren Jian Sun Kaiming He, Xiangyu Zhang. 2015. *Deep Residual Learning for Image Recognition.* https://arxiv.org/abs/1512.03385.
[5] Andrew Zisserman Karen Simonyan. 2014. *Very Deep Convolutional Networks for Large-Scale Image Recognition.* https://arxiv.org/abs/1409.1556.
[6] Alex M. [n. d.]. *Finding Beautiful Yelp Photos Using Deep Learning.*
[7] Maryam Khademi Mingming Fan. 2014. *Predicting a Business Star in Yelp from Its Reviews Text Alone.* https://arxiv.org/abs/1401.0864.
[8] Bengt O. Muthen. [n. d.]. *Beyond SEM, General Latent Variable Modeling.*
[9] Vraj Shah Ojas Gupta, Sriram Ravindran. [n. d.]. *Text Based Rating Prediction on Yelp Dataset.* https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a041.pdf.
[10] Zhen Yuan Zhou Lijuan Wang Hua, Chen Nin-ning. 2011. *The Data Processing Based on Factor Analysis.* https://ieeexplore.ieee.org/document/6003113.
[11] Y. Bengio P. Haffner Y. Lecun, L. Bottou. 1998. *Gradient-based learning applied to document recognition.* https://ieeexplore.ieee.org/document/726791.
[12] Chenghu Zhang Ying Yue, Xinghua Ma. 2010. *Comprehensive Performance Evaluation of the Listed Companies in Coal Mining Industry Based on Factor Analysis and Cluster Analysis.* https://ieeexplore.ieee.org/document/5480444.