TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

PULCHOWK CAMPUS

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

The undersigned certify that they have read, and recommended to the Institute of Engineering for acceptance, a project report entitled INTEREST RATE PREDICTION OF BANKS submitted by Bidhya Nandan Sharma, Everest K.C., Sushant Kafle and Swapnil,Sneham in partial fulfilment of the requirements of Major Project for the Bachelor's degree in Computer Engineering.

_____

Supervisor, Dr. Arun Timalsina

Professor

Department of Electronics and Computer Engineering

Date of Approval: August 24, 2014

COPYRIGHT

ACKNOWLEDGEMENT

We owe our sincere gratitude to Dr. Arun Timalsina and Dr. Aman Shakya of Department of Electronics and Computer Engineering, Pulchowk Campus for their kind effort and help in guiding continuously during the project. Kind help and insightful suggestions of Dr. Arun Timalsina during the project supervision worth a millions of thanks.

Our special thanks goes to Er. Prabindra Kharbuja and Er. Moti Ram Ghimire of Yomari Incorporated Private Limited for providing us with constant co-operation and helpful guidelines in the project.

Our final thanks goes to all our friends, seniors, teachers and all those people who have helped us directly or indirectly in completing this project.

With warm regards,

-Bidya Nandan Sharma (067BCT507)

-Everest K.C. (067BCT514)

-Sushant Kafle (067BCT545)

-Swapnil Sneham (067BCT547)

ABSTRACT

Data mining has drawn much attention in generating useful information from the Web data. It is a powerful technology that carries a great potential to discover information within the data that queries and reports cannot reveal effectively.

This data mining techniques has been used in this project to predict the Interest Rate of Banks. This project provides a portal for the prediction of interest rate of one month ahead which can be useful for large mass of people for making wise decision. It mainly concentrates on the use of the news information to build the knowledge base using cognitive map providing the relative strength of news information to the neural network to build the model for the prediction of the interest rate. It also intends to analyze various factors that plays important role in the prediction of the interest rate so that it can be used for useful decision making. Moreover its web plus mobile version provides users with the easy to use interface and easy to interpret results. The major outcome is the predicted numerical value of interest rate and the various visualization of trend of fluctuation of interest rate with respect to the contributing factors.

Correlation was used to calculate the relationship between various data parameters; regression analysis to predict the data model of interest rate prediction which best fits the given sets of data.

Keywords:

Information retrieval, Sentiment Analysis, Cognitive Map, News Mining, Regression Analysis , Correlation Analysis

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ADT	Android Development Tools
AI	Artificial Intelligence
AIDE	Android Integrated Development Environment
API	Application Programming Interface
CM	Cognitive Maps
CSS	Cascading Style Sheets
CSV	Comma Separated Values
D3	Data-Driven Document
DB	Database
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DOM	Document Object Model
DSF	Django Software Foundation
EMI	Equated Monthly Installment
ERD	Entity Relationship Diagram
FOREX	FOReign EXchange
GDP	Gross Domestic Product
GNI	Gross National Income
GNU	GNU is Not Unix
HTML	Hyper Text Markup Language
HTTP	Hyper Text Transfer Protocol
ID3	Iterative Dichotomiser 3
IDE	Integrated Development Environment
IR	Information Retrieval
JSON	JavaScript Object Notation
KB	Knowledge Base
KBNMiner	Knowledge Based Miner
KR	Knowledge Representation and Reasoning
LAMP	Linux, Apache, MySQL, Perl/PHP/Python
MAE	Mean Absolute Error
MM	Moving Mean
MSE	Mean Square Error

NLTK	Natural Language ToolKit
PHP	Hypertext Preprocessor
POS	Parts of Speech
QEMU	Quick EMUlator
RDBMS	Relational Database Management System
Regexp	Regular Expression
RMSE	Root Mean Square Error
SCM	Source Code Management
SDK	Software Development Kit
SVG	Scalable Vector Graphics
SVM	Support Vector Machine
TDIDT	Top-Down Induction of Decision Tree
TSV	Tab Separated Values
UI/UX	User Interface/ User eXperience
URL	Uniform Resource Locator
US	United States

Chapter 1

1. INTRODUCTION

1.1 Background

Forecasting the term structure of interest rates plays a crucial role in portfolio management, household finance decisions, business investment planning, and policy formulation. Knowing what will be the interest rate of loan in future will help the borrower as well as the lender in efficient decision making like when to borrow loan, when to lend, whether to borrow in floating interest rate or fixed interest rate and so on.

With the advancement in technology we now have the ability to generate and collect enormous amount of data. There are many databases that are used for storing business data. We can readily use them for data mining application but the data such as news information are not stored in databases. The news articles are scattered in various websites of the World Wide Web. Those news are to be collected and classified for us to be able to use them in data mining application. Many researchers have attempted to predict interest rates by using the time series model, neural networks model , the integrated model of neural networks and case-based reasoning . Meanwhile another approach was attempted in the prediction of the stock price index where Kohara, Ishikawa, Fukuhara, and Nakamura took into account non-numerical factors such as political and international events from newspaper information.[1]

This project aims to predict the interest rate based on past interest rates, numerical factors like Petrol Price, FOREX rates, Gold Price etc. and other several non-numerical factors like political scenario of the country, change in government policy, unemployment rate and so on.

