Impact of News and Social Media on Financial Markets

Final year Project Report

By
Anshul Goyal (147106)
Kiran Konduru (147126)
Ibrahim Shaik (147148)

Guided by
Dr. K. V. Kadambari
Dept. of Computer Science and Engineering
NIT Warangal

To
Dr. B. B. Amberkar
Project Coordinator B.Tech 4/4 Section-A
Dept. of Computer Science and Engineering
NIT Warangal

Contents

Problem Statement

The objective of this study is to develop a market sentiment model based on news and social media data for financial markets using machine learning and see its impact on various financial market indicators like market indices, trading volumes, market volatility etc.

Abstract

Financial market analysis on the basis of financial news and social media sites like Twitter etc. has drawn a lot of attention recently. Due to the volatility of the financial market, price fluctuations based on news reports and social media sentiment are common. Traders draw upon a wide variety of publicly-available information to inform their market decisions.

Sentiment analysis can use natural language processing, machine learning, text analysis and computational linguistics to identify the attitude of a writer with respect to a topic. It's an important cornerstone of behavioral finance, where theorists believe that markets are irrational and that asset prices are driven by human emotion (e.g., fear, greed, hope and overconfidence, among others).

The efficient market hypothesis (EMH) asserts that financial market valuations incorporate all existing, new, and even hidden information, since investors act as rational agents who seek to maximize profits. Behavioral finance has challenged this notion by emphasizing the important role of behavioral and emotional factors, including social mood, in financial decision-making. As a consequence, measuring investor and social mood has become a key research issue in financial prediction and can be done efficiently using machine learning.

The proposed work is currently focused on finding association between financial indicators and sentiment based on social media, financial news and general world news.

Efficient Market Hypothesis – An Introduction

The **efficient market hypothesis** (EMH) asserts that financial markets are "informational efficient", or that prices on traded assets (e.g., stocks, bonds, or property) already reflect all known information, and instantly change to reflect new information. Information or news in the EMH is defined as anything that may affect prices that is unknowable in the present and thus appears randomly in the future. Stock market prediction brings with it the challenge of proving whether the financial market is predictable or not, since there has been no consensus on the validity of Efficient Market Hypothesis (EMH).

Stock market prediction has been an important issue in the field of finance, engineering and mathematics due to its potential financial gain. As a vast amount of capital is traded through the stock market, the stock-market is seen as a peak investment outlet. Researchers have strived for proving the predictability of the financial market. Henceforth, Stock Market prediction has always had a certain appeal for researchers. While numerous scientific attempts have been made, no method has been discovered to accurately predict stock price movement. Even with a lack of consistent prediction methods, there have been some mild successes.

Types of Data

We are planning to use three types of data for our analysis.

1. **General World News** – Using the general world news to predict the impact on financial parameters. For example – How would a terrorist in some ‘X’ country affect the DJIA index etc.? We will be analyzing this news in two forms -
    a. General World News Headlines
    b. General World News Articles

2. **Financial News** – Using news specifically from the financial domain to predict the impact on financial parameters. This should intuitively present better results than the general world news. We will be analyzing this news in two forms -
    a. Financial News Headlines
    b. Financial News Articles

3. **Social Media Data** – Using data from popular microblogging sites like Twitter which are being extensively used in real time sentiment tracking and public mood modeling.

Data Collection – Datasets

1. **Reddit News Dataset:** This dataset consists of top 25 daily general world news headlines collected for the period 2008-2016. The news source is Reddit News. No news articles are available.

