|  |
| --- |
| **CSE 5334 – Data MINING** |
| **Performing Aspect Based Sentiment Analysis On Restaurant Reviews Using Naive Bayes Classifier** |
| Vidyadhar Angadiyavar and Rajesh Nayak |
|  |
|  |
| **5/5/2015** |

|  |
| --- |
|  |

**MOTIVATION AND OBJECTIVES**

Twitter has a section called Trends, which displays the various trending current topics. We planned to calculate similar ‘trending’ restaurants by using the Yelp dataset. These trending restaurants would be selected based on the ‘weights’ calculated from a list of positive and negative reviews. Based on these trending restaurants, there could well be a website or a mobile application, which will display the trending restaurants for every week, for a particular city.

By using the Yelp data set, we plan to determine which of the reviews for a restaurant is positive or negative. A user can select a particular city and will be able to view a list of restaurants in that city. Based on the restaurant which the user selects, he will be able to view a list of positive, negative and neutral reviews. A user can also view which aspect of the restaurant selected is good or bad. The various aspects which are considered are ambience, price, food and service.

**DATA MINING/ANALYSIS TASKS TACKLED** [1]

A Naïve Bayes Classifier was used on the training data set obtained at the SemEval website (<http://alt.qcri.org/semeval2015/task12/> ) and the PHP/ir website (<http://www.phpir.com/bayesian-opinion-mining/> ). Sentiment Analysis was used to determine whether the review is positive or negative based on the above training data. The SemEval training data consists of various aspects which we could use as train data, such that we could calculate the positive and negative polarity for the new incoming test data based on the aspects which it consists.

Following are the steps which are used to implement the Naïve Bayes Classifier –

We perform the following tasks on the train data before the actual test data is processed [1].

* Firstly, we find out the number of sentences which have the ‘positive’ tag on them. We scan the train data and store all the sentences which are positive, and maintain the count for those sentences. (Say, the count obtained is P)
* Secondly, we repeat the above step for all the sentences with a ‘negative’ tag on them. We obtain the count for this as well, and store it. (Say, the count obtained is N)
* Having found the count of both positive and negative sentences, we calculate the probability of a sentence being positive or negative.
  + - Probability (Positive) = P/P+N
    - Probability (Negative) = N/P+N
* Maximum likelihood smoothing Naïve Bayes estimate is calculated by using the below formula [1] -

**γ(ω| α) = (count(ω,α) + 1) / (count(α) + | V |)**

where **count(ω,α)** : Number of occurrences of ω in training documents from class α

**count(α)** : Number of words in that class

| V | : Number of terms in the vocabulary.

This method (also known as Laplace Smoothing) is used to eliminate the Zero Probability problem.

* Example [1]:
  + **γ (like | pos) = (3 + 1) / (25 + 31) = 4/56 = 1/14**
  + **γ (like | neg) = (0 + 1) / (12 + 31) = 1/43**
* Having calculated the estimate for all the words in the training data, the Posteriori Probability [1] is calculated for the sentences in the test data by multiplying the results obtained during Laplace smoothing (previous step) and the Probability of a Positive/Negative sentence (Step 1).
* The resulting value is nothing but the Positive/Negative polarity of that particular ‘test’ sentence. If the positive polarity value is higher, the sentence is considered positive. If the polarity of the negative sentence is higher than that of the positive one, the sentence is considered negative.
* Although the whole sentence is deemed positive or negative based on the above polarity value, there might be aspects in the sentence which say otherwise. For example, consider the sentence “I would not consider coming back to this restaurant if the waiter was not so courteous”. Overall, it gives a picture where the customer is not satisfied with any aspect of the restaurant but for the service.
* Thus, we also display the polarity of the aspects in a sentence, whether it is positive or negative. We do that by again considering the training data for positive reviews and negative reviews for each of those aspects. (Hence, a separate positive/negative training data for food, service, price etc)
* Bag of Words have been used as a filter to the aspect based sentiment analysis technique.

**DESIGN OF METHODS**

The design consists of the following modules -

* The restaurant data which is available on the Yelp data set is stored in a **database**. This database will contain all the information which will be required to obtain the reviews (For instance, the restaurant city, state, restaurant name etc.)
* The train data which is used for aspect-based sentiment analysis is read from a CSV file and the training data used for normal sentiment analysis Is stored in a TEXT file. The train data is read by the **PHP** code and is stored in **arrays**.
* The PHP file reads the Yelp restaurant data from the database and directly displays it on the front end. This would make the project run faster.
* The PHP file then displays all the reviews for a particular restaurant on a single page, and given an option to the user to check the polarity for each of those reviews.
* When the user clicks on the option of viewing the polarity, the code navigates to another PHP file where the aspect based sentiment analysis has been implemented. The PHP also makes use of the already existing train data for aspect-based sentiment analysis. This will display the reviews on the front end with all the aspects and their corresponding results (Positive or Negative).
* The user also has an option of viewing the polarity of the whole review, which redirects to a different logic. The PHP makes use of the already existing train data for regular sentiment analysis. The PHP with this particular logic will then display the polarity of the whole review on the front end.

