

## CSCI 5521: REPORT FOR PROJECT 3

Submitted by: Venugopal Mani

Student ID: 5212648

### 1: CENTROID BASED CLASSIFICATION

#### DESCRIPTION

The implementation of centroid based classification in this project, follows a very similar approach to the clustering algorithm of Spherical K Means clustering.

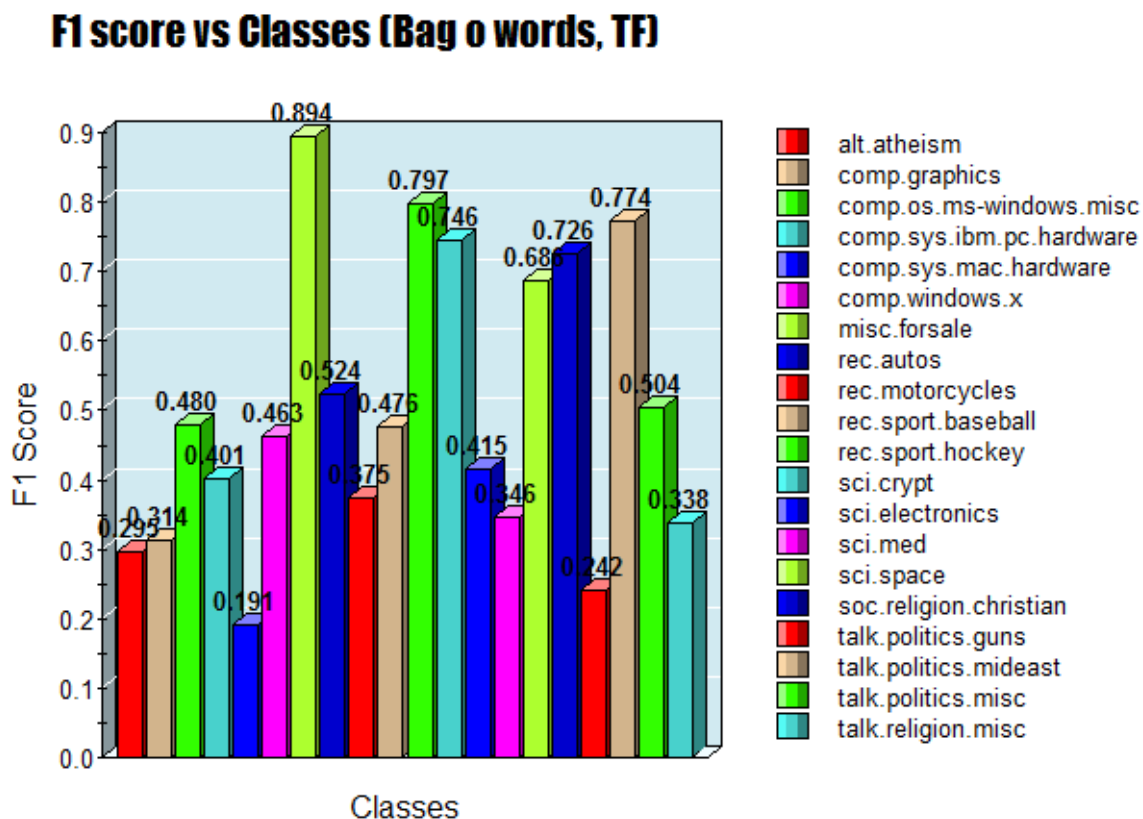
From the train set, classes are created – one for each label.

Each class stores in it – members and notmembers – two lists which have article objects which are members and not members respectively.