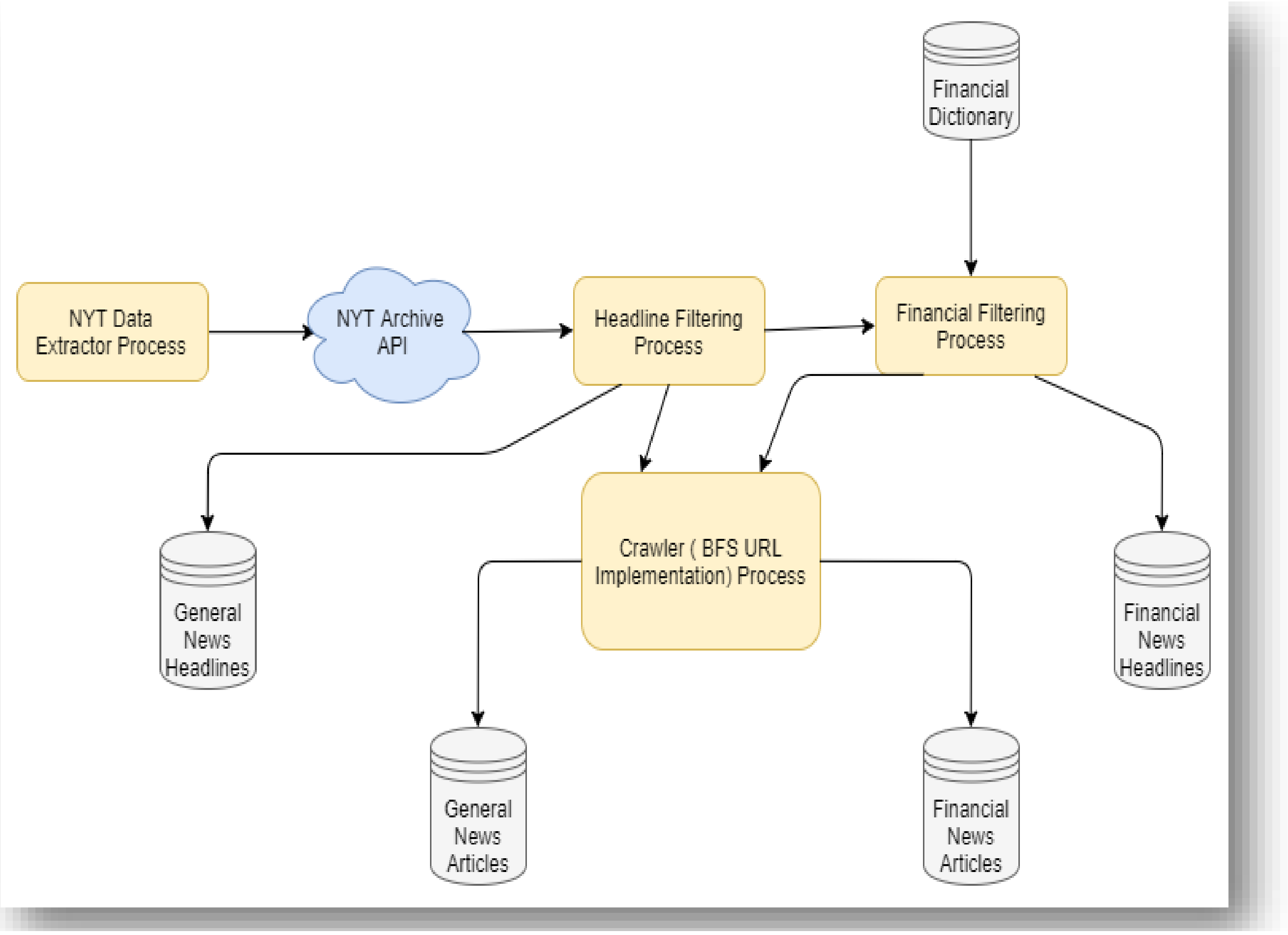
2. **New York Times Dataset:** Created our own data collection model in order to use the NYT API to gather news headlines and articles for the period 2008-2016. A financial dictionary was constructed to filter the financial news and a web crawler was written to get the news articles. Approximately, 3800 world news articles per month and 1000 financial news articles are being collected.

3. **Twitter Dataset:** No open source dataset required for our application was found. Dataset is being created on a weekly basis using the Twitter API.

