Sentiment Analysis of Restaurant Reviews

Kavyashree Anantha Raman, *The University of Texas at Dallas*

***Abstract*- Sentiment analysis is a prominent branch of Natural Language Processing. It deals with the process of identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc., is positive, negative, or neutral. This project concentrates on using bag-of-words model for extracting features and use this as input for machine learning models such as SVM, Logistic Regression and Naive Bayes in order to classify restaurant reviews (from Yelp) into positive or negative classes. The accuracy obtained for the three models is then compared for different number of input features.**

I. INTRODUCTION

With the vast amount of unstructured information available online, there is so much to be gained from the development of automated systems that can effectively organize and classify this data, so that the mined data can be leveraged by human users in a meaningful way. Text sentiment analysis involves the application of computational powers in understanding sentiment (Positive or Negative) implied in the writer's text. Natural Language Processing involves the task of finding the semantic meaning of the text content (in particular, of a topic, product, business category etc.) and thereby analyzing the text information. One such business category that can be considered is that of Restaurants. Restaurants constantly get feedback from customers via online portals. Yelp is one online channel which enables people to express their reviews. This project makes an attempt to use restaurant review data from Yelp and perform sentiment analysis on the text which would invariably help businesses to work on their strategies.

Machine learning also plays an important role in classifying the positive or negative sentiment for a review text analyzing the words. Machine learning can be categorized in two types known as supervised and unsupervised machine learning techniques. Supervised learning technique uses a labeled dataset where each document of training set is labeled with appropriate sentiment (class label). Whereas, unsupervised learning includes unlabeled dataset where class variables are not present. The current project involves incorporating supervised learning technique for binary classification (since two class labels i.e., Positive and Negative are involved).

Every review is considered as a document and sentence level classification is performed to determine the polarity of an individual sentence of a document. Also, aspect level classification which identifies different aspects (words, semantics) of a corpus is used to calculate the sentiment of each review with respect to the obtained aspects.

II. RELATED WORKS

This section briefly discusses about previous works related to sentiment classification of restaurant and movie reviews. The following paper (Predicting Yelp Restaurant Ratings from Review Text: Applying NLP and Machine Learning techniques to better understand the relationship between a restaurant’s rating and its review text by Parul Singh, Ilakya Palanisamy, Mukund Chillakanti, Abhinava Singh) discusses about guessing Yelp restaurant’s rating from just the text of its reviews. Upon learning the best mapping from a word vector to a business rating, they found the most essential and informative features for this classification. Further, accuracy on cross-validated data of predicting restaurant rating from review text was calculated and consequently, they concluded that Logistic regression consistently achieves better prediction accuracy than SVM.

For binary sentiment classification, Turney proposed counting the positive and negative terms and expressions in a review to determine its polarity. In this case, positive and negative terms were identified by querying a collection of dictionaries. In the same paper, a second method that uses Support Vector Machines (SVMs) was also proposed, and it was shown that this algorithm performs significantly better than the term-counting method with valence shifters.

In their paper, Pang et al. (2002) and Pang and Lee (2004) have compared the performance of various classifiers when determining the sentiment of a document, and also found that SVMs were generally the best approach. Unigrams, bigrams, part-of-speech (POS) tags and term positions were considered as features, and using unigrams alone yielded the best results. Thus, SVMs have repeatedly emerged as remarkably effective tools for performing sentiment analysis, even when using very simple features in the classifier.

Considering these related works, this project has been implemented using the classifiers SVM, Logistic Regression (since these two methods are considered to yield good results) and Naive Bayes for obtaining the sentiment of review texts after constructing a vocabulary of words from the training data using Bag-of-words or the Unigram model in Python.

III. BASIC IDEA AND DATASET

The project revolves around extracting sentiments for restaurant reviews using Bag-of-words model to extract the features from the text and giving it as input to machine learning classifiers namely Logistic Regression, SVM and Naive Bayes which classify the test reviews into Positive or Negative labels (Binary classification) based on the input features. The dataset used is from the Yelp Dataset Challenge (https://www.yelp.com/dataset_challenge) containing over 2.7M reviews for 86K businesses.

A. Input

The Yelp Dataset consists of several tables (.json format) namely business, checkin, review, tip and user. The review dataset is of interest to this project. It consists of various columns such as user_id, review_id, business_id, text, stars, date, type and so on. Since the dataset is huge, Hadoop Streaming with Python was used on Cloudera cluster to extract limited number of reviews of around 10 restaurants. During the process, an extra column called 'Sentiment' was introduced based the star rating given by the user ('Positive' for star rating between 3 and 5 whereas 'Negative' for ratings 1 and 2). This explains the labeled dataset required by supervised machine learning techniques. The restaurant with most number of reviews (5472 reviews) is being used as training data and the reviews of the remaining 9 restaurants (32657 reviews) as testing data.