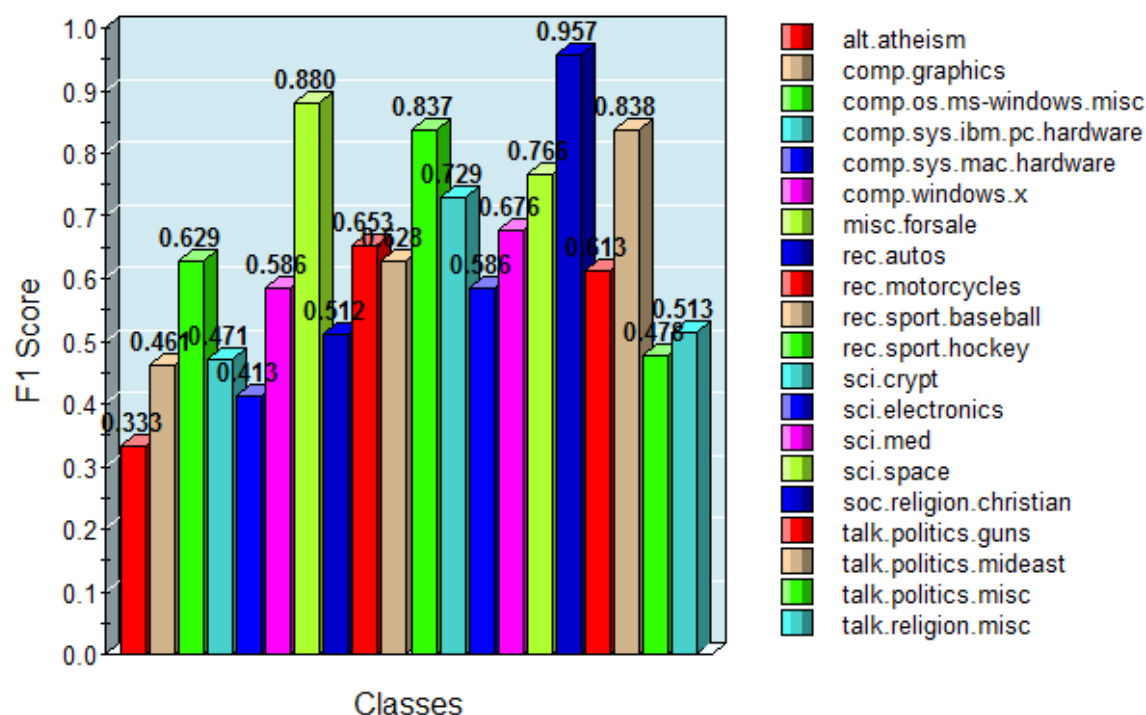
Now, the positive and negative centroid are computed using the computeCentroid method.

When an article from the test instance comes in , it is checked with all the classes and classified to the class which it has the greatest prediction score to.