1.2 Overview

This project is a web plus mobile application that provides the prediction of interest rate of one month ahead. It utilizes the news information from the e-kantipur.com news in order to subsume the non-numerical factor affecting the interest rate. To

define the polarity of the keywords in the news information we have used sentiment analysis algorithm that is Opinion phrase extraction which is considered as one of the best methods to extract the polarity of the opinion words. For numerical factors, different factors are taken into account like FOREX, petrol price, gold price, GDP (Gross Domestic Product) etc. Our application analyzes these data and presents the result of the predicted interest rate in the numerical value as well as through several insightful visualizations that can be used for wise decision making. We have applied a variety of data mining techniques in order to fulfill this goal. Several Machine learning techniques has been used to build the most appropriate neural network model that is trained with financial and non-financial information to predict the value future interest rate.

1.3 Problem Statement

With ever increasing demand and dependence of people with banks for their savings and loans, the interest rates of bank has been a continuous value of interest. Controlled by various factors such as unemployment rate, government policy for IR, unsuitability of political situation, the interest rate provided by the bank swings with time. It is almost impossible for human mind to account all those factors and predict the upcoming interest rates with sufficient accuracy. Many of household finance decisions, business investment planning, and policy formulation hugely depend on banks for loan and interest. So, it is very crucial for any organization or households to understand how the banks, trusted by their customers is performing now and will perform in near future. This sort of information can be of value to various investors as well.

1.4 Aims and Objectives

The aim of this project is to develop a useful tool that could solve the above mentioned problem by utilizing data mining techniques, which involves extraction of valuable information and training a decision making engine using those information.

The objective of the project is to mine financial data publicly available on web and other related available information which is possible factors affecting banking sector

so as to predict the interest rates. This involves extraction of related data, processing it and extraction of information which will be used to predict the interest rates. The objectives of the project are described as below:

i. Data Analysis: This aims to analyze all the available data for the extraction of financial information.

ii. Interest Rate Prediction: This is the main goal of this project which aims to predict the interest rate based on financial data and sentiment strength of news.

iii. Decision Making: The project also tends to support decision making financial scenarios for individual requiring financial decisions.

Data analysis is one of the major portion of this project. The term financial information, here, is the indicative which governs the changes of interest rates. The analysis largely involves natural language processing as it is required to know if given data is a positive or negative indicator of financial growth. Thus, this information shall be playing crucial role in task prediction. Interest rates vary as it depends on GDP, savings (supply of credit), investment (demand for credit), government spending, money supply (monetary policy) and taxation. The information related to these variables is extracted using available data and thus is used for the prediction of interest rates. The application can be used as decision making tool by both the banks and customers. Financial companies can use this to make decisions regarding the fixed interest rates. For instance, if in near future the interest rate is going to be at peak, the bank should mark the fixed interest rate near that predicted figures. For customers, it will help to decide whether to choose fixed interest rate or floating interest rate. The project consists of three core modules. Each of these helps to gain some expertise in different fields. Data analysis will help in acquiring some knowledge related to natural language processing and data mining. Interest Rate prediction will help to get some knowledge of statistical learning and use it for the prediction task.

1.5 Scope of Project

The work focuses on knowledge based data mining of news information on internet using different techniques, to predict the pattern of interest rates. The primary user of the software will be banks and customers of banks who wish to know the pattern of rise and fall of the interest rates so as to use it for their benefit. The scope of the work can be explained in two different scenarios. They are mentioned as below:

i. View point of Financial Company:

Financial company like banks makes money by providing capital to different firms and customers as loan. The loan is provided on certain interest rate. The interest rate has to be optimal so as they make good profit. The interest rate should not be set so high that no customer seeks for loan from them. The company also should consider the fact that the interest rate might go up or down by certain margin in days to come. So, considering these scenarios, the interest rate should be set to an optimal value. Thus the software will be helpful for making decision regarding interest rate.

ii. View point of Customers:

A firm or customer will always look for loan at the least available interest rate. One may go for either fixed interest rate which will be fixed for the return time period or for floating interest rate which varies with time. If it is known in advance that the floating interest rate in near future will go down, then one will probably go for floating interest rate while taking loan, but if it is the other way then it would be wise to go for fixed rate. Thus the software will be helpful to customers in making such decisions. Besides these, the tool that is aimed to be developed can be helpful in other various financial analyses.

1.6 Organization of Report

This chapter explains the project background, overview, problem statement, aims and objectives and scope of the project.

Chapter 2 explains about the literature review, which gives us the previous work done in the field of Interest Rate Prediction, Sentiment Analysis of Keywords and Machine Learning.

Chapter 3 provides the information about Current Economic Trends in Nepal. They contain various graphs that show the current economic trend in Nepal.

Chapter 4 is all about the Research in our project. It contains information regarding the research work in our project.

Chapter 5 provides the related theoretical backgrounds. Various theoretical information regarding scrapping, knowledge representation, cognitive map, keyword sentiment analysis and prediction models can be obtained from this section.

