Project Report

Title: Finding the Best Dish of a Restaurant using NLP Techniques

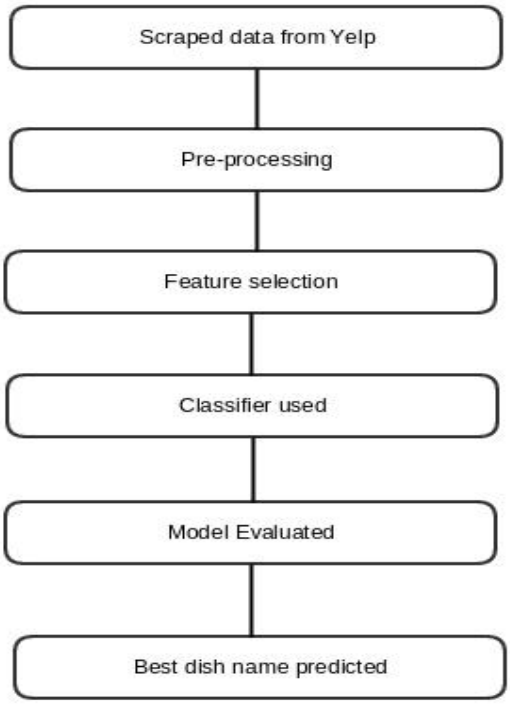
Author: Deepak Munjal (B00748375)

Course Name: CSCI 6509
Advanced Topics in Natural Language Processing

Abstract

Our goal for this project is to find the best dish of a restaurant using the various techniques of Natural Language Processing. As a subtask of this, we have to scrape data from online resource and further do sentiment analysis of it. User Review Data is scraped through online resources. For this task, we are considering the user reviews scraped for various restaurants of Halifax. This data, i.e. Restaurant's Yelp URL, name and reviews, is saved in a file. Manual labeling of each review is done. And the names of the dishes mentioned in each review are also written alongside. Then various experiments are done on this data using various classifiers and prediction of best dish of each restaurant is done. In the initial model designing, we considered a list of positive[2] and negative words[1], taken from online resources, to do sentiment classification and later we used character trigrams, word trigrams with random forest classifier and naive bayes classifier. At the end, we evaluated each model and predicted the best dish of each restaurant.

Introduction



Flowchart of the problem statement

Sentiment analysis[5] is the process of finding the emotional sentiment or opinion of the user. This process is also called opinion mining. In our case, the problem statement is to find the best dish of a restaurant by doing sentiment analysis of the user reviews. The motivation behind this is the importance of the user reviews for a restaurant on a business level. The solution to this problem statement can provide the restaurants with crucial information from the user point of view. The objective of this project is to do experimentation with various Natural Language Processing techniques that can be used for sentiment analysis, evaluating the various models and finally predicting the best dish for each restaurant.

First trial is done through matching the words in a review with the already collected list of positive[2] and negative words[1] and the sentiment of the review is predicted based on this. This is a pretty simple approach and the model designed is very basic in this case. So we proceeded further, and split each review into character trigrams. These trigrams are considered as features and term-document matrix is created. Data is split into training and test data with 70-30 percentages. Random Forest algorithm is applied on the training data. This model is evaluated with various parameters such as accuracy, confusion matrix, f score, precision, recall. Then further experimentation is done. Reviews are split into word trigrams and Random forest algorithm is applied. Similarly, Naive Bayes algorithm is applied on both character trigrams and word trigrams and evalaued all these models based on various factors. In each of the model, we finally predicted the best dish of restaurant.

Related Work

The first research work is from paper[3]- “Thumbs up? Sentiment Classification using Machine Learning Techniques” by Bo Pang and Lillian Lee, Department of Computer Science, Cornell University & Shivakumar Vaithyanathan, IBM Almaden Research Center. This paper discusses the problem of classification based on the overall sentiment of various documents. The classes being considered is positive and negative. The data on which classification is done is the movie review data. During experimentation, it is also found that the classifiers Naive Bayes, maximum entropy classification and support vector machines don’t perform well on the traditional topic based classification. This paper also discusses about the strength of sentiment of different words and does experimentation with the unigrams, bigrams and parts of speech as feature set. It considers these three and various combinations of these as feature set.