4. **DJIA Dataset:** Dow Jones Industrial Average (DJIA) index data was collected from Yahoo Finance for the period 2008-2016.

Data Collection – Data Collection Model

The following diagram shows the data collection model constructed.



Data Collection – Pseudo-code

```
1.0 BEGIN
2.0 Initialize Datasets:
        2.1 General News Headlines (GNH) = null
        2.2 General News Articles (GNA) = null
        2.3 Financial News Headlines (FNH) = null
        2.4 Financial News Articles (FNA) = null
3.0 Hit the New York Times Archive API to gather news for the specified period.
4.0 Collect all the headlines along with their ids: This will be the GNH dataset.
5.0 For all news ids :
        5.1 Call the Crawler (BFS URL Implementation) Process to collect the entire
            news article
        5.2 Add article to GNA dataset.
        5.3 End For loop
6.0 GNA dataset is generated.
7.0 For all news ids :
        7.1 For all words in the financial dictionary:
                7.1.1   If word is present in the headline:
                        7.1.1.1         Add to FNH dataset
                        7.1.1.2         Call the Crawler Process to collect the entire news
                                        article
                        7.1.1.3         Add the article to FNA dataset
                        7.1.1.4         Continue
8.0 FNH and FNA datasets are generated
9.0 END
```

Financial Parameter – DJIA

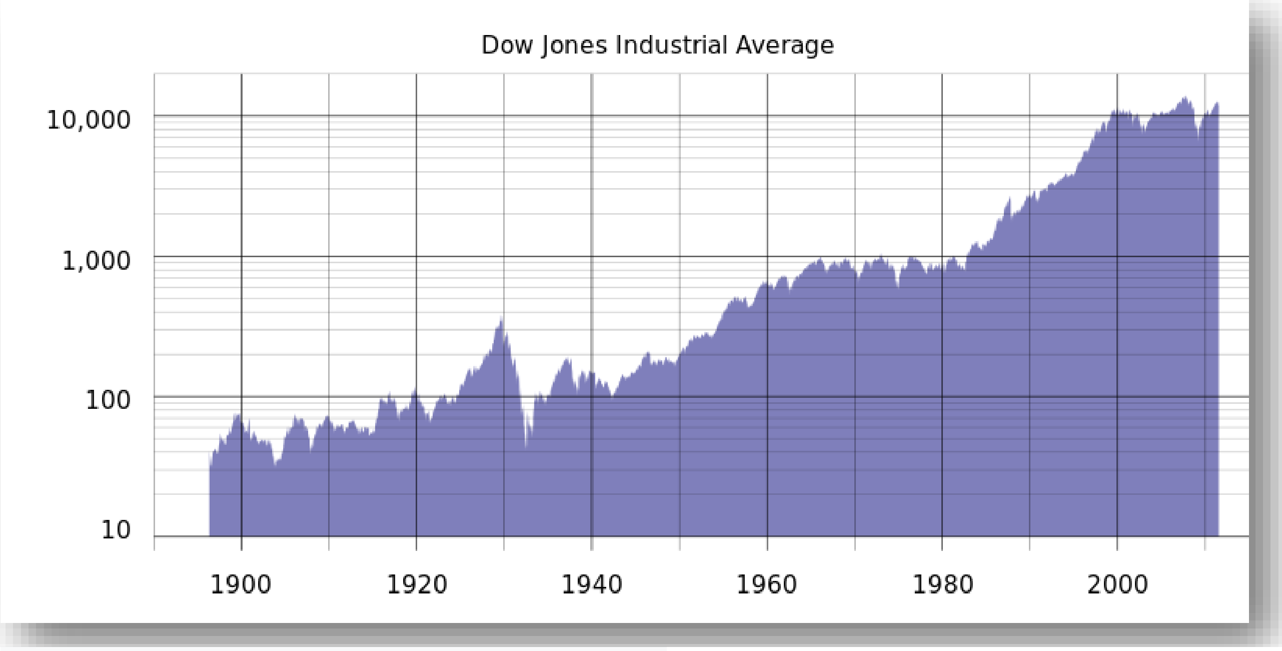
Currently we are working with DJIA as our financial parameter which would also be extended for other financial parameters in the near future.