Chapter 6 explains in detail about the technological backgrounds. The technical details of web scrapping, knowledge representation, keyword sentiment analysis and prediction models are explained in this chapter.

Chapter 7 and 8 contains details regarding the system design. It has several diagrams like state diagram, ERD , Activity Diagram etc. that effectively models the system.

Chapter 9 explains about the implementation part of the system. Topics like Overall System Workflow, Data Collection Implementation, Cognitive Map and Implementation of Interface are included in this chapter.

Chapter 10 contains brief descriptions about tools and technologies used.

Chapter 11 shows the results obtained from our project. This chapter contains various figures and screenshots that shows the output of our project.

Chapter 12 explains in brief the future enhancement that can be made on the project and also the conclusion derived from the project.

Chapter 2

2. LITERATURE REVIEW

2.1 Existing research on prediction using Artificial Intelligence

This section describes the existing research used for prediction with the use of AI and its methodologies.

2.1.1 Stock Market Prediction using Multiple Regression, Fuzzy Type-2 Clustering and Neural Network

In this research paper a three stage stock market prediction is introduced[2]. In the first stage Multiple Regression Analysis is applied to define the economic and financial variables which have a strong relationship with the output. In the second phase, Differential Evolution-based type-2 Fuzzy Clustering is implemented to create a prediction model. For the third phase, a Fuzzy type-2 Neural Network is used to perform the reasoning for future stock price prediction. Multiple Regression Analysis is for the initial and the most challenging steps of stock market prediction to determine the manageable amount of the input variables which have the strongest forecasting ability and is used as inputs to a prediction system (the Fuzzy type-2 Neural Network) . It is performed on 25 finance and economic variables to both reduce the dimensionality of the variable set and identify which have a strong relationship with the market price for subsequent testing period.

The Fuzzy clustering used in this system is a well established paradigm used to generate the initial type-2 If-Then rules. It removes the uncertainty in choosing the "*m* parameter" existing in Fuzzy c-means by suggesting a solution for a range of its values covering {1.4, 2.6}, a meaningful range for "*m*". This approach resulted in a lower prediction error as compared to a Fuzzy type-1 approach.

2.1.2 Type-2 Fuzzy Clustering and Type-2 Fuzzy Inference Neural Network for the Prediction of Short-Term Interest Rates

This paper discusses the use of a hybrid model for the prediction of short-term interest rates[3]. The model used consists of a differential evolution-based fuzzy type-2 clustering with a fuzzy type-2 inference neural network, after input preprocessing with multiple regression analysis. The model described in this paper forecast the US 3-Month T-bill rates. The results in this paper indicate that the artificial neural network hybrid model with differential the differentia evolution optimization-based fuzzy clustering and fuzzy inference achieves a very good performance with RMSE = 0.7952.

2.2 Existing Methods for Knowledge Representation

2.2.1 Corpus-Based Knowledge Representation

A corpus based knowledge representation system consists of a large collection of disparate knowledge fragments or schemas and a rich set of statistics computed over the corpus[4]. It argues that by collecting such a corpus and computing the appropriate statistics, corpus-based representation offers an alternative to traditional knowledge representation for a broad class of applications. The key advantage of Corpus based knowledge representation is that it avoids the laborious process of building a knowledge base. It describes the basic building blocks of a corpus-based representation system and a set of applications for which such a paradigm is appropriate. Corpus-based Knowledge Representation is an outgrowth of the work on schema and ontology matching. The matching problem is to find a semantic mapping between two disparate representations, be they database schemas or ontologies. This approach is based on analyzing the variations of representations in a corpus of schemas. It extends the idea of using a corpus to a general approach to knowledge representation.

3. ECONOMIC MARKET AND INTRESET RATE TRENDS IN NEPAL

Currently different commercial and development banks provide different loans. They provide loan in both floating rate and fixed rate. These loans are provided for different purposes. Some of the very popular types of loans are Education Loan, Home Loan, Auto Loan etc. The interest rate of loan also varies as per the type of loan.
Some of the key players that affect the economy of a country and ultimately also affect the interest rate were identified.

3.1 Economic Market

The current trends of these key players are listed below:

3.1.1 Foreign Exchange Rates (FOREX)

The trend of FOREX was plotted. The USD to NPR exchange rate was taken as the standard to measure FOREX. The graph below shows that the FOREX is highly fluctuating with time.