Other research work is from paper[4]- “Picking Out Good Dishes from Yelp” by Angela Gong and Jennifer Lu, Stanford University. This problem statement for this paper is to find the good dishes from the website yelp using sentiment analysis i.e. match up the menu items with the dish names mentioned in the user reviews. This research is of much importance because generally a user doesn’t write the exact dish name in the review while mentioning about any dish for a particular restaurant. So there is a need to do matching between the dish name mentioned by user in the reviews and the actual name included in the restaurant’s menu. This paper discusses about collecting the dish name list and various approaches of matching the dish names using partial match, SVM match, substring match, fuzzy match and exact match. Different models are compared using the F-score. This paper also tries to find some relation between Number of reviews and restaurant rating. Further, it mentions about rating of restaurants and the city. So it does some research about the reviews based on various locations. It considers 3 classes for sentiment analysis i.e. positive, negative and neutral. So overall both the papers discuss about various experimentation related to sentiment analysis, which was a good read and was useful in beginning sentiment analysis in our project task.

Problem Definition and Methodology

Our research problem here is to do sentiment analysis of scraped user reviews and find the best dish of each restaurant. For beginning of our work, we need to gather relevant data. This data is scraped through the famous website www.yelp.ca, which displays the user reviews for many restaurants. In our case, we scraped user reviews of 20 restaurants and 330 user reviews from Yelp. The class classification of labeled data is 276 positive and 54 negative. All these 20 restaurants belong to the Halifax in Nova Scotia, Canada. The Yelp URL of the restaurant, it’s name and the user reviews are scraped. This scraped data is saved in a CSV with these three columns. This task of scraping is done using Beautifulsoup, a Python library for scraping HTML web pages. This data is manually labeled for each review as positive or negative. Only these two classes i.e. positive and negative are considered for this project work. While labeling these reviews, the dish names mentioned in the reviews are also written alongside in a new column. So overall, we have five columns, the first three contain the Yelp URL of the restaurant, restaurant name, its user review respectively. The fourth column contains the labeled sentiment i.e positive or negative. The fifth column contains the dish names of the restaurants.

There are n number of natural language processing techniques possible for creating a machine learning model and providing a solution to the problem statement. In the initial step of this, we have to preprocess the data.

Preprocessing includes -

1. Special Symbols removal:
All special symbols and punctuation marks are removed in this step. Namely, the symbols: ?|$|.|!|,|\'|"|+|-|_|\(|\)|*|\^|#|@|~|`|/|;|:|<|>|- are removed from the reviews.

2. Conversion to lowercase:
All data in the user reviews is converted into lowercase in this step of preprocessing.

3. Removal of stopwords:
Stopwords are the words which don't have much context or importance in terms of sentiment analysis. So, all stopwords are removed from the reviews.

4. Stemming:
Stemming[6] is the process of reducing inflected or derived words to their word root form. In order to do sentiment analysis, stemming is an important step.

Preprocessing gives quite clean data to further work on. Further many Natural Language Processing techniques and classifiers are applied on this data -

Bag of words with positive and negative list of words:

In this approach, a list of positive and negative words are taken from online resources[1][2]. A list of all unique dish names is generated, which are taken from the labeled data of dish names. For each review, the review itself is broken down into bag of words. Each word is compared with the list of positive and negative words. In the similar way, count of positive words and negative words in each user review is counted. Negative words are given a weightage of negative one and positive words are given as weightage of positive one. For each review these two scores are added and a final score is calculated. This is then divided by the total number of words in the review to create a normalized score for each review. If this normalized score is zero or more, then it is considered as positive, otherwise it is considered as negative. For all reviews which have positive predicted class, based on the normalized score, count of all dish names present in the reviews is calculated. Based on this data, the dish name with the

maximum count is considered as the best dish of that particular restaurant. This model has good accuracy but is very basic and simple.