IV. METHOD

The approach is to solve the sentiment classification task for restaurant reviews using bag-of-words model and machine learning techniques. Under a machine learning framework, we have a dataset of m instances (each review is an instance) $((x_1, y_1),\ldots,(x_m, y_m))$, where x_i is the feature vector extracted from the ith data instance, and y_i is the label for that particular data instance. The set of words occurring most frequently in all of the reviews constitutes the vocabulary (i.e., each of the words in the vocabulary is considered as the input feature). The feature vector captures the number of appearances of each feature from the vocabulary (For example: if the vocabulary has the features

'great', 'bad', 'awesome', 'disappoint', 'food' and the review text is 'I was disappointed with the food', the feature vector for this instance would be [0,0,1,0,0,1] (Although, preprocessing techniques like removal of stopwords/punctuations, stemming are involved before arriving at the feature vectors)). Selection of features will be explained shortly. The labels in this task are "positive" or "negative" for binary classification. Below is the diagrammatic view of the method.

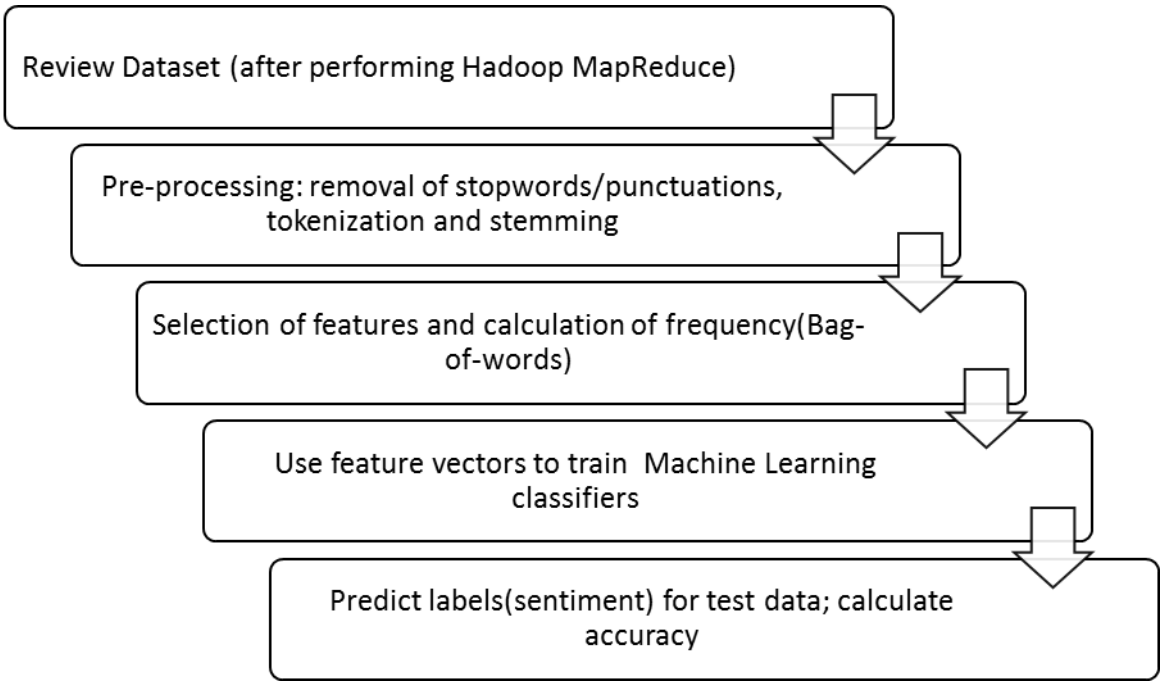

Figure 1. Method (Diagrammatic view).

A. Feature Extraction

In linguistics, a corpus is a large and structured set of texts (reviews in this case). They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory. This necessitates the processing of reviews using Bag-of-words model.

Bag-of-words model: this is a model that simplifies representation, generally used in Natural Language Processing and Information Retrieval. In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity. Prior to the selection of features, pre-processing steps such as removal of stopwords and punctuations, conversion to lowercase letters, tokenizing and stemming (reducing words to their lexical roots) are done. This is followed by the selection of the features (words that occur more frequently when compared to other words) and consequently, the feature vectors.