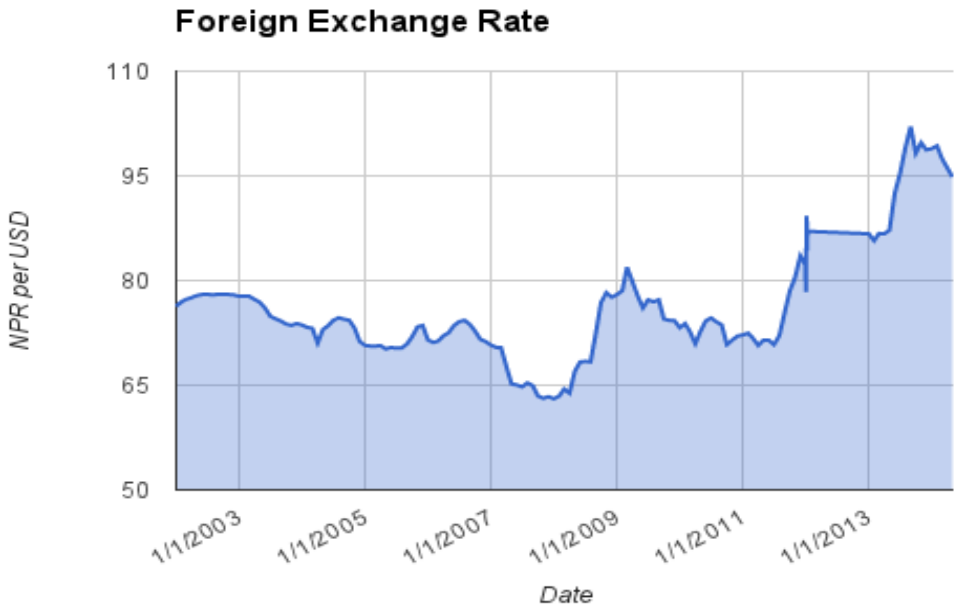

Fig 1: Area Graph of FOREX

3.1.2 Gold Price

The area graph of Gold Price was plotted against the time. The graph resembling with stair case was obtained. It can be seen that the Gold Price is increasing almost exponentially with time.

Fig 2: Area Graph of Gold Price

3.1.3 Petrol Price

The price of petroleum products has also been a key player in economy of Nepal. So to incorporate its effect the Price of Petrol per litre was included as a major economic factor. Expect some spikes it can be seen that the price is increasing exponentially.

Fig 3: Area Graph of Petrol Price

3.1.4 Gross Domestic Product (GDP)

GDP of Nepal is continuously increasing forming a staircase shape in area graph shown below. GDP is also increasing exponentially with time.

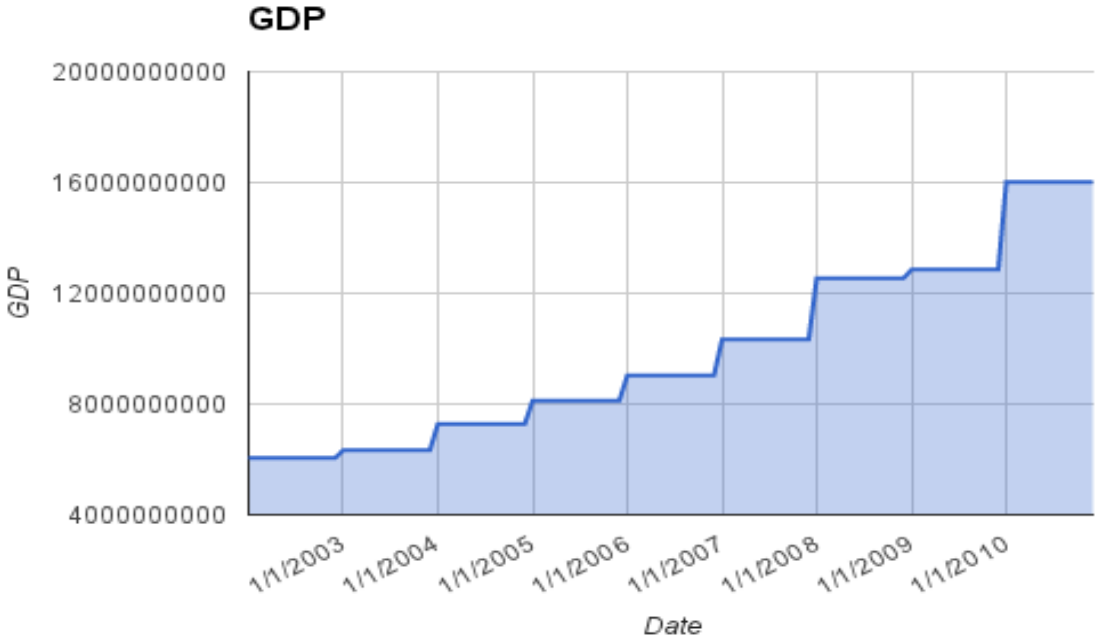

Fig 4: Area Graph of GDP

3.1.5 Gross National Income (GNI)

From the area graph plotted against time it can be clearly seen that GNI of Nepal is also continuously increasing. The trend of the graph is nearly exponential.