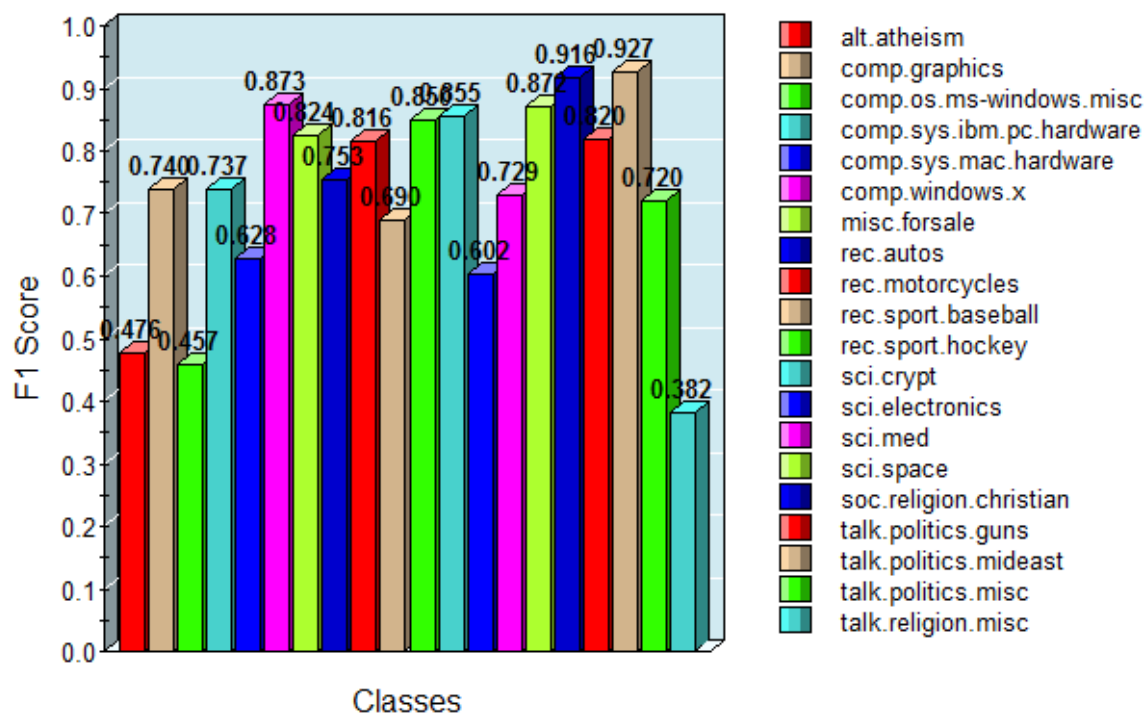
#### RESULTS



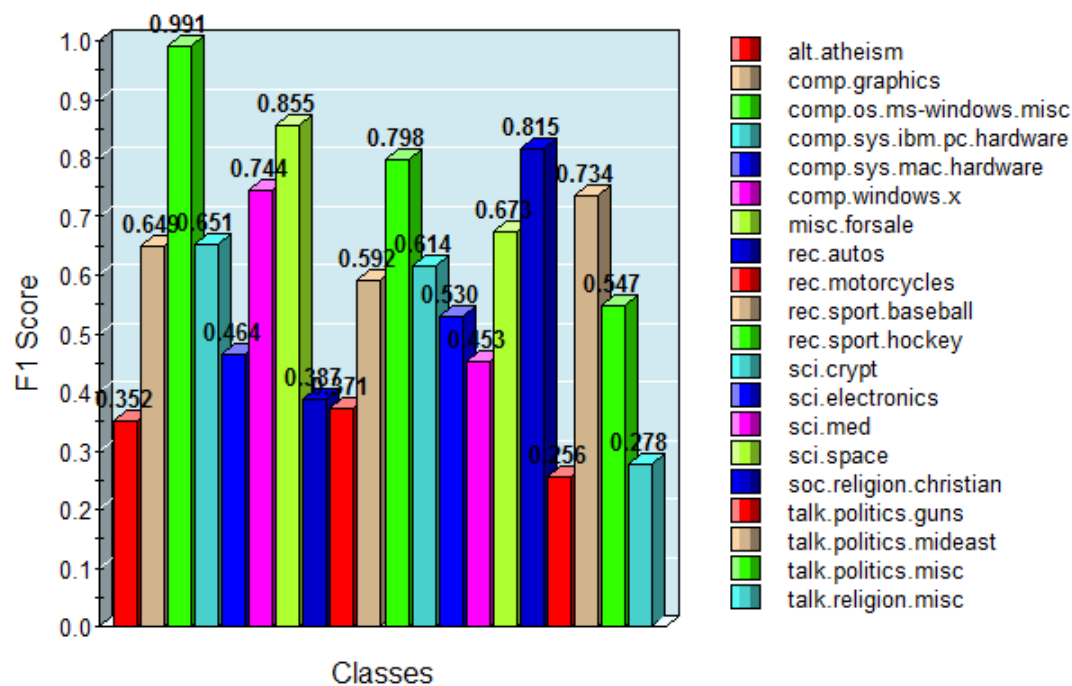
## F1 score vs Classes (Bag o words, Binary)



