CSE 842 Project: Hotel Reviews Classification

Saloni Pandya
Department of Electrical and Computer Engineering
Michigan State University
East Lansing, Michigan, USA
pandyas6@msu.edu

***Abstract*—In this age when people are travelling around the globe, the world has become a global village. Affordable and comfortable hotels have become a go-to for explorers. The concepts of online rating and giving one's opinions online have become a major trend in the hotel industry. Many investors rely upon platforms that allow customers to give their reviews. Such review data can have a large impact on impression on potential users and profits of a hotel. Natural Language Processing can be a great tool to process such data in order to classify the reviewed product into various categories such as sentiment classification, user nationality and average rating. The use of semantic analysis and the concept of word bag model are good means of creating such classification.**

***Index Terms*—CSE 842 Final Project; sentiment analysis; natural language processing; LogisticRegression.**

I. Introduction

There are two main categories in textual information, they are ***facts*** and ***opinions***. Facts are objective statements while opinions are subjective ones. A lot of research is being conducted to retrieve information from different sources to throw light on these two aspects of a statement. Opinions are the subjective statements and still rare in existing researches. Opinions reflect the people's sentiments or feelings about the product and events. Many of the existing research are based on mining and retrieval of factual information and not on opinions. Opinions are also important when someone wants to hear the other's viewpoint before they make a decision [1].

The dataset used in this project consists of 515,000 user reviews, extracted from the source www.booking.com, of 1493 lavish hotels located across Europe. The final classification of hotels based on their reviews provided by thousands of customers will provide a solid base for investors to judge the fame and customer satisfaction of the hotels.

II. Framework

A. The steps involved

The classification of data of online reviews of hotels mainly contains some steps: capture data, clean data, extract feature, sentiment analysis of reviews, splitting data into sets, training models with our training data and so on. The Figure 1 shows the process of classification of data and its modelling in steps.

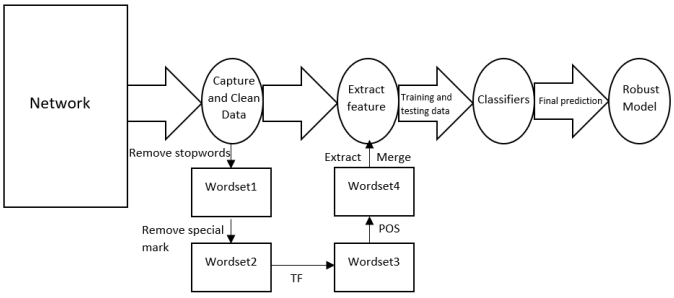

Fig. 1. The process of hotel and review classification

B. TF-IDF

1) Term Frequency (TF): the word W is the number of times that word is in the review.
The higher the frequency (TF), the higher the review, and the more attention it gets. The word frequency NW of a word W (i.e., the number of times that word W appears) is:

$$N_W = \sum_{i=0}^{N} W_i : W_i \in W \tag{1}$$

where $\boldsymbol{W_i}$ is the $\boldsymbol{i}^{th}$ time the word W comes up [2].

2) Comment Frequency (DF): It is called document frequency, refers to the proportion of a comment in the total comment. DF= Comments containing words W/the total number of comments, N is the total number of comments, and the comment frequency is calculated as shown,

$$DF = \frac{N}{W} = \frac{\sum_{i=0}^{N} W_i : W_i \in W}{N} \tag{2}$$

3) Inverse document frequency (IDF): Measure the review frequency of a word or phrase in all. The higher the frequency of the inverse document, the more important of the word. It is a measure of the universal importance of the word. What IDF think is: if the less of the comments with word W, the smaller is N_W, more significant is IDF, then the word W has a good classification ability. The IDF of a specific word W can divide by the total comment number by the comment containing the word,

and then the logarithm is obtained [2] The calculation formula is as shown,

$$IDF = LOG\frac{N}{N_W + 1} \tag{3}$$

$$= LOG\frac{|N|}{\left|\{\sum_{i=0}^{N} W_i : W_i \in W\} + 1\right|} \tag{4}$$

4) The Frequency-Inverse document frequency (TF-IDF): The frequency - inverse document frequency (TF-IDF) is an essential measure of a candidate keyword in combination with word frequency and inverse document frequency. The frequency - inverse document frequency (TF-IDF) is considered to be one of the most effective and commonly used features of all features. If a word or phrase in an article in the high frequency of TF1, and rarely appears in other materials, say the word or phrase to distinguishability, has the outstanding suitable for classification [2]. The calculation of TF-IDF is as shown:

$$TF_{IDF} = \frac{TF_W}{V_N(length)} \times LOG\frac{N}{N_W} \tag{5}$$

$$= \frac{\{\sum_{i=0}^{N}\sum_{j=0}^{M} W_{ij} : W_{ij} \in W\}}{\sum_{i=0}^{N}\sum_{j=0}^{M} W_{ij}} \times \tag{6}$$

$$LOG\frac{|N|}{\left|\{\sum_{i=0}^{N} W_i : W_i \in W\} + 1\right|} \tag{7}$$

where $\boldsymbol{V_N(length)}$ is the length of N comments. The TF-IDF value is proportional to the frequency of the word and inversely proportional to the number of times it appears in the entire review [2].