Random Forest Classifier[7]: This is an ensemble learning method for classification. It works by creating a collection of decision trees during training. The output class for this is the mode of the classes for classification. A k-fold cross validation[8] is used here. k-fold cross validation is where the complete data is divided randomly into k parts. One part is used for cross validation and rest k-1 parts are used for training the model.This process occurs k times, with a new part taken for cross validation every time and the rest k-1 parts for training the data.

Naive Bayes Classifier[9]: Naive Bayes classification algorithm is based on Bayes theorem. It merges prior knowledge with the observed data and practical learning algorithms. It estimates the probabilities for the hypothesis. Probabilistic models are developed in this, which help to derive the independent assumptions. Here also we will use k-fold cross validation method.

Random Forest and NaiveBayes Classifier with Character Trigrams as features:

In this approach, we split each review into character trigrams and considered the set of all unique character trigrams as the feature set. Document-term matrix is created with this feature set. This is divided into training and test data with 70-30 percentages. Then Random Forest algorithm is applied on the training data to train the model, where we assumed positive class to be 1 and negative class to be 2. We used k means algorithm with k as 10 for cross validation here. Further, we predicted the classes based on the test data. And followed by this, accuracy of the model is evaluated along with precision, recall, f-score etc. The count of predicted positive classes for each restaurant is calculated and based on the values, best dish of the restaurant is predicted. Same procedure is applied with the Naive Bayes classifier as well.

Random Forest and NaiveBayes Classifier with Word Trigrams as features:

In this approach, almost the same work is done except considering the word trigrams as feature set. Each review is split into word trigrams and the set of all unique word trigrams is considered as the feature set. This is followed by generating Document-term matrix. Then the data is divided into training and test data with 70-30 percentages. Then Random Forest algorithm is applied on the training data to train the model, where we assumed positive class to be 1 and negative class to be 2. We used k means algorithm with k as 10 for cross validation here. Further, we predicted the classes based on the

test data. And followed by this, accuracy of the model is evaluated along with precision, recall, f-score etc. The count of predicted positive classes for each restaurant is calculated and based on the values, best dish of the restaurant is predicted. Same procedure is applied with the Naive Bayes classifier as well.

Experiment Design

The various models are evaluated on many factors. These factors and their values are described below-

Confusion Matrix: A confusion matrix is also known as error matrix. It contains the data in a tabular format. Generally the left side has the actual class labels and top side has the predicted class labels.

Accuracy: Accuracy is the ratio of the count of data for which class predicted is correct and the total count of the dataset considered. In a confusion matrix, it is the ratio of the counts on diagonal elements and the total count of all values in the confusion matrix. Its value ranges from 0 to 1, with 1 being the most accurate and 0 the least.

Precision: Precision[10] is the ratio of count of true positives divided by total count of true positives and false positives together.

Recall: Recall[10] is the ratio of count of true positives divided by total count of true positives and false negatives together.

F-Measure: F-Measure is the harmonic mean of precision and recall.
F-Measure = 2 * (precision * recall) / (precision + recall)

We started with the first model i.e. the bag of words with positive and negative lists. Since the model is very basic and simple, so here we only evaluated the accuracy, which came out to be 0.84.

Bag of words with positive and negative list	
Accuracy	0.84

Further we evalaued the next model i.e. Random Forest Classifier with Character Trigrams as features. Here the accuracy came out to be 0.83. Values of Standard

deviation, Precision, Recall, F Measure and Confusion Matrix for an output are shown below-

Random Forest Classifier with Character Trigrams as features	
Accuracy	0.83
standard deviation	0.03
Confusion Matrix	[[82 1] [16 0]]
Precision score	0.84
Recall Score	0.99
F Score	0.91

Then evaluation of Random Forest Classifier with Word Trigrams as features is done. Here the accuracy came out to be 0.84. Values of Standard deviation, Precision, Recall, F Measure and Confusion Matrix for an output are shown below-

Random Forest Classifier with Word Trigrams as features	
Accuracy	0.84
standard deviation	0.01
Confusion Matrix	[[86 0] [13 0]]
Precision score	0.87
Recall Score	1
F Score	0.93