Database Implementation

(Mongo DB)

**Yelp Dataset**

NO SQL

Training Data (Aspect Based Sentiment Analysis)

Training Data (Sentiment Analysis)

Perform Aspect Based Sentiment Analysis using Reviews as Input

Perform Aspect Based Sentiment Analysis using Reviews as Input

Displaying Reviews

PHP Home Page

Restaurant Data and Reviews

**IMPLEMENTATION OF METHODS**

The implementation of the sentiment analysis has been done as follows –

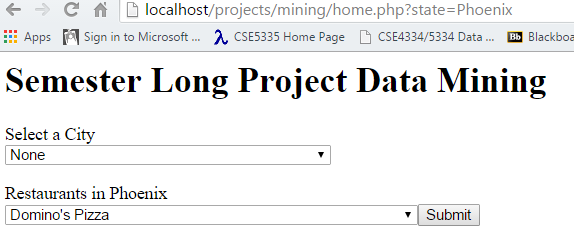
* The restaurant data which is available on the Yelp data set is stored in a **database**. This database will contain all the information which will be required to obtain the reviews (For instance, the restaurant city, state, restaurant name etc.)
* The train data which is used for aspect-based sentiment analysis is read from a CSV file and the training data used for normal sentiment analysis Is stored in a TEXT file. The train data is read by the **PHP** code and is stored in **arrays**.
* The PHP file reads the Yelp restaurant data from the database and directly displays it on the front end. This would make the project run faster.
* The PHP file then displays all the reviews for a particular restaurant on a single page, and given an option to the user to check the polarity for each of those reviews.
* When the user clicks on the option of viewing the polarity, the code navigates to another PHP file where the aspect based sentiment analysis has been implemented. The PHP also makes use of the already existing train data for aspect-based sentiment analysis. This will display the reviews on the front end with all the aspects and their corresponding results (Positive or Negative).
* The user also has an option of viewing the polarity of the whole review, which redirects to a different logic. The PHP makes use of the already existing train data for regular sentiment analysis. The PHP with this particular logic will then display the polarity of the whole review on the front end.

**RESULTS AND EVALUATION**

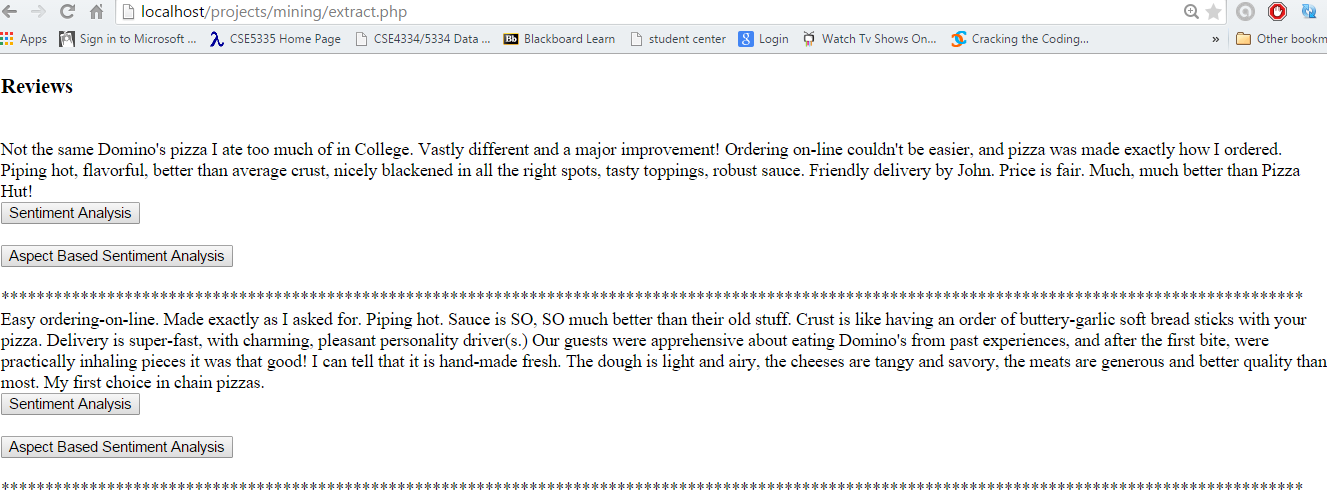
1. We have implemented sentiment analysis and aspect based sentiment analysis as 2 different techniques. The regular sentiment analysis takes the review as the input and comes up with a single polarity for the whole sentence.
2. On the other hand, the aspect based sentiment analysis displays the sentiment based on the aspect (depending on whether the aspect is applicable for the sentence). To determine whether an aspect is applicable for the sentence or not, we have used bag of words. If those bag of words are present in the review, the aspect sentiment is calculated.
3. We could achieve approximately 70% accuracy by using Naïve Bayes for sentiment analysis. We could have achieved more than 80% by implementing the Neural method on top of Naïve Bayes.
4. Due to time constraints, we could not display the list of trending restaurants by summing up the number of positive and negative reviews for every restaurant and implementing a weighting mechanism.
5. The performance of the application depends on the number of reviews a certain restaurant has. As of now, for a restaurant having approximately 20 reviews, the sentiment analysis takes approximately2-3 seconds for a review.
6. The correctness of the result can be improved by some percentage (5-10%) by adding more training data for the aspect based sentiment analysis. We could not find the data over the internet for aspect based restaurant reviews.