DJIA – Dow Jones Industrial Average

The Dow Jones Industrial Average (DJIA) is a stock market index, and one of several indices created by Wall Street Journal editor and Dow Jones & Company co-founder Charles Dow. The industrial average was first calculated on May 26, 1896.

It is an index that shows how 30 large publicly owned companies based in the United States have traded during a standard trading session in the stock market.

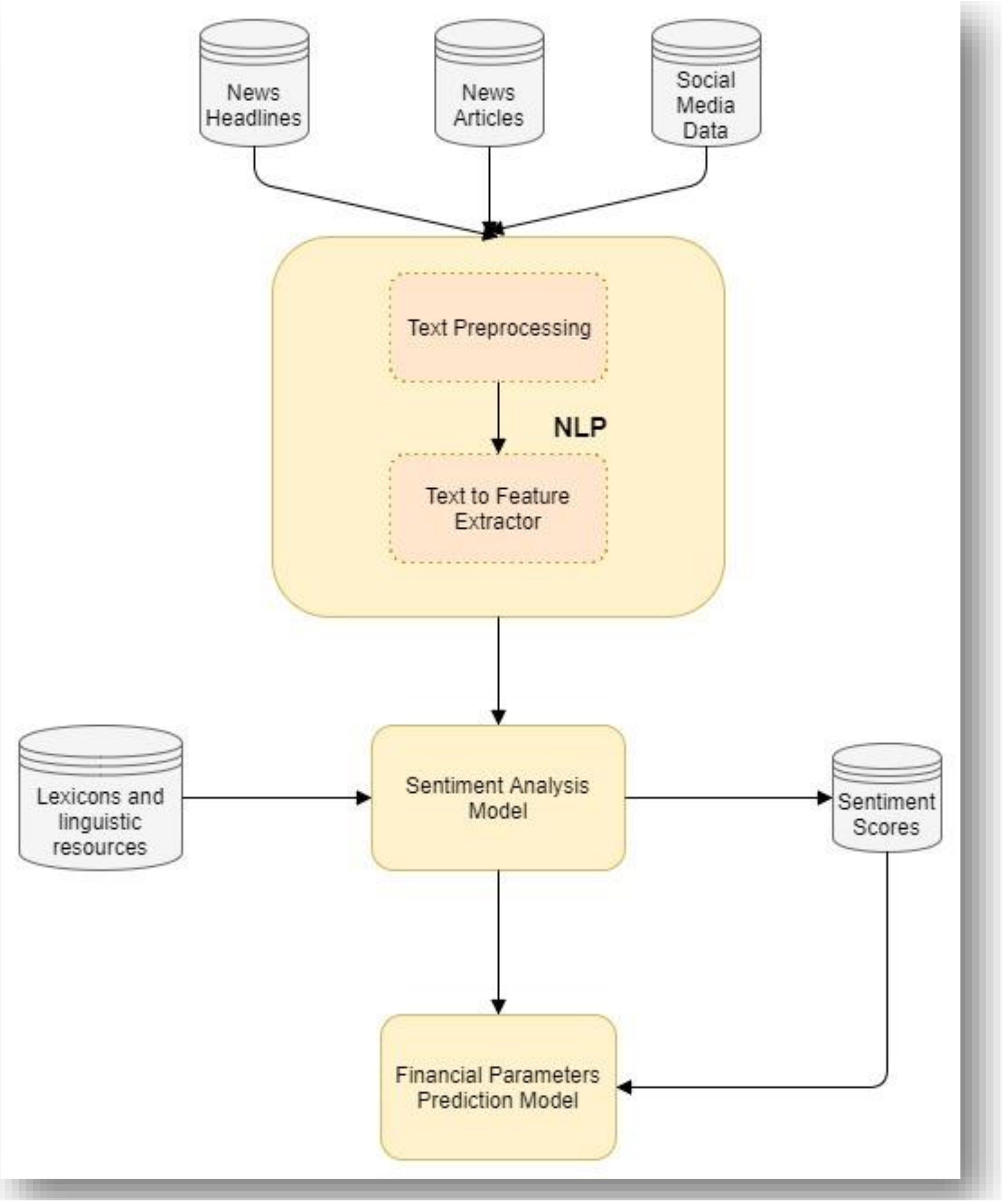
The value of the Dow is not the actual average of the prices of its component stocks, but rather the sum of the component prices divided by a divisor, which changes whenever one of the component stocks has a stock split or stock dividend, so as to generate a consistent value for the index.