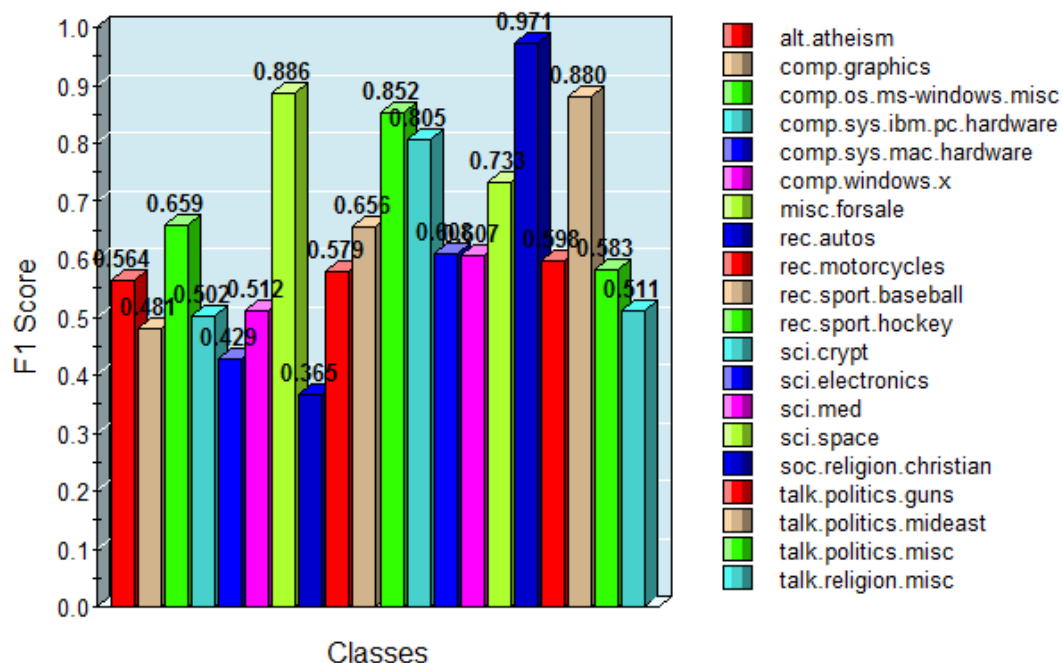
## F1 score vs Classes (Bag o words, TF-IDF)



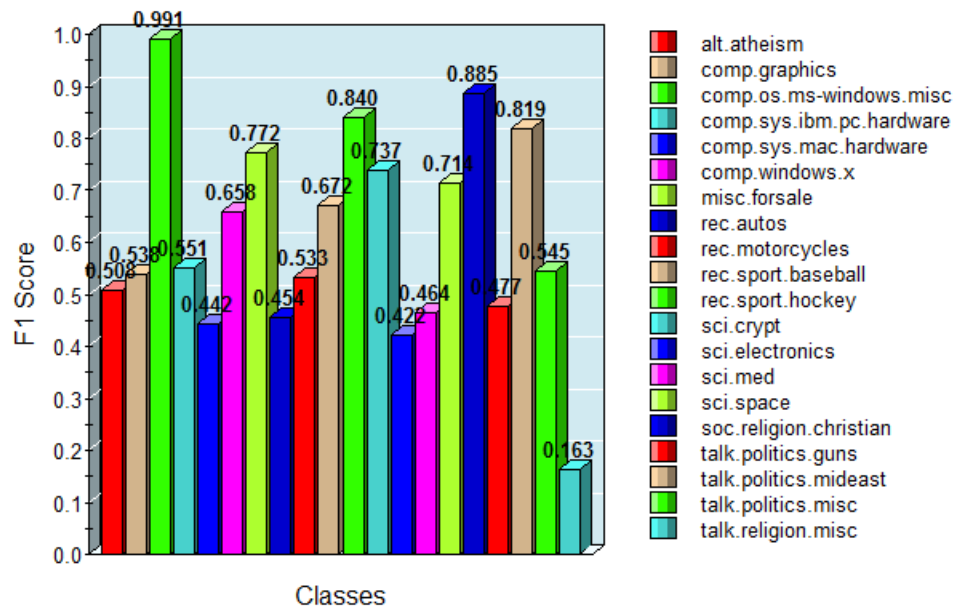
### F1 score vs Classes (n-grams, TF)