III. Methodology and Testing

A. Evaluation method to evaluate our performance

There are three indexes generally used in text categorization: Recall, Precision, and Accuracy. So we adopted these indexes to evaluate the performance of sentiment classification in our study [3]. The indexes such as accuracy, precision and recall are found in the Figure 2. The three indexes mentioned above are calculated as follows.

$$Accuracy(A) = \frac{a+d}{a+b+c+d} \tag{8}$$

$$Precision(p) = \frac{a}{a+b} \tag{9}$$

$$Recall(p) = \frac{a}{a+c} \tag{10}$$

$$Precision(n) = \frac{d}{c+d} \tag{11}$$

$$Recall(n) = \frac{b}{b+d} \tag{12}$$

Sentiments of the reviews are analyzed using SentimentIntensityAnalyzer and the good and bad reviews are segregated

	Actual positive	Actual negative
Predict positive	**a (Tp)**	**b (Fp)**
Predict negative	**c (Fn)**	**d (Tn)**

Fig. 2. Contingency Table for Performance Evaluation [3].

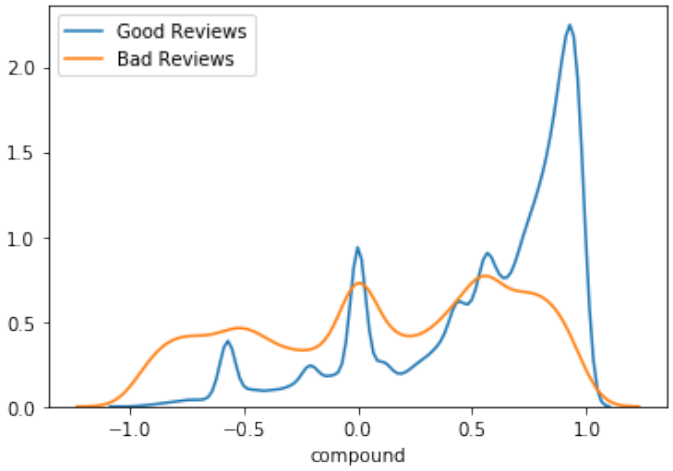

Fig. 3. The density plot of sentiment distribution for the positive and negative reviews.

with labels. The density of the sentiment distribution has also been plotted to get an idea of the distribution of good and bad reviews. It can be observed from Figure 3 that the good reviews have a largely positive *compound score* while, the negative reviews have a low value of the *compound score*. Term frequency-Inverse document frequency are also computed which are then fed to a variety of classifiers like the LogisticRegression and the RandomForrestClassifier to train and test performance of the models. The AUCROC (Area Under Curve - Receiver Operating Characteristics) curve shown in Figure 4, Precision, recall and accuracy computed in the notebook show that the LogisticRegression classifier with penalty *l1*.

IV. Conclusion and Future

The paper starting with online review data realizes the process of hotel classification by feature extraction of online review data, feature importance, and review segregation. The article focuses on the features of the online reviews which include words like “dirty”, ”room”, “bad”, “staff”, and ”rude”. Then, using classifiers and computinf their accuracy, this paper realizes the LogisticRegression classifier with penalty *l1* gives the best accuracy. Thus, the reviews can be predicted quite accurately with this model. Finally, using the result of sentiment analysis to classify hotels. This paper will lay a foundation for intelligent recommendation hotels. Future developments can include classifying the hotels based on their reviews taking into consideration details such as their location

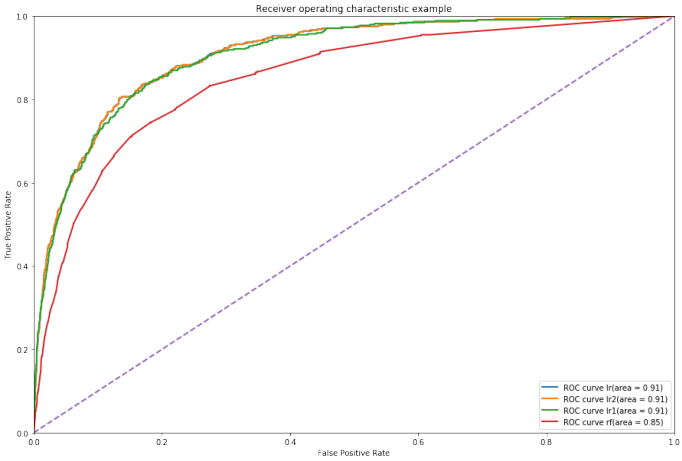

Fig. 4. The AUC-ROC curve plotted for all the classifiers used on the data after sentiment analysis.

on the map. One can also calculate the Norm weight to compute the importance of the words involved in the reviews.

REFERENCES

[1] K. Shein, "Ontology based combined approach for sentiment classification," pp. 112–115, 01 2009.

[2] H. Qin, X. Ye, Y. Zhao, and X. Cai, "Hotel classification based on online review data," in *2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 2018, pp. 264–270.

[3] T. Ghorpade and L. Ragha, "Featured based sentiment classification for hotel reviews using nlp and bayesian classification," in *2012 International Conference on Communication, Information Computing Technology (ICCICT)*, 2012, pp. 1–5.