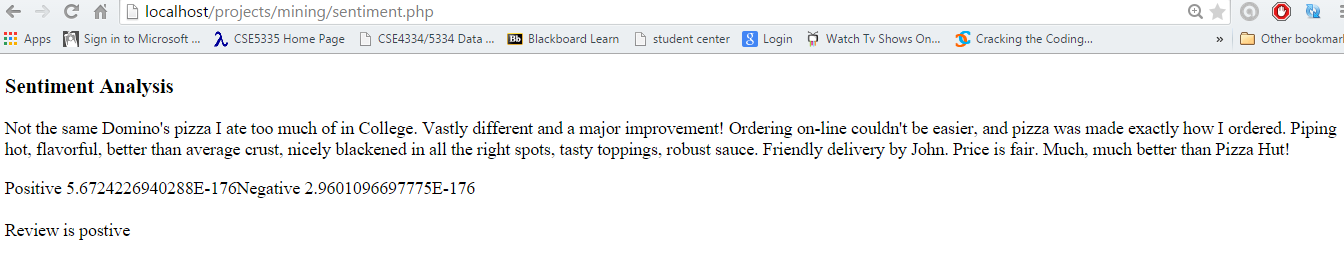
**PRESENTATION/VISUALIZATION OF THE OUTCOME**

1. **Home Page which displays the Cities and the corresponding restaurants in those cities.**

