**Assignment 04**

Hands-on Data Understanding and preparation

**Name: Dean D’souza**

**H.U. ID: 168424**

# Abstract

Sentiment analysis is the process of analyzing text data to identify and categorize opinions, especially to determine if the writer’s attitude towards a topic, product, etc. is positive, negative or neutral. The project aims at analyzing the sentiment contained within twitter data (also known as tweets). There are number of steps involved in sentiment analysis, but before this we discuss some important questions associated with the data.

# Data Cleansing

Data cleansing involves the removal of bad data and data that is too detailed for use as well as the correction of some bad data. Data cleansing can be performed in two ways:

1. Validating codes against a list of acceptable results
2. Deleting particularly “Dirty” records

The data set to be used for the project has already been cleansed and is ready for training a classifier model.

# Data Transformation

Data transformation involves conversion of the data into a form more suitable for use with the algorithm such as transformation of numerical values to a common range (such as z-value) or converting categorical data information into numerical by using dummy variables.

While the data is ready for use in a classifier through the values stored in the ‘Text’ variable (consisting of the raw text of the tweet), it also needs to be repurposed for answering the questions associated with why the sentiment value is important.

One such transformation would be for the ‘Query’ variable which needs to be converted to numeric form to correctly measure its correlation with the sentiment value.

# Data Imputation

Data imputation involves techniques for handling missing values in the data, such as, like-wise (or case-wise) deletion, pair-wise deletion and reasonable value imputation.

The data for the project does not contain missing values and so data imputation need not be applied to it.

# Data Weighting and Balancing

Data weighting and balancing is done for a number of algorithms in order to account for the effects that each variable has on the target variable.

The data for the project however will not require any balancing to be done in earlier phase but may require to be done to correctly model for the classifying each user.

# Data Filtering

Data filtering involves techniques used to remove or eliminate rows so as to remove unnecessary information which helps to reduce noise below the level that can confuse the analysis. It usually involves operations such as removal of outliers, Aggregation or selection to set the time-grain of the analysis.

While the data of the project does come with timestamps for each post, the initial analysis will not require the use of such parameters. However, in order to answer later questions, a classifier may be built (mostly through k-means clustering) and may require that a high pass filter be applied to filter out any possible outliers.

# Data Abstraction

Data abstraction involves techniques for transforming data to meet the needs of processing.

The data for the project would not require abstractions to be made for both the initial and later phases.

# Data Reduction (Data Sampling)

Data sampling is performed to reduce the number of cases to be used for modelling so that only cases in which the response patterns are relatively homogeneous are used, and to balance for occurrence of rare events for analysis by machine learning tools. There are many methods of performing Data Reduction or sampling the data.

As the data for the project is quite large (more than 1,048,576 entries), data reduction would be necessary to bring it down to a process able level for laptops. Approximately 20% of the data will be extracted from the corpus (numbers may vary so as to keep the data processable), which will be further divided into training, dev-test and test data sets for effective use.

Additionally, only the ‘Text’ variable will be used to get the sentiment value, so for the purpose of the initial classifier, only one variable will be used. However, for later purposes the ‘text’ variable will be left out and all other variables (including the sentiment value target variable) will be used.

# Data Derivation

Data derivation techniques usually involves obtaining new variables for aiding the analysis.

In the case of the project, the ‘timestamp’ values need to be divided so as to create new variables for year, day, month and day of the week. Additionally, our initial analysis will involve training a classifier to perform sentiment analysis on the text data which will produce the sentiment value. This will then be used for classification of each user.

# Conclusion:

While the data is mostly clean and ready to be used for the initial phases, additional cleaning and transformation may be required in order to successfully answer all the questions the project aims to answer.

# Reference:

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

[2] “Handbook of Statistical Analysis and Data Mining Applications”, John Elder, Gary Miner, Robert Nisbet. 2009