Yelp Dataset - Sentiment Analysis and Rating Prediction

Saksham Gupta
Virginia Tech.
saksham@vt.edu

Pronnoy Goswami
Virginia Tech.
pronnoygoswami@vt.edu

ABSTRACT

In today's data-driven world reviews play a significant role in impacting various business, products and services. Our preferences while buying products online or at a store are involuntarily affected by the reviews of that product. The business and revenue of all major e-commerce, travel, and restaurant business are largely impacted by the reviews their products or services receive. Yelp is also a platform that allows users to post reviews about various restaurants where they had an opportunity to visit. The review is a free-form text whereas the star rating is out of 5 stars. In this project proposal, we present a systematic methodology to analyze the sentiment of a review posted by a user about a restaurant and also predict the rating from that review. We can treat the sentiment analysis part as a binary classification problem, where the two classes are - positive and negative. On the other hand, the rating prediction task is a multi-class classification problem. In this proposal we explore the Yelp Dataset and possible machine learning methodologies which can be used to tackle the classification problem. Finally we will provide a detailed comparison of the best combination of feature extraction and prediction methodology.

KEYWORDS

Data Mining, Data pre-processing, Machine Learning, Model Evaluation

1 INTRODUCTION

Yelp [1] is an online public sourced review forum for local businesses. Founded in 2004, Yelp currently has more than 4.5 million crowd-sourced reviews. In the growing world of connectivity Yelp acts as platform to help new users by providing peer information about businesses and local markets.

Yelp annually holds a challenge where in it provides its text review and photo dataset with the aim of getting some meaningful insights and models from the research community. One of the use-case of learning is Sentiment Analysis, wherein given some review text the learned model tries to predict whether the user finds the business good or bad.

2 DATA DESCRIPTION

Yelp Dataset [1] consists of 5,996,996 text reviews about 188,593 businesses containing 280,992 pictures belonging to 10 metropolitan areas. For our project we use the reviews, business and user information to solve sentiment analysis and rating prediction problems. In reviews we kept the unique id of each review, user, business, text and stars, while for user we kept user ID, Name and review count. For Business we kept business ID, Name and address. Fig:1 shows a detailed relational view of the dataset, describing the type of attributes provided and their relations. For sentimental analysis we convert the star ratings from 1-2.5 stars as negative and 2.6-5 stars as positive reviews. For multi-class rating prediction we take round the ratings to form 5 classes ranging from 1 to 5 stars discretely.

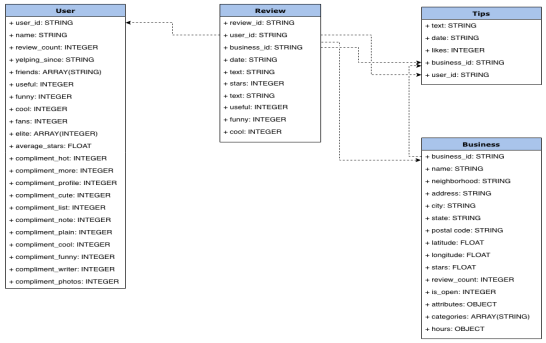

Figure 1: Yelp Dataset Schema: Depicts the attributes provided and the relationship between different tables

3 DATA PRE-PROCESSING

For our problem we focused on businesses in United States only, so we removed all the non-US businesses from the dataset. Due to this we had to remove reviews belonging to removed businesses from the review dataset. We also cleaned out those reviews which did not relate to any business in the dataset. We had some concerns regarding the fake reviews present in the dataset, but for now we didn't cleaned them because that would require to resorting to unofficial dataset sources leading to false results.

Tokenization: Tokenization describes the general process of breaking down a text corpus into individual elements that serve as input for various natural language processing algorithms.

Stop Words: Stop words are words that are particularly common in a text corpus and thus considered as rather un-informative (e.g., words such as so, and, or, the). One approach to stop word removal is to search against a language-specific stop word dictionary. An alternative approach is to create a stop list by sorting all words in the entire text corpus by frequency. The stop list after conversion into a set of non-redundant words is then used to remove all those words from the input documents that are ranked among the top n words in this stop list.

Stemming and Lemmatization Stemming[2] describes the process of transforming a word into its root form. The original stemming algorithm was developed my Martin F. Porter in 1979 and is hence known as Porter stemmer.

Sometimes stemming can create non-real words. In contrast to stemming, lemmatization aims to obtain the canonical (grammatically correct) forms of the words, the so-called lemmas. Lemmatization is computationally more difficult and expensive than stemming.

