Team : AustereSpaiens **Email** : sanchitagarwal108@gmail.com,
rao.vidyadhar@gmail.com

1) Software Used: **Python 2.7,NLTK 2.0**

2) Features: **Most Frequent word list**

3) Similarity/Distance Measures: **Frequency Distribution,Chi-square Test**

3) Classifier :-**Naive Bayes Classifier**

4) References :HTTP://nltk.org/install.html,http://stackoverflow.com

---

Please describe the algorithms in details:

1) Pre-processing step:-

a)Read the Tweets. Removed the **Stop words, sms Abbreviations,URL** after tokening the tweets
b)Remove the substring of **#,@** and other punctuations.
c)Separated the tweets of the two kinds into separate list of sports and politics tweets.

2) Training Algorithm:Naive Bayes Classifier

a)separated the tweets into Politics and sports
b)Generated the tokens for each type of tweets
c)Calculated the **score** for both the politics and sports words from the list of tokens
d)generated bi-grams from the tokens for both sports and Politics features
e)selected the **best words pairs** from the generated features of bi-gram and unigrams based on **chi-square test**
f)Learned the Classifier.

Input Format:-<tweeetid label tweet> in txt file
Tunable Parameters:-None
Output Format:-<tweetid tweet> in txt file

3) Validation and Parameter Tuning:

a)Divided the Training data into **1:4 folds**
b)**4 folds** used for **training** and **1 fold** for **testing**
c)generated the **scores** for **sports feature set** and **politics feature set**.
d)validated the training set against the **best scores**

4) Testing Algorithm:

a)decomposed the test tweet into features.
b)classified the generated feature based on the learned classifier and returned the label

---

Explanation of results on validation data:

Naive Bayes Classifier is based on the probability of occurrence of a word in a feature set, the generated labels for the testing data are based on the probability of occurrence in the given tweet.

Team : AustereSpaiens **Email** : sanchitagarwal108@gmail.com,
rao.vidyadhar@gmail.com

More over the occurrence of a bi-grams of a specific class with a high score in a given tweet increases the probability of its being classified as that class.

The Chi-square test based score evaluation for the bi-grams accounts for the results from the algorithm.

The accuracy of **84.4579** on the validation data is due to **the 3 step procedure** of the algorithm which include:

1) The removal of all the stop words including the SMS slangs used in tweets.
2) Separation of the two types of tweets and learning features for them separately.
3) The computation of bi grams based on chi-square has increased the accuracy.

The algorithm can be further improved if n-grams are used and some important **keywords for sports and politics** are given high weights and are included in during the scoring.
The above **biasing of score** based on keywords can significantly improve the results.
