**Stock Market Prediction Using Data Mining**

**Abstract –**

Data mining is well founded on the theory that the historic data holds the essential memory for predicting the future direction. This technology is designed to help investors discover hidden patterns from the historic data that have probable predictive capability in their investment decisions. The prediction of stock markets is regarded as a challenging task of financial time series prediction. Data analysis is one way of predicting if future stocks prices will increase or decrease. Also, it investigated various global events and their issues predicting on stock markets. The stock market can be viewed as a particular data mining problem. Text mining approach is also used for measuring the effect of real time news on stock. It uses different techniques and strategies to predict ups and downs in stock market. In this paper, we present a model that predicts the changes of stock trend by analyzing the influence of non- quantifiable information namely the news articles which are rich in information and superior to numeric data.

**Existing system:**

The rapid progress in digital data acquisition has led to the fast-growing amount of data stored in databases, data warehouses, or other kinds of data repositories. Although valuable information may be hiding behind the data, the overwhelming data volume makes it difficult for human beings to extract them without powerful tools. Easy and quick availability to news information was not possible until the beginning of the last decade. In this age of information, news is now easily accessible, as content providers and content locators such as online news services have sprouted on the World Wide Web. Continuous availability of more news articles in digital form, the latest developments in Natural Language Processing (NLP) and the availability of faster computers lead to the question how to extract more information out of news articles. Financial analysts who invest in stock markets usually are not aware of the stock market behavior. They are facing the problem of stock trading as they do not know which stocks to buy and which to sell in order to gain more profits. All these users know that the progress of the stock market depends a lot on relevant news and they have to deal daily with vast amount of information. They have to analyze all the news that appears on newspapers, magazines and other textual resources. But analysis of such amount of financial news and articles in order to extract useful knowledge exceeds human capabilities. Text mining techniques can help them automatically extracting the useful knowledge out of textual resources. We would develop a system which is able to use text mining techniques to model the reaction of the stock market to news articles and predict their reactions. By doing so, the investors are able to foresee the future behavior of their stocks when relevant news are released and act immediately upon them. As input we use real-time news articles and intra-day stock prices of some companies in Bombay Stock Exchange.

The overall purpose of study can be summarized in the following research questions:

• How to predict the reaction of stock price trend using textual financial news?

• How data and text mining techniques help to generate this predictive model? In order to investigate the impact of news on a stock trend movement, we have to make a prediction model.

**PROPOSED WORK**

Methodology for NLP module To exactly predict the stock price is very complex task till the date. Here we are proposing to make a prediction based on news articles using one of the Text Mining concepts like sentiment analysis. We would like to make the prediction system for Indian Stock market. Implementation steps to be followed to make a prediction system are:

1. Gathering of news articles.

2. Perform sentiment analysis on news articles

3. Get Polarity of the text

4. Make a prediction based on current stock price and calculated polarity of the text.

To Get the News Articles To collect the news articles R.S.S feed is the main source. As R.S.S feed is used for news article collection process. Here the Times of India’s R.S.S feed is used for business and market related news. It will give results by retrieving top news of Indian stock market. We have to just specify the R.S.S feed address in our code.

To Perform Sentiment Analysis and Get Polarity of the text Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level - whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral.

For sentiment analysis and calculating polarity of text two things are used: 1. POS tagger 2. SentiWordNet\_3.0.0 POS Tagger A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally computational applications use more fine grained POS tags like 'noun-plural'. This software is a Java implementation of the log-linear part of-speech taggers.

Part-of-speech tagging (POS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context—i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc. POS-tagging algorithms fall into two distinctive groups: rule-based and stochastic. E. Brill's tagger, one of the first and widely used English POS-taggers, employs rule-based algorithms.

Part-of-speech tagging is harder than just having a list of words and their parts of speech, because some words can represent more than one part of speech at different times, and because some parts of speech are complex or unspoken. This is not rare in natural languages (as opposed to many artificial languages), a large percentage of word forms are ambiguous.

In part-of-speech tagging by computer, it is typical to distinguish from 50 to 150 separate parts of speech for English. For example, NN for singular common nouns, NNS for plural common nouns, NP for singular proper nouns. Several downloads are available. The basic download contains two trained tagger models for English. The POS tagger which I have used is developed by Stanford University natural language processing group [8]; it is licensed under the GNU general public license as it is an open source.

**SYSTEM CONFIGURATION:**

**HARDWARE REQUIREMENTS:**

Hardware – Pentium

Speed - 1.1 GHz

RAM - 1GB

Hard Disk - 20 GB

Floppy Drive - 1.44 MB

Key Board - Standard Windows Keyboard

Mouse - Two or Three Button Mouse

Monitor - SVGA

**SOFTWARE REQUIREMENTS:**

Operating System : Windows

Technology : Java and J2EE

Web Technologies : Html, JavaScript, CSS

IDE : My Eclipse

Web Server : Tomcat

Tool kit : Android Phone

Database : My SQL

Java Version : J2SDK1.5