Twitter Sentiment Analysis

Devanshu Shah COMPS-A

Type

Sentiment analysis falls under the category of ***supervised learning*** as we are going to provide labelled data to the machine learning algorithms and expect a binary output. We will be doing ***binary classification*** on tweets and output whether they are positive or negative.

Data

The data that we will be using is going to be two files named short_pos.txt and short_neg.txt. The short_pos file contains all positive, short, single lined statements. Similarly the short_neg file contains all negative statements.

Data Pre-Processing

We will use the ***natural language toolkit (NLTK)*** for processing the data. We apply the following steps to our data:-

- Tokenize all words using the word_tokenize command.
- Find the 3000 most used adjectives using part of speech tagging and FreqDist features available in NLTK for both the text files combined and add them to a list named word features.
- Next we will label all the single lined reviews as either positive or negative by grouping a dictionary of word features(features as key and true/false as values depending on whether they are present in the review) found in that review and its category(positive/negative) in a tuple.
- In this way we obtain a training set which helps the algorithm predict which words are more frequently used in a positive comment and which of them are part of negative comments.

In this manner we have obtained a labelled DataSet which can be used to train algorithms for predicting sentiments.

Training Algorithms

We will be using the following algorithms to train our data set.

1. **Bernoulli NaiveBayes Algorithm**

   The Bernoulli NaiveBayes algorithm is based on the ***Bernoulli Distribution***. The main feature of Bernoulli NaïveBayes is that it accepts features only as binary values like true or false, yes or no, 0 or 1. This works best for us as the feature set we have defined is a dictionary which maps the words against their Boolean value i.e. whether the word exists in the review or not. Same reasoning follows for not using Multinomial NaiveBayes theorem.
   The Bernoulli NaiveBayes Classifier is based on the following rule:

   $$P(x_i|y)=P(i|y)x_i+(1-P(i|y))(1-x_i)$$

   In Bernoulli Naive Bayes each feature is treated independently with binary values only, it explicitly gives penalty to the model for non-occurrence of any of the features which are necessary for predicting the output y.

2. Logistic Regression

   Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more independent variables.

3. Linear SVC

   SVM is a supervised machine learning algorithm that can be used for both classification or regression challenges

   The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyperplane that divides or categorizes your data into two classes. For our case the two classes are positive and negative.

Pickling

After the classifiers have finished training we pickle the classifiers into a (.pickle) file which converts the data into bytes. By doing this, we do not have to train the data everytime we run the python script. We only need to open the pickle file and set the

classifier variable to the data in the file. This greatly decreases the processing time of our file.

Creating a Function to classify text

Then we create a voted classifier which return the mode of the votes given by all the classifiers that we have created and also the confidence of the vote.

Finally we create a function named sentiment() which takes the text as a parameter and uses the voted classifier to return the classification of text along with the confidence of the vote.

Using Twitter Data

For working with twitter data we will be using the python module tweepy.
We will need to create an app at apps.twiiter.com and obtain the consumer & access key and consumer & access secret to be able to work with the twitter API.

Now we create an auth variable which is an instance of the OAuthHandler class and pass in ckey and csecret to initialize. And set the access token to the auth variable.

We can now stream data from the twitter API and convert data to python data type by running json.load() on the data. After this we access the text of the tweet and pass it into the sentiment() function that we defined above and obtain the sentiment value and the confidence of the vote.

Conclusion

Thus we have conducted sentiment analysis on twitter data using the Bernoulli naïve bayes, logistic regression and Linear svc classifiers.