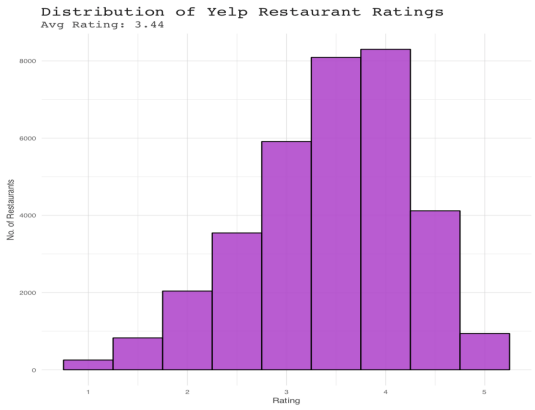

Figure 2: Overall Distribution of Yelp Restaurant Ratings

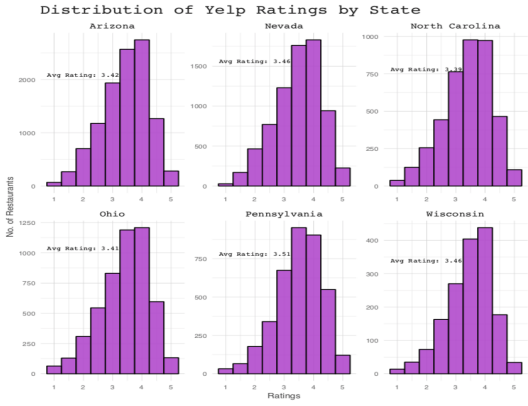

Figure 3: State-wise Distribution of Yelp Restaurant Ratings

Figure 4: Distribution of Yelp Ratings for highest Variance restaurant categories

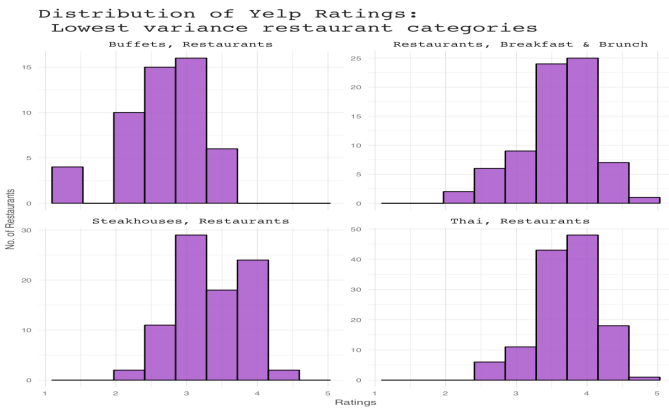

Figure 5: Distribution of Yelp Ratings for lowest Variance restaurant categories

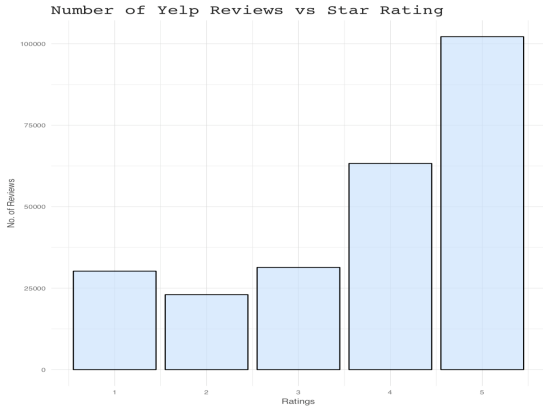

Figure 6: Number of Yelp Reviews vs Star Ratings

4 DATA EXPLORATION

Fig.3 depicts the ratings aggregated on different US States. From this we can observe that the distribution of ratings is similar to the overall distribution of ratings in US depicted by Fig.2.

We also observed that the highest variance restaurants categories(Fig.4) are the fast-food, pizza, sandwiches, burgers, and hot dogs restaurants. On the other hand, the lowest variance restaurant categories(Fig.5) are buffet, breakfast and brunch, steakhouses and Thai restaurants.

We also analyzed the types of ratings that people leave for restaurants at Yelp and found that the number of good reviews were greater than the number of bad reviews. It implies that people are more likely to leave good reviews on Yelp than bad reviews. By this analysis, we can have an initial estimate of the distribution of classes (in this case ratings) of our dataset.

5 MODEL BUILDING

5.1 Feature Selection

When dealing with high volume text data we cannot directly apply machine learning algorithms as it will complicate the model, add redundancy and might lead to learning of low variance and dependent attributes. Thus feature selection and reduction is one of the most important step in any data analytics algorithm.

