Machine Learning (COMP30027)
Report

1. Introduction

In today's time, there a large amount of restaurant reviews that can be found on the web, which can prove to be very beneficial in forming a detailed and descriptive sentiment analysis. This can be done by analysing the opinions and sentiments which are presented in these reviews. A restaurant gets its rating through its **customers' reviews** and hence it is able to develop new and better business strategies through these **sentiment analyses** (Doan & kalita, 2016). The useful keywords can be further extracted from these sentiments and thus after being analysed they can be put in respective category depending on the whole customer review. **Yelp** is the web service that provides all the relevant data to conduct the sentiment analysis.

In this project, the focus is on the task of **sentiment categorization** to rate the numbers of stars given to a restaurant. This can be useful for the customers selecting a restaurant based on the reviews and star ratings given by other customers.

Sentiment analysis categorizes sentences to either **negative** or **positive** polarity and **factual** or **opinionated** subjectivity.

In the given Yelp Dataset, there are only **five** relevant features that are essential in the evaluation of the sentiment analysis. The **reviews** are needed to calculate the polarity and the subjectivity of the sentiment analysis. Apart from the reviews, other four features are needed to analyse and accurately predict the rating column using predictive models such as the logistic regression model, KNN classifier, Naïve Bayes and so on, which is then added to the test data.

2. The Dataset

The datasets provided from Yelp consist of **4 csv files** for the purpose of this project. These are then divided into meta and text. The meta dataset consists of 8 columns: **Date, review ID, reviewer ID, business ID, vote_funny, vote_cool and vote_useful** and the test dataset contains all the **text reviews** of the people in text form. Furthermore, the datasets are also divided for training and testing purposes. The text and meta train have **28068 rows** implying they have 28068 reviews and the text and meta test have **7018 rows** meaning they have 7018 reviews. The meta train dataset also contains an extra column ***'rating'*** which contains all the rating given by the reviewers. This assists in calculating the accuracy of the data.

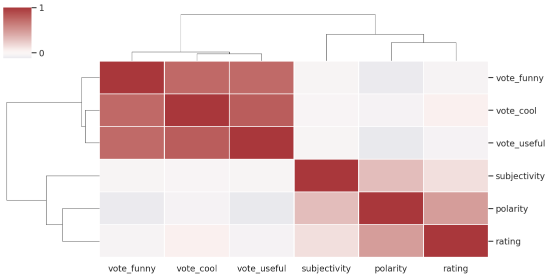

Figure 1 – Correlation of features

2.1 Cleaning the data

There are **two main** steps involved in cleaning of the data:

The first step was to **merge** both meta and text datasets to one merged train dataset along with one merged test dataset.

The second step was the **preprocessing** of the data where I **regex** was used to clean the data that involves the process of removing all the symbols, numbers and '\n' from the review column.

2.2 Frequency of the data:

To find the number of words used in a sentence the **CountVectorizer** function is used (learn, 2007-2019). It creates a matrix of text and frequency of them. Through that matrix created we then do an **exploratory data analysis (EDA)** in which we find out the **most common** words of the review data and plot as a bar graph and generate it as a word cloud (Jain , 2018) .

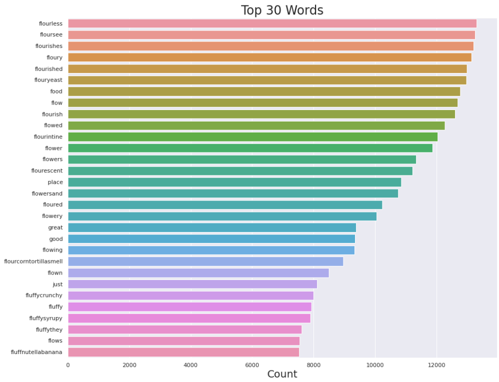

Figure 2- Top 30 words

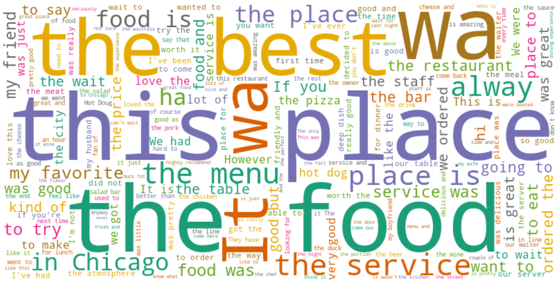

Figure 3- Word Cloud for the most common words

3. Sentiment Analysis

Sentiment analysis is the automated process of understanding the **sentiment or opinion** of a given text. It gives an insight into the given text, determining if it's **positive, neutral, or negative** in nature (Ghazvinian, 2010). About 80% of the world's data is unorganized which makes it hard to understand and analyse it. It would be very costly to arrange and organize all this data, not to disregard the amount of time it will take to do it. However, sentiment analysis labels this **unstructured data** and makes it easier to interpret.

