Research Methods in Social Science and Design
Sentiment Analysis Assignment

Himanshu Singh
2017291

In this assignment, nltk.NaiveBayesClassifier is trained and tested using labeled dataset. Initially the data is obtained from the third column of the files comprising of fake and real news headlines and labels of ‘fake’ or ‘real’ were given to them accordingly. This data was then pre processed. To feed the data to the classifier, featureset was made using all the words (varied to show its effect on accuracy) in the data. This feature set was then divided randomly into 800 and 60 rows of data for training and testing purposes respectively. Accuracy was recorded after each testing. This whole process was repeated about 10 times from the start to take an average of the accuracy. At the end of each round, accuracy is displayed along with the 5 most informative features.

Preprocessing

In preprocessing the data (headlines) undergo a lot of processes. These processes are

- Removal of HTML (components)
- Removal of punctuations
- Removal of stopwords (eg me,he etc, because stopwords have low predictive power)
- Lemmatizing of text (option of Stemming of text is also there in code)
- Transforming all the words to lowercase
- Splitting of lines into list of words

Assumptions

- There are no null columns

Observations

The accuracy seemed to wary for different combination and selection of processes in preprocessing.
Without any preprocessing (except splitting of lines into list of words), the accuracy was recorded as 63.3 %.
With preprocessing, the accuracy was recorded as 84.0 %

Accuracy improved by preprocessing the data as words like html tags, punctuation marks etc were not helping in classification of headlines into fake and news, They were not informative and had low predictive values.

Diving deeper in the preprocessing steps, accuracy was recorded for every with commenting/ removing of one of the preprocessing steps -
Without punctuation - 76.67%
Without stop words - 81.5 %
Without lemmatizing - 83.0 %
Without lowercase - 81.5%
It can be seen that removing punctuations had the maximum effect on the dip of the accuracy.

Selecting word features = all = 83%
Selecting word features = only top 1000 = 78 %
In the feature set, accuracy was seen to increases if all the word features were included.

Accuracy value

Condition	Accuracy
Without Preprocessing	63.3%
With Preprocessing	84%