5.1.1 ***Unigram, Bigram and Trigram.*** In unigram model we consider each word appearing in the review corpus as a feature just like a bag of words model. [3] One of the disadvantage of only using unigram as features in natural language processing is the presence of word relations like "not good" in the reviews. To cover such relations we extend the unigram approach by taking into consideration bigram and trigram features as well. Finally we multiply these features with their TF-IDF Term Frequency and Inverse Document Frequency to give more weight-age to word that appear less frequently in the reviews and act as a good distinguisher in comparison to words that occur very frequently and does not help in identifying the class.

5.1.2 Latent Semantic Indexing. One of the major problem in Natural Language processing is the presence of large number of features extracted from text which leads to complex models and high computation requirements, ultimately leading to low accuracy and poor results. LSI match topics instead of exact words i.e. words occurring in similar context. We utilize the matrix generated by unigram[4] and apply Singular Value Decomposition(SVD) to yield three matrices USV out of which V forms our final feature matrix.

5.2 Learning

5.2.1 Decision Trees - C4.5.
The C4.5 decision tree algorithm [5] uses normalized information gain as the splitting criterion and splits on the attributes with the highest value to produce child nodes that have homogeneous distribution of classes.

5.2.2 Naïve Bayes Classification.
The Naïve Bayes Classifier makes the assumption that given a class the conditional probability between any two features is independent of each other. [6] We calculate the posterior probability given a feature using this assumption and build the model. Then, when a new feature is encountered we compute all the joint probability values of the class for that feature and the highest probability is the output as the final class label for this new feature.

5.2.3 Support Vector Machine(SVM). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. [7] New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. For our approcach we will be using linear SVMs.

5.2.4 Long-Short Term Memory Networks(RNN + LSTM). Since reviews are sequential data and order of words encode lot of useful information, we can use Recurrent Neural Network(RNN's) with Long-Short Term Memory Units(LSTM's) to learn these sequential patterns in the data. [8] We can design a network that can learn some low level embeddings which can be passed through LSTM units to add recurrent information about the sequence of words in the reviews. For better understanding a simplified depiction is given in Fig.7.

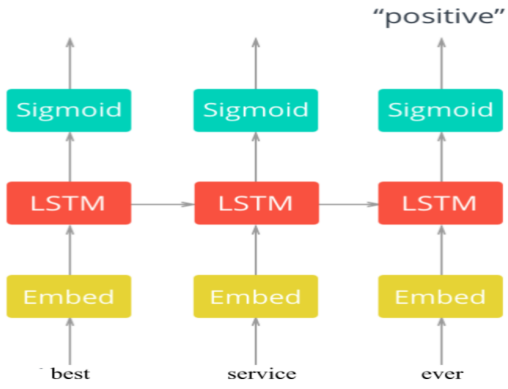

Figure 7: Recurrent neural networks with LSTM units

6 MODEL EVALUATION

We are planning to use 70% of the dataset for training, 10% for creating a validation set, and 20% for testing. We will be using Receiver Operating Characteristic (ROC) Curve as the primary metric to evaluate our model.

6.0.1 ROC Curve.
The ROC Curve[9] is the plot of the *True-Positive Rate (TPR)* versus the *False-Positive Rate(FPR)* for various threshold settings. The *TPR* is also called the *sensitivity*, *recall* or *probability* of detection whereas the *FPR* is also called the *fall-out* or *probability of false alarm*. The area under the ROC curve is the measure of the correctness of classification of the model given an unseen sample point. An ideal model should have $AuC = 1$.

REFERENCES

[1] Yelp. Yelp dataset challenge'2018.
[2] B. P. Pande, Pawan Tamta, and H. S. Dhami. Generation, implementation and appraisal of a language independent stemming algorithm. *CoRR*, abs/1312.4824, 2013.
[3] Riadh Bouslimi, Abir Messaoudi, and Jalel Akaichi. Using a bag of words for automatic medical image annotation with a latent semantic. *CoRR*, abs/1306.0178, 2013.
[4] Nabiha Asghar. Yelp dataset challenge: Review rating prediction. *CoRR*, abs/1605.05362, 2016.
[5] Kemal Polat and Salih GÃijneÅ§. A novel hybrid intelligent method based on c4.5 decision tree classifier and one-against-all approach for multi-class classification problems. *Expert Systems with Applications*, 36(2, Part 1):1587 – 1592, 2009.
[6] I. Rish. An empirical study of the naive bayes classifier. Technical report, 2001.
[7] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML'98, pages 137–142, Berlin, Heidelberg, 1998. Springer-Verlag.
[8] Zachary Chase Lipton. A critical review of recurrent neural networks for sequence learning. *CoRR*, abs/1506.00019, 2015.
[9] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861 – 874, 2006. ROC Analysis in Pattern Recognition.