Historical logarithmic graph of the DJIA from 1896 to July 2011

High Level Flow Diagram

The following diagram depicts the high level formal model.



Text Preprocessing

The following text preprocessing techniques are being tested for our application.

1. **Noise Removal -** Any piece of text which is not relevant to the context of the data and the end-output can be specified as the noise. A general approach for noise removal is to prepare a dictionary of noisy entities, and iterate the text object by tokens (or by words), eliminating those tokens which are present in the noise dictionary. Ex: Stop words filtering etc.

2. **Lexicon Normalization -** Another type of textual noise is about the multiple representations exhibited by single word. Normalization is a pivotal step which converts the high dimensional features (N different features) to the low dimensional space (1 feature).

    1. **Stemming -** It is a rudimentary rule-based process of stripping the suffixes (“ing”, “ly”, “es”, “s” etc.) from a word.

    2. **Lemmatization -** It is an organized & step by step procedure of obtaining the root form of the word, it makes use of vocabulary (dictionary importance of words) and morphological analysis (word structure and grammar relations).

3. **Object Standardization** – Text data often contains words or phrases which are not present in any standard lexical dictionaries. Example – Removing colloquial slangs from tweets etc.

Text to Features

The following feature engineering techniques for text data are being tested for our application.

1. **Named Entity Recognition (NER) -** The process of detecting the named entities such as person names, location names, company names etc. from the text is called as NER. A typical NER model consists of three blocks:
    1. Noun phrase identification
    2. Phrase classification
    3. Entity disambiguation

2. **Topic Modeling -** Topic modeling is a process of automatically identifying the topics present in a text corpus, it derives the hidden patterns among the words in the corpus in an unsupervised manner. Topics are defined as "a repeating pattern of co-occurring terms in a corpus".

3. **N-grams as features** – A combination of N words together are called N-Grams. N grams (N > 1) are generally more informative as compared to words (Unigrams) as features.

4. **Term Frequency – Inverse Document Frequency (TF – IDF) -** TF-IDF is a weighted model commonly used for information retrieval problems. Formula for TF-IDF calculation is :

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Sentiment Analysis

Sentiment analysis refers to a wide range of areas of natural language processing, text mining and computational linguistics. Opinion mining, a sub discipline within data mining and computational linguistics, refers to the computational techniques for extracting, classifying, understanding, and assessing the opinions expressed in various online news sources, social media comments, and other user-generated content. Sentiment analysis is often used in opinion mining to identify sentiment, affect, subjectivity, and other emotional states in online texts. The aim of Sentiment Analysis is to develop a machine learning technique for determining the polarity of a document. The key objective here is to design an algorithm that can learn ‘certain’ information from the pre-classified data set (learning/training data set) and then classify a document into its predicted class.

The sentiment found within news articles and social media data provide useful indicators for many different purposes. These sentiments can be categorized either into two categories: positive and negative; or into an n-point scale, e.g., very good, good, satisfactory, bad, very bad. In this respect, a sentiment analysis task can be interpreted as a classification task where each category represents a sentiment. Sentiment analysis of news articles and social media data provides a means to estimate the movement of a particular index in either direction.