It is divided in two classifiers: **Polarity and Subjectivity**. Since polarity refers to sentiment orientation, it is the **major metric** of measurement.

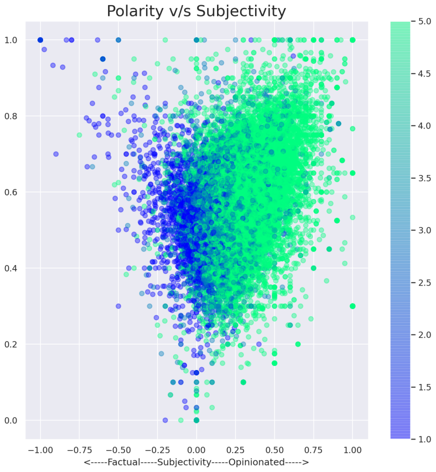

Figure 4- Polarity V/S Subjectivity

3.1 Polarity

This classifier calculates float value ranging from **-1 to 1**. A **positive** statement will be a positive number (above 0) and a **negative** statement will be a negative number (under 0) (Thakkar, 2017).

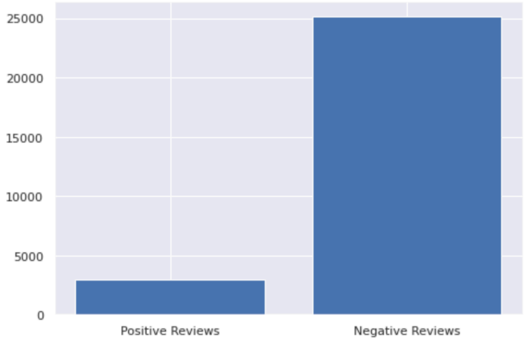

Figure 5- Positive V/S Negative Reviews

3.2 Subjectivity

This classifier calculates float value ranging from **0 to 1**. A **factual** expression will be under 0.5 and an **opinionated** expression will be a negative number (under 0) (Thakkar, 2017).

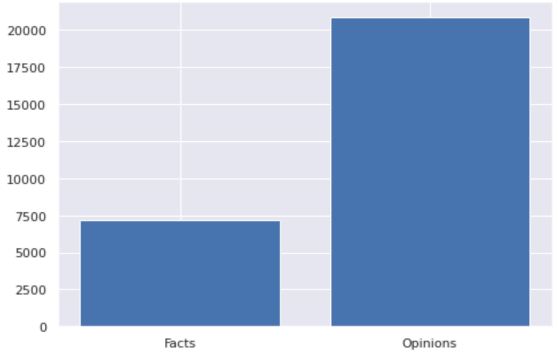

Figure 6- Factual V/S Opinionated reviews

4. Predictive Model

In this project, 4 models were implemented in order to find the accuracies from the training and data. This was done to see if those **predictions** could be **implemented** to the test data with **reasonable accuracies** from the models (Asghar, 2016). The models chosen for this classification are logistic regression model, KNN classifier and Naïve Bayes.

4.1 Logistic Regression Model

Logistic regression is a predictive analysis and is used to interpret data. It explains the relationship between a **binary variable and one or more ordinal, nominal, interval or ratio-level independent variable** (Li Chen & Zhang, 2014). In this case, there was the task of finding the dependencies of features like vote_funny, vote_cool, vote_useful and the polarity and subjectivity that were calculated to the binary variable rating (1,3 and 5) of the train data. The former gives an accuracy of **72.355%** which is reasonable. Thus, this model is selected to make a prediction on the test data.