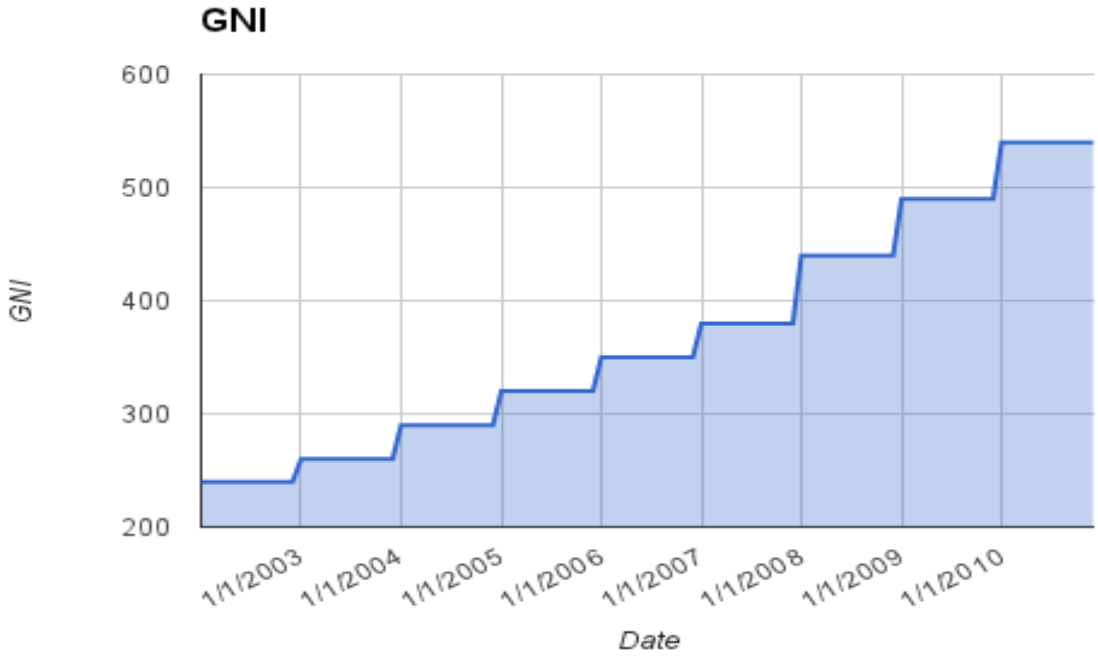

Fig 5: Area Graph of GNI

3.1.6 Interest Rate

The average interest of banks can be seen in the graph below. The interest was highest in the year 2002. It gradually decreased and was lowest between the year 2007 and 2008. It again increased gradually.

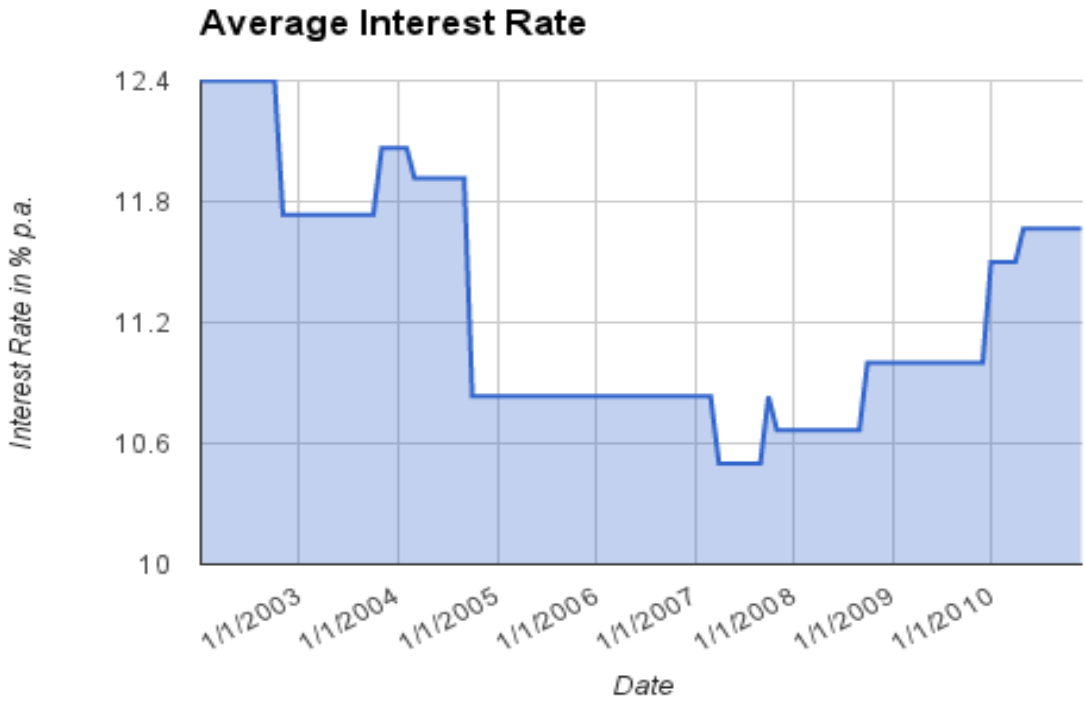

Fig 6: Area Graph of Interest Rate

3.2 Interest Rate Trends in Nepal

Though the projects for Interest Rate Prediction is not new in the global context, it is first of its kind for Nepal. Different projects are going on globally related to financial forecasting. The motto of these projects is to earn huge amount of profit as possible by accurately predicting the future and making decisions accordingly.

In the beginning of the project similar works were searched. Several works were found that used financial data affecting interest rates. These financial data were used to train a model and using the model the interest rate was predicted.