Some of the basic steps involved are:

- Generating a Sentiment Dictionary/Lexicon:

  A new sentiment dictionary would be generated specifically for financial domain sentiment analysis. The initial seed lists would be recursively expanded to complete the dictionary. In this research, the lexicon will be created based on the norms of business and financial domain. The process of constantly adding new words to the lexicon will be needed in order to produce larger domain specific lexicon. A list of new words from financial related articles will be generated and each of the words found will be tagged as either positive or negative polarity.

  In this work, the values positive, negative and neutral will be assigned to general terms, which express some kind of sentiment (e.g. 'benefit', 'positive', 'danger') and to financial terms (e.g. 'risk capital', 'rising stock', 'bankruptcy').

  Example:

  Words like bear and bull have different meanings in finance than their usual meanings. These are just names of animals in general domain whereas in financial domain, bear conveys a negative sentiment and bull conveys a positive sentiment.

- Classification:

  There are various classifiers that can be used for sentiment analysis. Some of the classification methods for sentiment analysis are:

  1. Rule based Classification :

  A rule consists of an antecedent and its associated consequent that have an 'if-then 'relation:

  antecedent =⇒ consequent

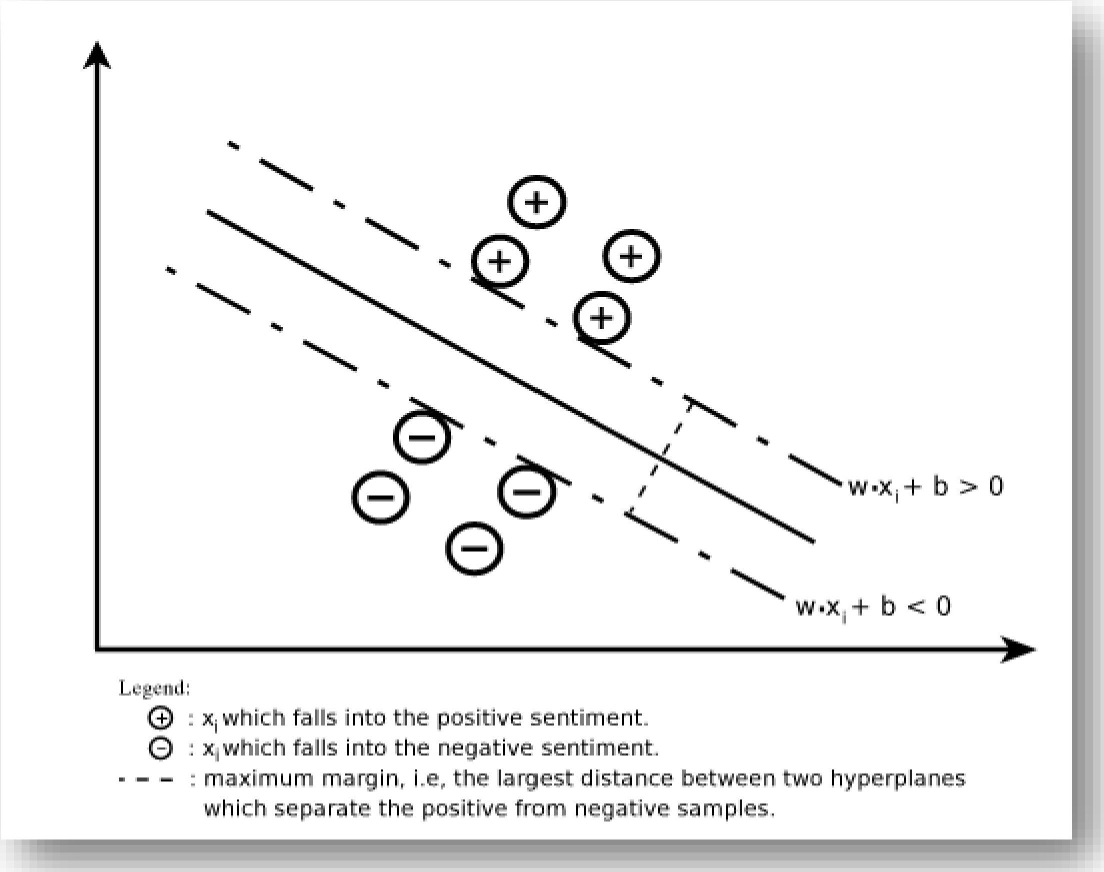
  An antecedent defines a condition and consists of either a token or a sequence of tokens concatenated by the ∧ operator. A token can be either a word, '?' representing a proper noun, or '#' representing a target term. A target term is a term that represents the context in which a set of documents occurs. A consequent represents a sentiment that is either positive or negative, and is the result of meeting the condition defined by the antecedent.