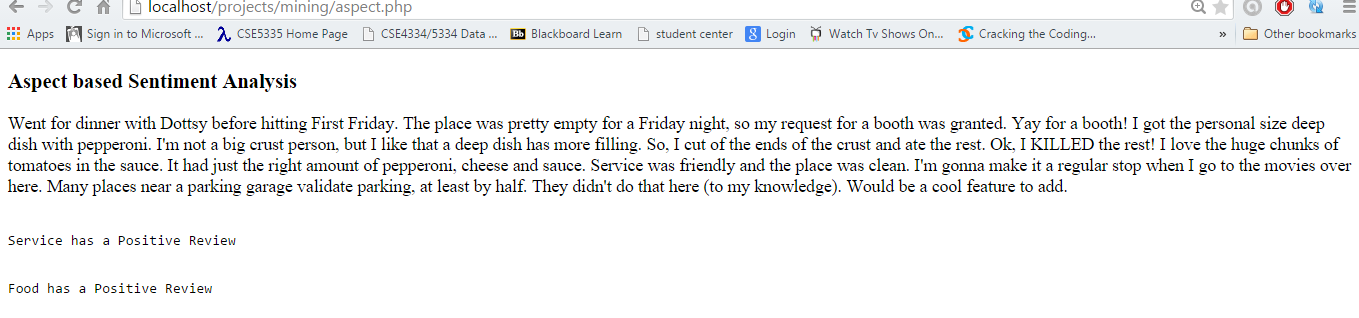
****

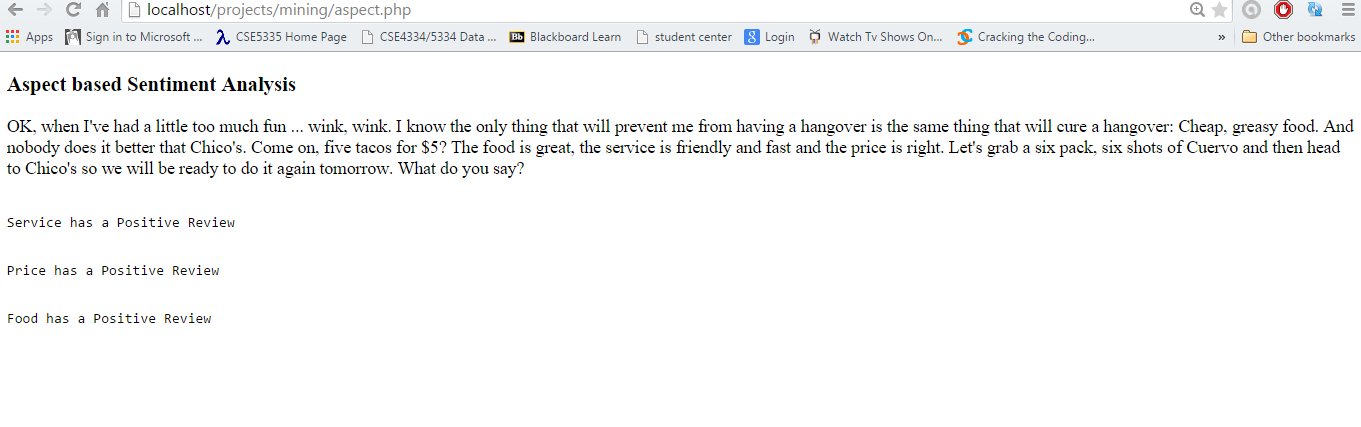
1. **The Extract.php displays all the reviews that that particular restaurant. From here, the user can choose to see the Aspect Based Sentiment Analysis and Regular Sentiment Analysis for those reviews**

****

**3. Upon selecting the regular sentiment analysis, the review and the polarity of the statement being positive and negative is displayed.**

**4. If the user selects Aspect Based Sentiment Analysis, the review is displayed and the result (Positive/Negative) of aspects of the review is displayed, based on which aspect is applicable.**

****

****

**CHALLENGES FACED**

* Adequate training data was not available to perform aspect-based sentiment analysis. Hence, the Naïve Bayes classifier for Aspect-Based sentiment analysis is not working as accurately as we had expected.
* We were not able to achieve the grammatical based sentiment analysis (For example “The place was good…not!” would be considered positive because ‘good’ has more weight toward the positive than ‘not’ has on the negative side.)
* Sarcasm cannot be detected using the method that we have used.

**POSSIBLE AREAS OF IMPROVEMENT**

* As we are displaying the positive and negative reviews of all the restaurants which are selected by the user, we could calculate the percentage of positive and negative reviews and then based on some weighting mechanism we could find out the trending restaurants for that particular city.
* Combining Naïve Bayes classifier with another method, such as neural networks will improve the correctness of the sentiment analysis.
* We could build a search engine for the project which will display the reviews based on the city and state selected.
* We could display a pie-chart/bar graph of the trending restaurants and the history of previous trending restaurants, so that the users will have an idea which place to visit often.
* We couldn’t deploy the web application on any servers except on the local machine because of additional billing for mongodb instances.

# Works Cited

1. *Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier.* **Lina L. Dhande, Dr. Prof. Girish K Patnaik.** 4, SSBT’s College of Engineering and Technology, Bambhori, Jalgaon, Maharashtra, India : International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2014, http://www.ijettcs.org/Volume3Issue4/IJETTCS-2014-08-25-138.pdf, Vol. 3. 2278-6856.

2. Yelp Dataset Challenge. *Yelp.com.* [Online] [Cited: 04 03, 2015.] http://www.yelp.com/dataset\_challenge.

3. SemEval 2015. *alt.qcri.org.* [Online] [Cited: 04 03, 2015.] Aspect Based Sentiiment Analysis : Idea obtained from here. http://alt.qcri.org/semeval2015/task12/.

4. SemEval-2014 ABSA Restaurant Reviews - Train Data. *Meta Share.* [Online] [Cited: 04 03, 2015.] Training Data for Aspect Based Sentiment Analysis Obtained here. http://metashare.ilsp.gr:8080/repository/browse/semeval-2014-absa-restaurant-reviews-train-data/479d18c0625011e38685842b2b6a04d72cb57ba6c07743b9879d1a04e72185b8/.

5. **Barber, Ian.** Bayesian Opinion Mining. *PHP/ir.* [Online] 01 01, 2010. [Cited: 04 03, 2015.] Training Data set for regular Sentiment Analysis and the Naive Bayes concept of Sentiment Analysis learnt from here. http://www.phpir.com/bayesian-opinion-mining/.