But only one work was found that used news data to predict the interest rate. On searching further, our hypothesis that "news affects the interest rate" matched with one of six basic tenets of the Dow Theory for stock price. Dow Theory tells that "**The**

stock market discounts all news." and the same theory is being tested for interest rate prediction.[5]

One work was done similar to ours by Taeho Hong and Ingoo Han in his paper "Knowledge-based data mining of news information on the Internet using cognitive maps and neural networks".[6] They have developed a Knowledge-Based News Miner (KBNMiner), which is designed to represent knowledge of interest rate experts with cognitive maps (CMs), to search and retrieve news information on the Internet according to prior knowledge, and to apply the information, which is retrieved from news information, to a neural network model for the prediction of interest rates.

3.2.1 Interest Rate in Nepal

Banks in Nepal offer different interest rates for depositing and borrowing money. There are different interest rates for different kind of loans:

i. Types of Loan

The commercial banks and development banks here in Nepal provide different types of Interest Rates. The interest Rates can broadly be divided into two different categories:

a. Floating Interest Rate

This is a type of Interest Rate which does not have a fixed rate of interest. The interest rate varies with time.

b. Fixed Interest Rate

This is a type of Interest Rate which has a fixed rate of interest over the life of the loan. This allows the borrower to accurately predict their future payments.

3.2.2 Data Source

In this project we are dealing with two different kinds of data. They are:

a. Numerical Data

The numerical data were collected as follows:

- **FOREX**

  The data of FOREX was collected from the website of Nepal Rastra Bank.

- **Gold Price**

  Though we were unable to find Gold Price of Nepal we have taken Gold Price of India from a website.

- **Petrol Price**

  Petrol Price was collected from the website of Nepal Oil Corporation.

- **GDP and GNI**

  The data regarding the GDP and GNI of Nepal was collected from the official website of World Bank.

- **Average Interest Rate**

  The interest rates were collected from three different banks viz. Nepal Bank Limited, Asian Development Bank Limited and Rastriya Banijya Bank. We then took average of their interest rates.

b. News Data

As we only found archived news in the website of ekantipur, we have collected the required news data from the website **ekantipur.com**. We wrote a script to scrape the news and store in our database.

3.2.3 Data Storage

The collected data were stored in database. We've used MySQL database to store our data.

Chapter 5

5. THEORETICAL BACKGROUND

5.1 Data Scraping/Web Scraping

As per wikipedia “Data scraping is a technique in which a computer program extracts data from human-readable output coming from another program.”. Web scraping is the branch of data scraping in which we use computer program to extract data from web-pages to the format we want (like xml, json, csv, tsv and etc.). There are several techniques used for web scraping. Some of them are as follows:

i. Human Copy and Paste

It refers to human’s manual examination and copy-and-paste. Sometimes this may be the only workable solution when the websites for scraping explicitly set up barriers to prevent machine automation.

ii. Text grepping and regular expression matching

A simple yet powerful approach to extract information from web pages can be based on the UNIX grep command or regular expression-matching facilities of programming languages (for instance Perl or Python).

iii. HTTP programming

Static and dynamic web pages can be retrieved by posting HTTP requests to the remote web server using socket programming. In UNIX we also use bash commands like curl to retrieve web pages by posting HTTP requests.

iv. HTML parsers

Many websites have large collections of pages generated dynamically from an underlying structured source like a database. Data of the same category are typically encoded into similar pages by a common script or template. In data mining, a program that detects such templates in a particular information source, extracts its content and translates it into a relational form called a wrapper.

v. DOM parsing

By embedding a full-fledged web browser, such as the Internet Explorer or the Mozilla browser control, programs can retrieve the dynamic content generated by client-side scripts. These browser controls also parse web pages into a DOM tree, based on which programs can retrieve parts of the pages.

vi. Web-scraping software

There are many software tools available that can be used to customize web-scraping solutions. This software may attempt to automatically recognize the data structure of a page or provide a recording interface that removes the necessity to manually write web-scraping code, or some scripting functions that can be used to extract and transform content, and database interfaces that can store the scraped data in local databases.

vii. Vertical aggregation platforms

There are several companies that have developed vertical specific harvesting platforms. These platforms create and monitor a multitude of “bots” for specific verticals with no man-in-the-loop, and no work related to a specific target site. The preparation involves establishing the knowledge base for the entire vertical and then the platform creates the bots automatically.

viii. Semantic annotation recognizing

The pages being scraped may embrace metadata or semantic markups and annotations, which can be used to locate specific data snippets. If the annotations are embedded in the pages, as Microformat does, this technique can be viewed as a special case of DOM parsing. In another case, the annotations, organized into a semantic layer, are stored and managed separately from the web pages, so the scrapers can retrieve data schema and instructions from this layer before scraping the pages.

ix. Computer vision web-page analyzers

There are efforts using machine learning and computer vision that attempt to identify and extract information from web pages by interpreting pages visually as a human being might.

5.2 Knowledge Representation

Knowledge is an interesting concept that has attracted many philosophers since a long time back. In recent times, particularly many efforts have been made to represent knowledge in a more applied way with an aim to bring life to machines. Although Artificial Intelligence (AI) has contributed a lot to extract useful knowledge from the raw data , knowledge is an invisible concept to represent. There are mainly two difficulties in representing knowledge. First and more severe problem is that knowledge is built differently among different individuals corresponding to their own perspective. Each individual has his own views on things and events and the complete communication of the entire experience is something very difficult. Another problem is knowledge is invisible. Knowledge may be differently represented in the process of visualization despite the fact that they might be originating from the same concept. Despite these difficulties AI has developed different techniques to represent different knowledge of data or human beings.