$$\{\text{token1} \wedge \text{token2} \wedge \ldots \wedge \text{token n}\} \Longrightarrow \{+|-\}$$

For Example: Bull => {positive sentiment i.e. +}

Bear => {negative sentiment i.e. -}

2. Support Vector Machines:


w•x_i + b > 0
w•x_i + b < 0
Legend:
⊕ : x_i which falls into the positive sentiment.
⊖ : x_i which falls into the negative sentiment.
- - - : maximum margin, i.e, the largest distance between two hyperplanes which separate the positive from negative samples.

Given a category set, C = {+1, −1} and two pre-classified training sets, i.e., a positive sample set, $T_r^+ = \sum_{i=1}^{n}(d_i, +1)$ and a negative sample set,

$T_r^- = \sum_{i=1}^{n}(d_i, -1)$ the SVM finds a hyperplane that separates the two sets with maximum margin (or the largest possible distance from both sets), as illustrated in the figure.

3. Hybrid Classification:

Hybrid classification means applying classifiers in sequence. For example:

RBC -> SVM (i.e. Rule Based Classification followed by the SVM Classifier)

4. Manual Classification:

Manual classification involves reading each article, headline and tweet and assigning it a sentiment tag: positive, neutral, or negative. Some general guidelines will be used when manually classifying articles. Mergers will be generally considered positive because they indicate companies have cash on hand. Technology and general interest articles will be considered neutral as they are not directly related to stocks. Lawsuits will be generally considered negative, as was corruption. Rising interest rates will be considered negative and declining interest rates will be considered positive because they indicate more cash in the general economy.

Example:

1.) News Headline: Bank of America Profit Misses Expectations

Classification: Negative

2.) News Headline: British Activists Press Tax Case Involving Goldman Sachs

Classification: Negative (especially for DJIA as Goldman Sachs is one of the 30 companies represented in the DJIA index)

3.) News Headline: Coca-Cola Posts Strong Profit on Emerging-Market Sales

Classification: Positive (especially for DJIA as Coca-Cola is one of the 30 companies represented in the DJIA index)

- Sentiment Scoring:

  Sentiment scores can be evaluated for each sentence (for sentence-based sentiment analysis), for entire document (for document-based sentiment analysis), or for specific aspects of entities (for aspect-based sentiment analysis). Sentiment scores are used to annotate the document and these annotations are the output of the system. As mentioned, each word in the lexicon is tagged for its polarity of either positive or negative. If the token matches with a word in the lexicon, then the token will be tagged as positive or negative polarity according to the pre-tagged words in the lexicon. Tokens that do not match with the words in the lexicon will be ignored in the system.

  For document classification to prepare training data set, we will apply some algorithm to aggregate the sentiment of words (L). So the sentiment of overall document (global sentiment) can be calculated by applying some aggregation function as follows:

$$\mathbf{G = F\ (L_i)}$$

  This global sentiment will define sentiment of overall news article.

Basic Implementation

Current implementation is done on the Reddit general world news headlines dataset without using the sentiment analysis model using different classifiers to predict the direction of DJIA index.

