**Project Proposal**

Sentiment Analysis of Twitter data

**Name: Dean D’souza**

**H.U. ID: 168424**

# The Data:

The data was obtained from the Sentiment140 project, which is a tool for the sentiment analysis of twitter data, and is in the form of a Microsoft Excel Comma Separated file.

It was built by the authors using a variety of technologies including the Twitter API, Amazon EC2, Google Gadgets, etc. The Twitter Search API was used to collect all the tweets in the dataset using keyword search and was automatically created. The authors used emoticons as indicators of positive and negative sentiment, particularly ☺ to indicate a positive tweet and ☹ to indicate negative tweets.

However, the dataset which was free to download consisted of training and test data which had all emoticons stripped out of the tweets.

The data set consists of the following variables:

1. Tweet Polarity (0 = negative, 2 = neutral, 4 = positive)
2. Id of the tweet (2087)
3. Date of the tweet (Sat May 16 23:58:44 UTC 2009)
4. The Query (For example: lyx). If there is no query, then this value is NO\_QUERY.
5. The user that tweeted (For example: robotickilldozr)
6. The text of the tweet (For example: Lyx is cool)

The reason I picked this dataset is so that I can get a view of sentiment analysis on twitter data.

# Questions of Interest:

Some of the questions to be answered include:

1. Which words in the corpus provides a greater weight for determining sentiment?
2. What is the average length of a tweet and does the length of the tweet help in indicating sentiment? (though tweets are limited to 140 characters, it would be interesting to evaluate this question)
3. Is there a relation between certain users and the polarity of the tweet?

# Methods:

A popular tool for sentiment analysis is the NLTK library for python. This toolkit aloo provides a number of ways to create corpora and also allows to easily read comma separated files using the csv.reader() function.

It also provides a number of algorithms for classification such as the Naïve Bayes classifier, Decision Tree classifier and Maximum Entropy classifier. For the purpose of this project, I would like to build two classifiers namely the Naïve Bayes Classifier and the Decision Tree Classifier so as to perform sentiment analysis on the test data and also to compare performance between the two.

Additional techniques would also be used to answer some of the questions if basic statistics could not provide sufficient evidence.

# References:

[1] Sentiment140, (<http://help.sentiment140.com/for-students/>)

[2] “Natural Language Toolkit”, (<http://www.nltk.org/>)

[3] “Twitter Sentiment analysis using python and nltk”, Laurent Luce, (<http://www.laurentluce.com/posts/twitter-sentiment-analysis-using-python-and-nltk/>)