Knowledge representation and reasoning (KR) is the field of artificial intelligence (AI) devoted to representing information about the world in a form that a computer system can utilize to solve complex tasks such as diagnosing a medical condition or having a dialog in a natural language. Knowledge representation incorporates findings from psychology about how humans solve problems and represent knowledge in order to design formalisms that will make complex systems easier to design and build.

The knowledge that is used in the tasks like the recognition of visual images, the understanding of natural language, medical diagnosis or robot arm control is very different. Nevertheless, there are some common traits of these different kinds of knowledge:

1. The Knowledge is largely qualitative (symbolically expressed), consisting mainly of rules (*if x then y),* relations (*x is a y)*, procedures (*do x then y)* and the properties with values on a qualitative scale (*the person x is severely ill* vs. *the variable x has the value 8.5*).
2. The knowledge includes abstract concepts. A simple collection of data is not considered to be knowledge. Knowledge includes capacity of generalization (*all x are like y*).
3. The Knowledge is closely related with the process that uses it. Because of

the qualitative nature of knowledge, the process which uses it will be based on inference rather than on arithmetic.

There are some of the requirements that must be taken into consideration while representing knowledge:

1. **Understandability**

A knowledge processing system should be able to represent his knowledge in such a way that people can understand it. This can be realized by using Natural Language or Graphical Communication, but also by structuring knowledge in such a way that it feels natural to the people.

2. **Expressive Power**

A representation formalism should be powerful enough to absorb all the knowledge that the person wants to express. This criterion does not seem to be realized by any current formalism. Moreover people often have to express their knowledge in a formalism which is different from the way they normally express it.

3. **Modularity**

Knowledge should be structured in a modular way , so that the system is flexible enough to add or change knowledge at any point. This is particularly important in a environments dealing to rapidly changing knowledge bases.

From the perspective of knowledge representation Cognitive Map (CM) is a proper tool by which human knowledge can be captured.

5.2.1 Cognitive Map and Causal Relationship

Cognitive Map was introduced by Axelrod which was originally used to represent cause effect relationship which may exist between the elements of environment. The basic elements of cognitive map are simple. Each concept is represented as *points* of cognitive map whereas the *arrows* represent the causal relationship between these concepts. This graph of *points* and *arrows* is called cognitive map. Causal relationship can take on values + (where one concept effects positively to another concept, like enhancing, improving, helping etc), - (where one concept effects negatively to another concept, like harms, retards, decreases, etc) and 0 (has no

relationship or does not affect) . This type of representation makes it easy to analyze how concepts and causal relations are related to one another. For instance, Figure 1 shows the CM studied by Welman and prepared by Levi and Tetlock [7] explains how Japanese made decision to attack Perl Harbor. This map shows that *remaining idle* promotes the *Japanese attrition* while enhancing the preparedness of U.S. army, both of which decrease the prospect of *Japanese success in war.* Here, the CM shows a set of concepts like : *remaining idle, Japanese attrition, U.S. preparedness, Japanese success in war* and a set of assigned edges representing causal relationship like *promotes, enhances, decreases etc.*

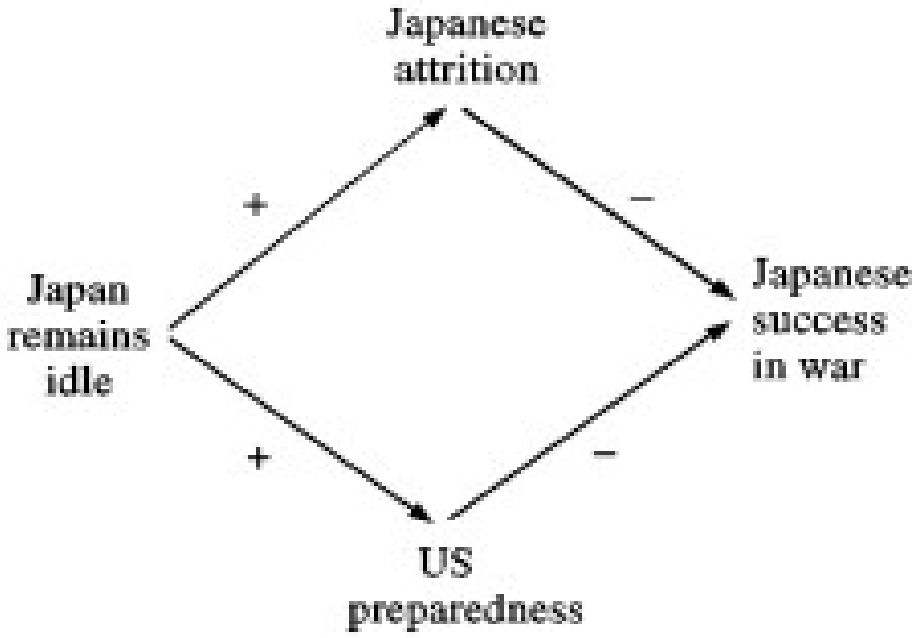

Fig 7: Sample Cognitive Map

As any cognitive map, CM of Figure 1 can be transformed in a matrix called an adjacency or valency matrix. A valency matrix is a square matrix with one row and one column for each concept (node) in a CM. For the above figure 1 we can represent the valency matrix as follows:

$$\begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 0 & + & 0 & + \\ 0 & 0 & - & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & - & 0 \end{pmatrix} \end{array}$$

With the valence matrix as shown above, we draw intuitive framework to form decisions. Thus , to reiterate, CMs are a power tool which allows users to represent and reason on causal relationships as reflected in realistic dynamic systems.