### F1 score vs Classes (n-grams, Binary)



## F1 score vs Classes (n-grams, TF-IDF)



## 2. RIDGE REGRESSION

### DESCRIPTION

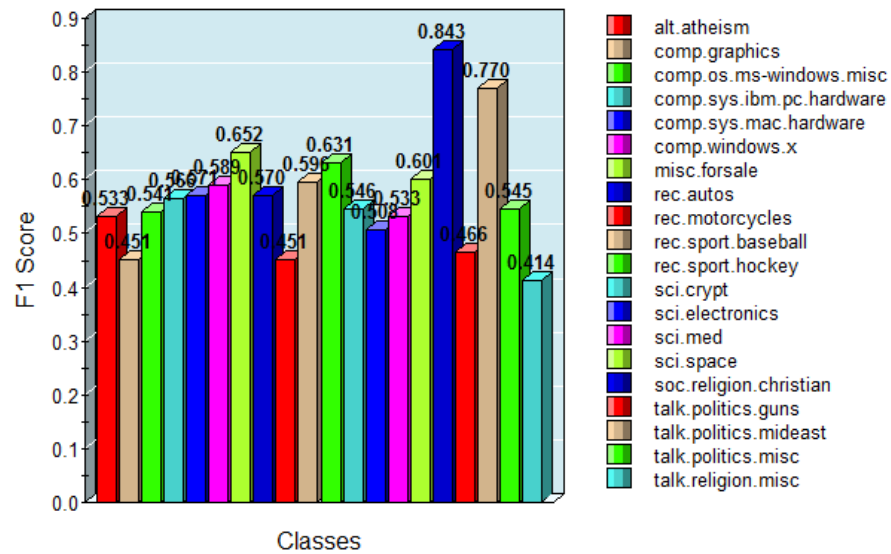
The term to compute is

$$w_k = (\sum a_{ik} - \sum a_{ik} \sum (a_{ij} * w_j) / \sum a_{ik} + \text{lambda})$$

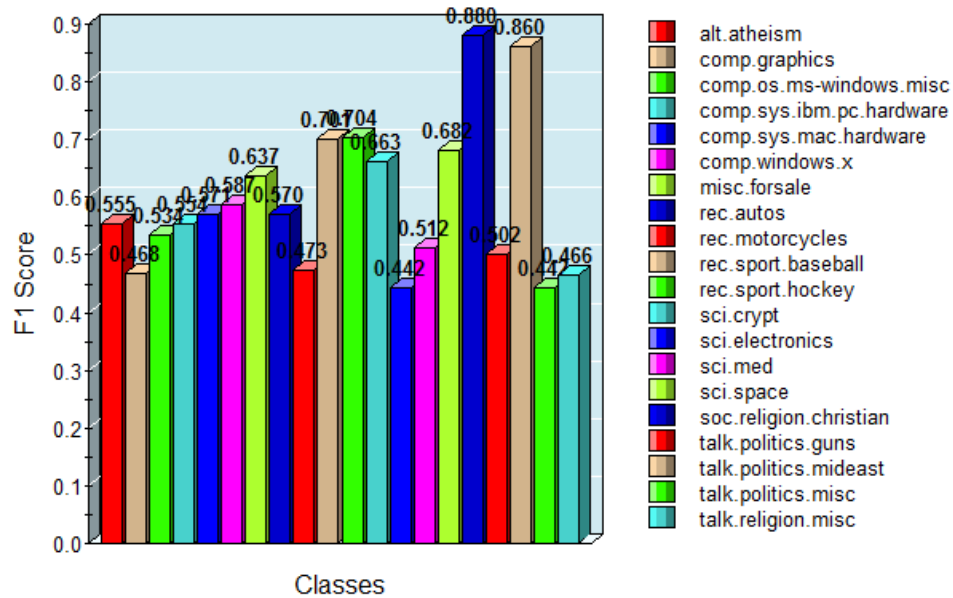
The model selection is not being done by the program itself and can be changed manually. Manually, the best average lambda was found to be 1.0

The following runs were done with f1 score vs classes for lambda value of 1.0 and bag of words representation.

### F1 score vs Classes (Bag o words, TF)



### F1 score vs Classes (Bag o words, Binary)



## F1 score vs Classes (Bag o words, TF-IDF)