B. Machine Learning

As a recap, once a vocabulary of n words is obtained, for each text review an n-dimensional feature vector is extracted, where the i-th index is the number of occurrences of the i-th word in the document (if it exists), and 0

otherwise. These n-dimensional feature vectors form the input to the machine learning classifiers namely Logistic Regression, SVM and Naive Bayes.

Logistic Regression: According to [5], a binomial logistic regression (often referred to simply as logistic regression), predicts the probability that an observation falls into one of two categories of a dichotomous dependent variable based on one or more independent variables that can be either continuous or categorical. Logistic regression can be seen as a special case of the generalized linear model and thus analogous to linear regression.

SVM: Support Vector Machines are highly effective in many research and application domains, including text categorization. They are known to be one of the best classifiers. The basic idea behind SVMs is to find a separating hyperplane with the largest margin in a given higher-dimensional feature space. The search for this hyperplane corresponds to a constrained optimization problem.

Naïve Bayes: Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is assumed that the value of a particular feature is independent of the value of any other feature, given the class variable.

V. RESULTS

Below is the table showing the comparisons of the precision, recall and accuracies of the three classifiers Logistic Regression, SVM and Naïve Bayes. A total of 32657 reviews have been considered in the testing dataset among which 4478 express Negative sentiments while the remaining 28179 express Positive attitudes (The true labels were obtained from the initial phase of MapReduce where they were assigned based on the star rating given by the user). Based on the labels predicted by the classifiers and the true labels, accuracies were calculated for three cases (for differing number of input features) and from the table, it can be seen that Logistic Regression and SVM perform way better than Naïve Bayes.

IF - INPUT FEATURES ; NEGATIVE(N) - 4478 ; POSITIVE(P) - 28179

CLASSIFIER			PRECISION	RECALL	F1-SCORE	ACCURACY
LOGISTIC REGRESSION	IF-200	N	0.51	0.40	0.45	86.53
		P	0.91	0.94	0.92	
	IF-500	N	0.55	0.49	0.52	87.52
		P	0.92	0.94	0.93	
	IF-1000	N	0.56	0.57	0.57	87.93
		P	0.93	0.93	0.93	
SVM	IF-200	N	0.47	0.39	0.42	85.69
		P	0.91	0.93	0.92	
	IF-500	N	0.50	0.47	0.48	86.37
		P	0.92	0.93	0.92	
	IF-1000	N	0.45	0.58	0.51	84.39
		P	0.93	0.89	0.91	
NAIVE BAYES	IF-200	N	0.34	0.61	0.44	78.43
		P	0.93	0.81	0.87	
	IF-500	N	0.30	0.77	0.43	71.66
		P	0.95	0.71	0.81	
	IF-1000	N	0.28	0.85	0.42	68.08
		P	0.96	0.65	0.78	

Figure 1I. Tabular form of accuracies of the three classifiers Logistic Regression, SVM and Naïve Bayes.

VI. CONCLUSION

In this project, an attempt has been made to classify the sentiments of restaurant reviews using the Natural Language technique bag-of-words and three different Machine Learning Techniques Logistic Regression, SVM and Naïve Bayes for binary classification. Things that I learned include: Extracting relevant columns from big data using Hadoop Streaming in Python, extracting features and classifying data using simple and efficient data analysis tools in Python namely scikit-learn, working with data frames using pandas and performing predictive analysis of large-scale data using the open source distribution Anaconda. Also, I understood why it is important to perform tokenization and stemming while extracting features into the vocabulary.

ACKNOWLEDGMENT

I express my gratitude and thanks to Prof. Dan I. Moldovan at The University of Texas at Dallas for his immense support in implementing this project.

REFERENCES

[1] Andr´es Cassinelli, Chih-Wei Chen(2009), "Boost up! Sentiment Categorization with Machine Learning Techniques"
[2] Parul Singh, Ilakya Palanisamy, Mukund Chillakanti, Abhinava Singh, "Predicting Yelp Restaurant Ratings from Review Text: Applying NLP and Machine Learning techniques to better understand the relationship between a restaurant's rating and its review text"
[3] Bo Pang and Lillian Lee, Shivakumar Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques"
[4] Abinash Tripathya, Ankit Agrawalb, Santanu Kumar Rath(2015), "Classification of Sentimental Reviews Using Machine Learning Techniques"
[5] Wikipedia for Logistic Regression, Naïve Bayes, SVM classifiers.
[6] Monisha Kanakaraj and Ram Mohana Reddy Guddeti(2015), "NLP Based Sentiment Analysis on Twitter Data Using Ensemble Classifiers"