Applied Machine Learning- Lab Report

Group 10
Group members:
G Rohit (1PI13IS038)
Ekta G Dharamshi (1PI13IS037)
Raksha P Rao (1PI13IS081)
Swetha B (1PI13IS114)

Accuracy:
Naive Bayes Classifier: 99%
Decision Trees: 98.33%

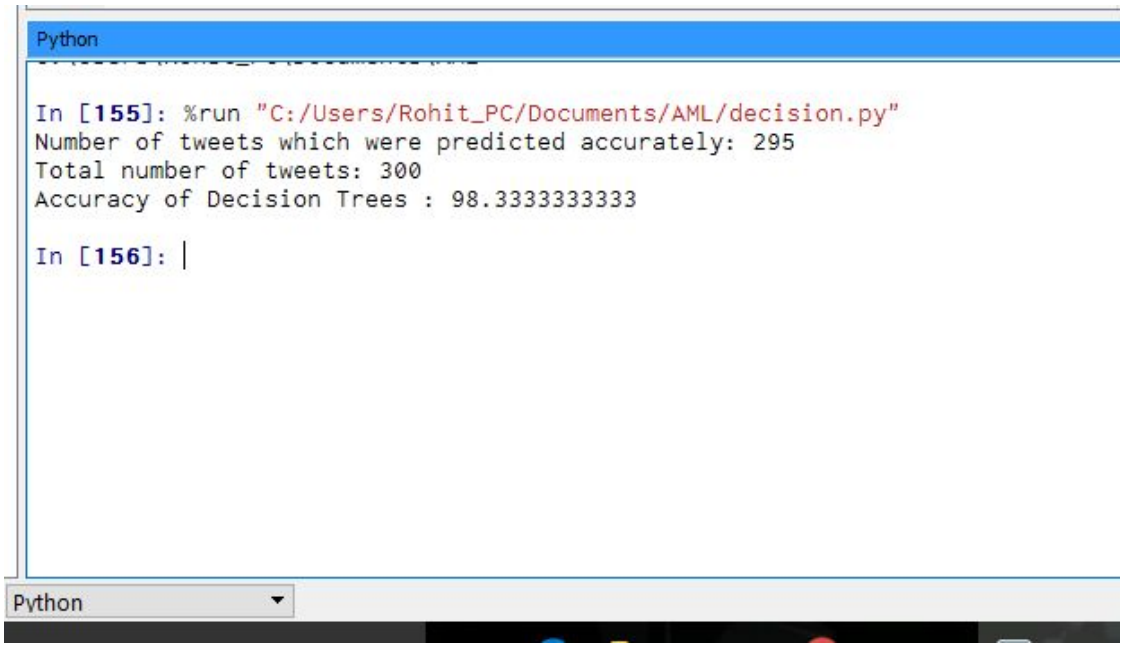
Procedure

Pre-processing:

The dataset provided was noisy, i.e, some of its contents do not help in classification, for example, retweets, hashtags, hyperlinks etc. So we removed retweets, converted #word to word and removed hyperlinks. Then we removed punctuations, whitespaces and stopwords from the tweets. We used this filtered data for training and testing, where 80% of the given 1500 tweets were used to train and the rest to test.

Decision Trees:

Decision Trees are based on the reduction of entropy in a dataset. Entropy is defined as the unpredictability of information content. Decision trees are based on splitting the dataset on the basis of all the values of an attribute, while choosing the attribute in such a way that reduction in entropy (i,e, information gain) is maximum.
We implemented the ID3 algorithm. We achieved an accuracy of 98.33% in decision trees. We referred to "http://www.onlamp.com/lpt/a/6464".



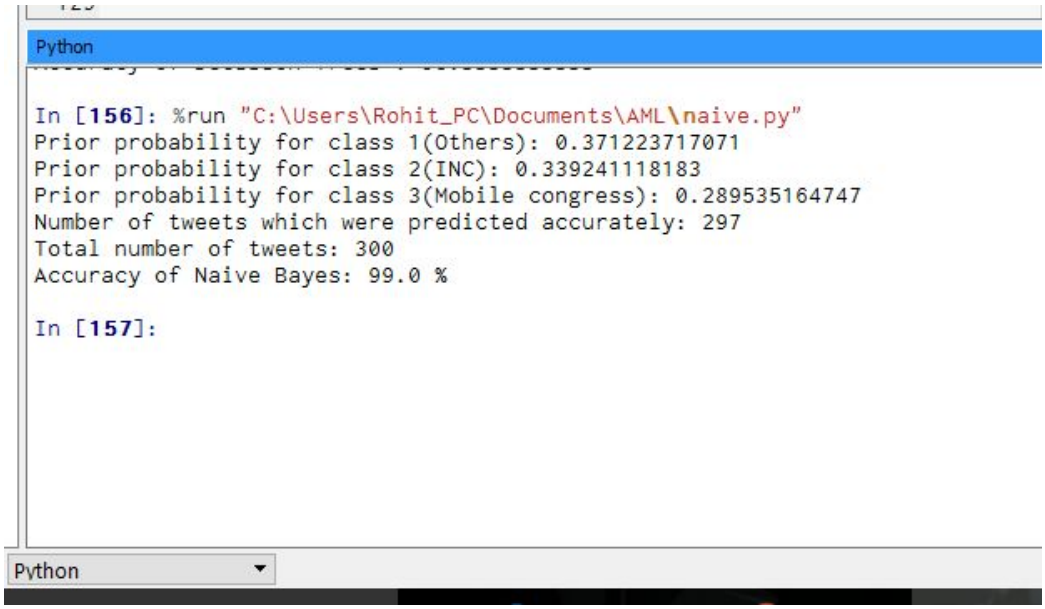
Naive Bayes Classifier:

The Naive Bayes Classifier is based on the Bayes' theorem, which can be stated as follows:

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

Here, suppose A represents the class of the particular tweet. B represents the feature vector of the tweet. Then, given the feature vector of a tweet, the probability that the tweet belongs to class A is given by the above formula.
P(A) is said to be prior probability. It represents the probability of class A in the training set. P(A|B) represents likelihood, i.e., the chance that the test tweet belongs to class A (based on similar tweets). P(B) represents evidence, which is the summation of prior probability * likelihood over all the classes.
This theorem assumes that the value of a particular feature is independent of the values of other features. This is not always true, and hence this assumption is said to be naive.

We achieved an accuracy of 99% with this classifier. We found out the prior probability of each class and likelihood of each feature. We then calculated posterior probability of the test tweet. Argmax was obtained across the three classes to find the best matching class of the test tweet.



Conclusion

After implementing both the algorithms, we found out that both gave good results. We achieved slightly greater accuracy in the Naive Bayes classifier. We found that the Naive Bayes classifier is easier to implement and is a faster algorithm than Decision Trees. This was based on the observation that Decision Tree took nearly 3 mins to give us the final answer whereas the NB classifier did the same in a few seconds.