Further evaluation of Naive Bayes Classifier with Character Trigrams as features is done. Here the accuracy comes out to be bit low i.e. 0.76. Values of Standard deviation, Precision, Recall, F Measure and Confusion Matrix for an output are shown below-

Naive Bayes Classifier with Character Trigrams as features	
Accuracy	0.76
standard deviation	0.06
Confusion Matrix	[[67 15] [15 2]]
Precision score	0.82
Recall Score	0.82
F Score	0.82

Finally evaluation of Naive Bayes Classifier with Word Trigrams as features is done. Here the accuracy comes out to be 0.82. Values of Standard deviation, Precision, Recall, F Measure and Confusion Matrix for an output are shown below-

Naive Bayes Classifier with Word Trigrams as features	
Accuracy	0.82
standard deviation	0.03
Confusion Matrix	[[76 5] [18 0]]
Precision score	0.81
Recall Score	0.94
F Score	0.87

Conclusion

After all the experimentation with different classifiers we found that Random Forest Classifier with Word Trigrams as features, has the best F-Measure (0.93) and Accuracy (0.84). Though accuracy of the first model i.e. the bag of words with positive and negative lists, is also 0.84. But since it’s a very basic model. Hence we consider the model, Random Forest Classifier with Word Trigrams as features, as the best for this purpose. So we can conclude that we experimented with different classifiers, evaluated the models and predicted the best dish for each restaurant using each model.

Future possibilities

This work can have many future possibilities. From using more classifiers to using different n gram approaches and so on. Some of them are mentioned below-

- Currently we have used the Random Forest classifier and Naive Bayes Classifier for our task. These are used with the character and word trigrams as feature set. But further it's possible to use few more classifiers and evaluate the different models.

- The code written is pretty complex in terms of algorithmic complexity now. But it is possible to further dive deep into the research related to algorithmic complexity and improve it.

- We considered the whole review when doing the sentiment analysis i.e. we considered one class of sentiment either positive or negative for each review. But it is possible to perform the sentiment analysis on partial sentences in reviews because a review can be partially positive and partially negative. Eg. if a user talks positively about few dishes and negatively about others.

- Other possibility is to consider unigram and bigrams as well while considering the feature set. There is a possibility of considering a combination of these as well.

References

[1]N. analysis, "Negative (–ve) words and adjectives list for sentiment analysis", Dreference.blogspot.com, 2017. [Online]. Available: http://dreference.blogspot.com/2010/05/negative-ve-words-adjectives-list-for.html
[2]"shekhargulati/sentiment-analysis-python", GitHub, 2017. [Online]. Available: https://github.com/shekhargulati/sentiment-analysis-python/blob/master/opinion-lexicon-English/positive-words.txt
[3]"Thumbs up? Sentiment classification using machine learning techniques", Cs.cornell.edu, 2017. [Online]. Available: https://www.cs.cornell.edu/home/llee/papers/sentiment.home.html
[4]"Gong, Angela, and Jennifer Lu. "Picking Out Good Dishes from Yelp."", 2017. [Online]. Available: https://nlp.stanford.edu/courses/cs224n/2015/reports/4.pdf
[5]"Sentiment Analysis | Lexalytics", 2017. [Online]. Available: http://www.lexalytics.com/technology/sentiment

[6]"Stemming", En.wikipedia.org, 2017. [Online]. Available: https://en.wikipedia.org/wiki/Stemming
[7]"Random forest", En.wikipedia.org, 2017. [Online]. Available: https://en.wikipedia.org/wiki/Random_forest
[8]"Cross-validation (statistics)", En.wikipedia.org, 2017. [Online]. Available: https://en.wikipedia.org/wiki/Cross-validation_(statistics)
[9]"Naive Bayes classifier", En.wikipedia.org, 2017. [Online]. Available: https://en.wikipedia.org/wiki/Naive_Bayes_classifier
[10]"Precision and recall", En.wikipedia.org, 2017. [Online]. Available: https://en.wikipedia.org/wiki/Precision_and_recall