# Sentiment Analysis for Stock Market Prediction Proposal

**Project Overview**

The aim of the project is to create a 2- layered machine learning system in which the first layer performs the retrieval and extraction of sentiment from news articles. The sentiment is then pooled with other relevant attributes as features into a classifier to result in a predicted stock market price.

After the completion of the 2 layered classifier, a web interface will be made. With this web interface, users can simply view the sentiment classification of news articles and timelines of the classification and predictions.

**Project Background**

It is well understood that stock market price is determined by the value that the market places on the stock and in turn, this is determined by what the perceived value of the company’s assets are. Hence, predicting the stock market can be broken down further into “predicting the perceived value of assets.” In the age of blogs and online news websites that publish news articles as soon as they are known, determine the perceived value becomes a lot easier. The main outcome of the system will be to determine whether or not the performance of SVM based stock market prediction can be aided by the inclusion of sentiment as one of the features.

**Objectives**

1. Build a sentiment analyser that will take as input news from online news blogs and will produce as an output the sentiment polarity and strength.
2. Build a classifier which will take in as input the output from the sentiment analyser as well as other suitable features in order to make a prediction of the stock market value
3. Build a web system which will act as an interface to the classifier and sentiment analyser and will provide suitable visualisations in order that the effect of the market sentiment can easily be measured against the performance of the stock market
4. Determine the performance of the classifier against that of traditional classifiers that do not take into account the sentiment.
5. Determine the performance of classifier against that of newer stock market price classifiers that do not take into factor other features such as momentum, relative strength index, commodity channel index and other features as identified.

**Programming Languages and Libraries**

Python will be used to program all parts of the system including the nltk library and libsvm. This is because the Python programming language make the manipulation of raw data very easy. In addition, web frameworks such as Django, Flask make the transmission from Python objects into web page graphs very easy. For web development, technologies such as HTML, CSS, D3 along will be used.

**Project Workflow**

Develop Client

Develop Server

Literature Search

Analyse Labelled Data

Literature Review

Collect, Generate Data, preprocess data

Label data

Train SVM on Data

Pre-process data: (Convert to unigram, remove stopwords, perform feature selection).

Evaluate results.

Write Complete Report

Credit Scoring Application

Collect features for prediction

Pre-process Data

Train SVM

Evaluate results.

**Time Estimates**

OCT

JAN

NOV

FEB

MARCH

DEC

APRIL

**3 Dec 2014 – Presentation**

* Aim to show developed SVM, regardless of results
* Aim to show labelling results
* Aim to show results of literature review.

**28 Feb 2015 – Dissertation Outline**

* Completed Analysis on Prediction
* Written dissertation sample for preliminary review

**08 May 2015 – Dissertation Complete**

Submit Dissertation