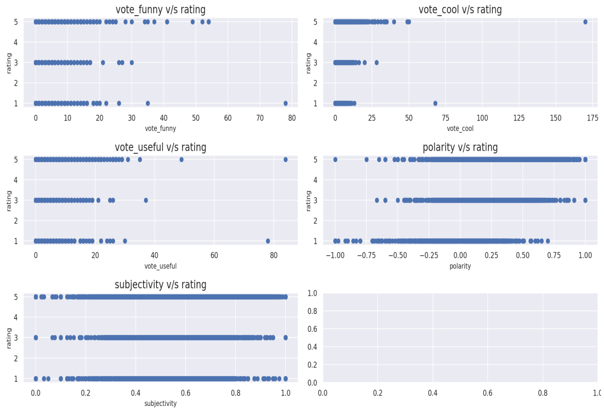

Figure 7 – Data dependencies distribution

4.2 KNN Classifier

KNN Classifier is based on feature similarity. It checks how much does the features resemble to give a final data (Bronshtein, 2017). In this case the **resemblance and similarities** of the features like vote_funny, vote_cool, vote_useful, the polarity and subjectivity were observed, and a rating is predicted which is then used to check the accuracy of the train dataset's rating. This gives an accuracy of **72.818%** which is again reasonable and hence this model is used to make a prediction on the test data. The KNN classifier is also dependent on number of neighbours it can check as that gives an accurate prediction groups of nodes created. It is good to have an odd number of neighbours as that avoids a conflict of nodes thereby leaving the prediction unaffected by any conflict.

4.3 Naïve Bayes Model

Naïve Bayes works on the concept of Bayes Theorem of **probability to predict the class of unknown datasets**. Naïve Bayes is dependent on the number of features it uses for predictions (Ray, 2017).

4.3.1. Gaussian naïve Bayes

Gaussian Naïve Bayes is used to classify dataset which are **continuous** and not discrete hence it was not chosen to be the classifier for naïve Bayes and had an accuracy score of 62.487% (Ray, 2017).

4.3.2 Multinomial Naïve Bayes

Multinomial Naïve Bayes works well with dataset that are **discrete** and have more than 2 number to predict. For the dataset Multinomial should perform the best as it has to predict 1,3 and 5 stars of the rating, but in this case, Bernoulli **outperforms** Multinomial Naïve Bayes. The accuracy score of Multinomial is 70.004% (Singh, 2018).

4.3.3 Bernoulli Naïve Bayes

Bernoulli Naïve Bayes works the best when the feature vectors are **binary** i.e. when it is zeros and ones (Singh, 2018).

It works well in this case as the Bernoulli Naïve Bayes algorithm is relevant and can more efficiently predict the data as compared to the Multinomial Naïve Bayes Algorithm as the Multinomial Naïve Bayes algorithm can predict the number of times an element **'x' occurs in 'n' trials.** Hence, Bernoulli Naive Bayes will outperform Multinomial Naive Bayes as it is used **for predicting values between 0 and 1**. The accuracy score of Bernoulli Naïve Bayes is **71.286%** and is hence chosen as the accuracy of Naïve Bayes.

5. Conclusion

Model	**Accuracy**
Logistic regression	72.355%
KNN classifier	72.818%
Naïve Bayes	71.286%

Table 1- Accuracies of the models

From the above table (Table 1), it is justified that **KNN classifier** gives the best accuracy score and is hence the ideal model to be used to predict the ratings of the test dataset.

Although, by taking into account a larger dataset with a **greater number of features/ information** about the reviews the predictive power of the model may improve marginally.

6. References

Asghar, N., 2016. Yelp Dataset Challenge: Review Rating Prediction, Waterloo: University of Waterloo.

Bronshtein, A., 2017. A Quick Introduction to K-Nearest Neighbors Algorithm. use journal.

Doan, T. & kalita, J., 2016. Sentiment Analysis of Restaurant Reviews on Yelp with Incremental Learning, Colorado Springs: University of Colorado.

Ghazvinian, A., 2010. Star Quality: Sentiment Categorization of Restaurant Reviews, s.l.: Stanford University.

Jain , S., 2018. Natural Language Processing for Beginners: Using TextBlob. Analytics Vidhya.

learn, s., 2007-2019. [Online]Available at:https://scikitlearn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

Li Chen & Zhang, J., 2014. Prediction of Yelp Review Star Rating using Sentiment Analysis, s.l.: Stanford University.

Mukherjee, A., Venkataraman, V., Liu, B. & Glance, N. What Yelp fake review filter might be doing? 7th International AAAI Conference on Weblogs and Social Media, 2013.

Ray, S., 2017. 6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R. Analytics Vidhya.

Rayana, S. & Akoglu, L. Collective opinion spam detection: Bridging review networks and metadata. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015. 985-994.

Singh, A., 2018. What is the difference between the the Gaussian, Bernoulli, Multinomial and the regular Naive Bayes algorithms?. *Quora.*

Thakkar, D., 2017. *quora.* [Online]
Available at: https://www.quora.com/What-is-polarity-and-subjectivity-in-sentiment-analysis