HACETTEPE UNIVERSITY COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
BBM 495
RESTAURANT SENTIMENT ANALYSIS
PROGRESS REPORT

Nehil Danış 21327876
İlayda Çavuşoğlu 21327805
Ebru Uçgun 21328524

1.1. Task

The model to perform sentiment analysis based on Twitter tweets depending on restaurants is developed. English is choosen as a language of this task. The model will be trained on positive and negative tweets, then when a user asks for a spesific restaurant, the program that will be implemented by us will collect the latest tweets from Twitter and tell the user whether they are positive or negative.

1.2. Datasets and Evaluation

In this project the data that is retrieved from Yelp dataset is used. For the evaluation yelp_academic_dataset_business.json and yelp_academic_dataset_review.json files is used. The necessary attributes from the yelp_academic_dataset_business.json file are starts and catagories and from the yelp_academic_dataset_review.json file are stars and text.

1.3. Methods

In this project the .json files that are retrieved from Yelp dataset will be used for both training and testing as well. But the data that is used for training can not be used for testing. All the emoticons is transformed into its characters. For example :) :D :P :(:S etc. For the emoticon characters, we will create an enum dataset. In this way we will specify each emoticons whether its positive or negative. Every misspelling and abbreviation will be fixed. For this process, NLTK will be used. The negative 'not' in the words will be seperated and wil be added to the next word by using NLTK library. For example: "**I don't like this.**" will be turned into "**I do not_like this.**" The stopwords which occurs in the text will be deleted by using NLTK library.

In the Yelp dataset we have reviews and ratings for the comments. By using this rating one can understand whether the comments is possitive or negative. However to compare this rating with the other user's rating, the ratings has to be normalized. We will use a basic baseline for this process.

x = u + r +c
x --> normalized rating
r --> restaurant's average rating – average raiting of all reviews
u --> customer average rating

Three way classification in one step involves estimating a probability distribution overall categories. For this classification process the Naive Bayes classifier is implemented by us. In Naive Bayes classifier an object's class is found.It helps to find out to which class the object is beloged to.

To compare accuracy of the Naive Bayes method, SVM(support vector machine) will be used. To implement SVM sklearn library is going to be used.

While we are implementing Naive Bayes and SVM methods, we are going to use unigram and bigram language models.

The small part of Yelp dataset will be used only for testing process. In this way, we will compute the accuracy of our program.

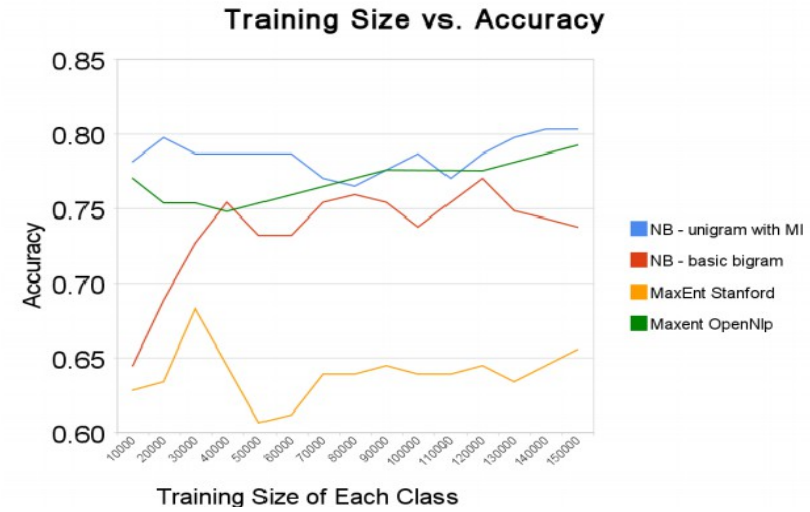
Based on the accuracy of our program, we will be able to say the comment that we received from Twitter is either negative or positive.

Basically the aim of this task is classification. The comments about a restaurant will be scaled as positive or negative.

1.4. Related Work [2]

An article that covers nearly the same subject with our project is found. In this article, they build an algorithm that can accurately classify Twitter messages as positive or negative. Their hypothesis is that we can obtain high accuracy on classifying sentiment in Twitter messages using machine learning techniques.

They collected their own set of data for the training. They marked tweets manually. They used several different classifiers.



They also explored Support Vector Machines using Weka software. They tested SVM with a unigram feature extractor, and achieved only 73.913% accuracy. They used unigram and bigram language models.

They thought that POS tags would be a useful feature since how you made use of a particular word. They gave an example, ‘over’ as a verb has a negative connotation whereas ‘over’ as the noun, would refer to the cricket over which by itself doesn’t carry any negative or positive connotation. On the Stanford Classifier it brought their accuracy up by almost 6%. The training required a few hours however and they observed that it only got the accuracy down in case of NB.

They extended the Naive Bayes Classifier to handle 3 classes: positive, neutral, and negative. However their results were terrible. The classifier only obtained 40% accuracy. They thought that this is probably due to the noisy training data for the neutral class.

1.5. Progress

So far the datasets are used in this task are found and reviewed. The .json files transformed into .csv files. The code was implemented in Python. json and csv libraries are imported. Other than that the methods that will be used on the project are decided.

1.6. References

[1] Saif, Hassan, Yulan He, and Harith Alani. "Semantic sentiment analysis of twitter." *The Semantic Web–ISWC 2012* (2012): 508-524.

[2] Go, Alec, Lei Huang, and Richa Bhayani. "Twitter sentiment analysis." *Entropy* 17 (2009): 252.