On internet opinion (sentiments on topic) mining is helping users in knowing the quality of any organization or products, if any user has good experience on any product or company then he will express good reviews/opinion and by seeing this opinion others users can know the quality of the product, in today’s online social networks like twitter all peoples expressing their opinions and social networking sites developing new techniques to detect sentiments from this opinions, all existing techniques used to discover either Positive or Negative or Neutral sentiments from topics but this paper proposes 5 levels of sentiments detection such as High Positive, Moderate Positive, Neutral, High Negative and Moderate Negative. To detect sentiments author is using 4 Ordinal Regression machine learning algorithms such as Softmax, Decision Tree, Random Forest and Support Vector Regression.

Ordinal Regression means classifier used many independent variables to predict class of given data, In this paper also we give tweets as input and classifier predict sentiment by using all independent words from this tweets. Ordinal regression is a statistical technique that is used to predict behaviour of ordinal level dependent variables with a set of independent variables. The dependent variable is the order response category variable and the independent variable may be categorical or continuous.

From above algorithms Decision Tree is giving better prediction result and to train all algorithms we are using publicly available twitter dataset from NLTK library. We are using many features from NLTK (Natural Language Processing Tool Kit) library such as cleaning tweet text by removing special symbols, removing stop words (such as the, then, where etc.), word stemming which means removing ing, tion etc from words. After cleaning tweets then we will convert all tweets to BOG (Bag of Words Dictionary) and then convert BOG to vector by calculating TF/IDF (Term Frequency/Inverse Document Frequency).

TF/IDF = number of times word occur in tweet / total number of times word occur in all tweets.

For example

Tweet1 = An apple a day keep the doctor away.

Tweet2 = Apple is good for health.

From above 2 tweets if we want to calculate TFIDF then we need to find all unique words from 2 tweets and count them. Remove all stop words and in below example I am keeping all unique words in column and their count in rows

Apple day keep doctor away good health (unique words from 2 tweets)

T1 1 1 1 1 1 0 0 (count of all words from tweet1)

T2 2 0 0 0 0 1 1

In above table we can see if word appear in tweet then I put value 1, if word not appear in tweet then I put value 0. Now from above table we can calculate TFIDF

TFIDF of word Apple in tweet1 = ½ = 0.5

Apple word appear in tweet1 is 1 time and Apple appear in all tweets (means tweet1 and tweet2) is 2 times, so ½ will give 0.5 TFIDF for Apple Word. Similarly application calculate TFID for all words and form a vector. This vector will be trained with SVR, Random Forest and Decision Tree algorithms with all positive and negative tweets dataset. Whenever we give new tweets then application convert new tweets also in TFIDF vector and then apply new tweet vector on train vector to predict sentiment type. Once we got sentiment value then we can calculate High positive or moderate positive or negative by calculating polarity. Polarity value will come from TFIDF, if new tweet polarity percentage is closer to positive (80 to 100%) then tweet is HIGH Positive, if 50 to 80 then moderate positive else neutral. Similarly negative ratio also calculated.

To run this project we need to download NLTK library, to download it just double click on ‘download.bat’ file then one window will appear and you just click on ‘download’ button and wait for 10 minutes to allow the library to download.