1. Without Text-Preprocessing

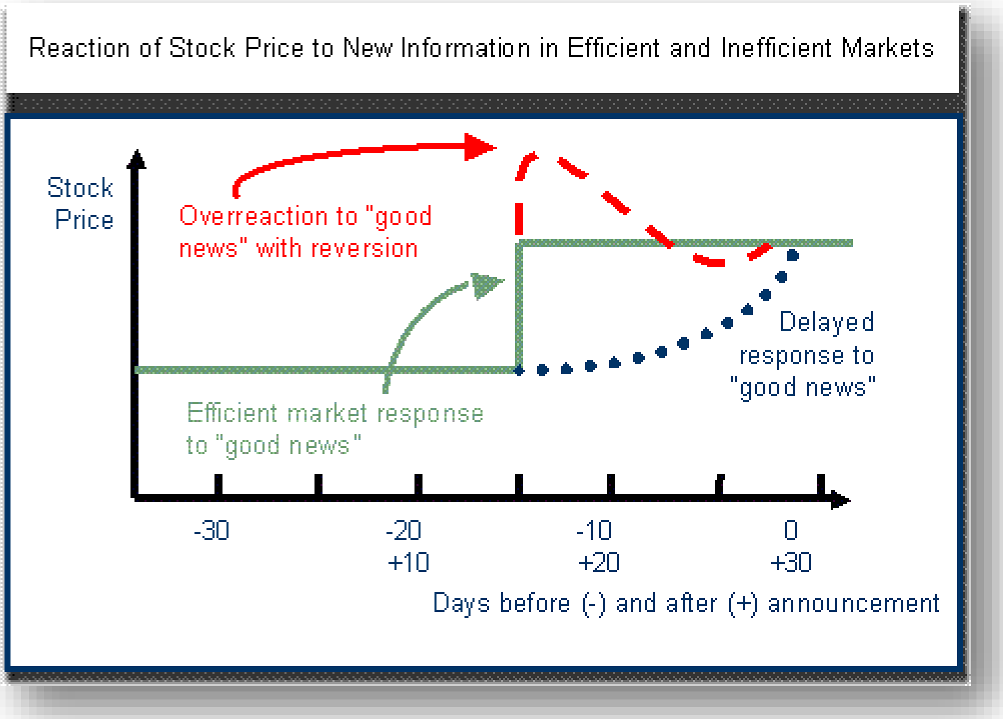
Classifier Used	**Accuracy**
Logistic Classifier	51.58 %
SVM	42.32 %
Random Forest	49.73 %

2. With Text-Preprocessing

Classifier Used	**Accuracy**
Logistic Classifier	56.35 %
SVM	55.82 %
Random Forest	56.61 %

Future Plan of Action

- Designing and Implementing a Sentiment Analysis Model specifically for use in financial domain.
- Collecting Social Media Data using Twitter API
- Incorporating other types of data for analysis : Financial News and Social Media Data
- Using other financial market parameters apart from DJIA for prediction like commodities, company stocks etc.
- Tackling the effect of news sentiment on (T+1) day.


Reaction of Stock Price to New Information in Efficient and Inefficient Markets
Stock Price
Overreaction to "good news" with reversion
Delayed response to "good news"
Efficient market response to "good news"
-30
-20
+10
-10
+20
0
+30
Days before (-) and after (+) announcement

References

- Desheng Dash Wu, David L. Olson - Financial Risk Forecast Using Machine Learning and Sentiment Analysis
- Sunandan Chakraborty, Ashwin Venkataraman, Srikanth Jagabathula and Lakshminarayanan Subramanian - Predicting Socio-Economic Indicators using News Events
- Chuan-Ju Wang, Ming-Feng Tsai, Tse Liu , Chin-Ting Chang - Financial Sentiment Analysis for Risk Prediction
- Jinjian Zhai, Nicholas Cohen, Anand Atreya - Sentiment analysis of news articles for financial signal prediction
- Huina Mao, Scott Counts, Johan Bollen - Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data