5.3 Sentiment Analysis

Sentiment analysis (also understood 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. Sentiment analysis is used in various applications such as polarity analysis of sentences, information retrieval from text, question-answering system, summarization of texts and many more. Our system finds its usages in polarity analysis of sentences.

5.3.1 Polarity Analysis of Sentence

Polarity analysis of sentence means the problem of classifying a sentences into to class of three main classes: Positive, Negative or Neutral. Sentence polarity analysis has been a topic of research for years, and over these years different approaches for the analysis has been described. This popularity of the analysis reflects its need and usage in various fields, such as:

1. Applications to Review-related Websites
2. Applications as a Sub-Component technology
3. Applications in Business and Government Intelligence
4. Applications across different Domains

Most existing techniques for polarity analysis sentiment classification are based on

supervised learning, for example, n-gram features and three machine learning methods (Naïve Bayes, Maximum Entropy classification, and SVM) can be applied to achieve the high performance[8].

i. Machine Learning Approach

Machine Learning approach treats sentiment classification simply as a special case of topic-based categorization (with the two "topic" being positive sentiment and negative sentiment). Some the primary algorithm used in this approach are:

1. Naive Bayes Classification
2. Maximum Entropy
3. Support Vector Machines(SVM)

ii. Semantic Orientation Approach

The semantic orientation approach performs classification based on positive and negative sentiment words and phrases contained in each evaluation text. It does not require prior training in order to mine the data. Two types of techniques have been used in previous sentiment classification research using the semantic orientation approaches.

1. Corpus-Based Techniques
2. Dictionary-Based Techniques

5.3.2 Semantic Analysis of Keyword in a Sentence

Semantic Analysis of keyword in sentence is the problem of sentiment analysis that involves classification of the orientation of a keyword in a sentence. The orientation can be classified in various different classes depending on the context of a problem. The most common orientation categorizes are the "Positive", "Negative" and "Neutral" classes. Some other semantic orientation could be "Increasing" , "Decreasing" or "None" classes. Our system is based on the later one. They are different approaches to semantic analysis of keyword, they are as follows:

5.3.2.1 n-gram Matching Technique

An n-gram is a contiguous sequence of n items from a given sequence of text. The

items can be letters, words or base pairs according to the application. An n-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram" (or, less commonly, a "digram"); size 3 is a "trigram". Larger sizes are sometimes referred to by the value of n, e.g., "four-gram", "five-gram", and so on.

For a sentence:

"The glass is half full."

3-gram based on letters would be:

{__T,_Th,The, gla, las, ass, ssi, sis, ish, sha, hal, alf, lff, ffu, ful, ull,ll_,l__}

2-gram based on words would be:

{(The glass), (glass is), (is half), (half full)}

n-gram Matching technique is a simple matching process where by a sentence is broken down into a set of n-grams which are then matched with the combination of keywords and polarity words.

For a sentence:

"The unemployment rate is decreasing."

The n-gram {unemployment rate and decreasing} matches the sentence. Thus, it can be inferred that the keyword "unemployment rate" in the sentence is decreasing. For this process to work efficiently, all the synonyms of the class "increasing" and "decreasing" are stored for the references, increasing broader semantic references[9].

5.3.2.2 Opinion Phrase Extraction Techniques

Opinion phrase in a sentence is described as the sentiment phrases in the sentence that are responsible for the semantic orientation of the sentence. Opinion phrase are the combination of sentiment words such as verbs, adjectives and adverbs. Opinion Phrase extraction techniques works by finding out such combination of verbs, adjective and adverb in sentences. The sentiment of keyword is then determined by the sentiment of nearest Opinion phrase from the keyword.

For example, in sentences:

```
The economic development of the country has been slow.

Opinion phrases: {"been slow"}

The painting is beautifully crafted thus aesthetically
pleasing.

Opinion phrases: {"beautifully crafted","aesthetically
pleasing"}

In other areas Nepal has made commendable progress.

Opinion phrases: {"commendable progress"}
```

Extraction of opinion phrase making is easier to handle negation in sentences as well. The assignment of opinion phrase to keyword is done by calculating its distance from the keyword, measured as word distance.

Opinion Phrase extraction depends on the use of a POS tagger for the sentence. This provides the required distinction between the part-of-speech of words in the sentences.

i. POS Tagger

A Part-Of-Speech Tagger (POS Tagger) is a 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'.

For example:

```
"Actor and comedian Robin Williams dies at the age of 63."
```

POS tagger output:

```
Actor/NN and/CC comedian/NNP Robin/NNP Williams/NNP dies/VBZ
at/IN the/DT age/NN of/IN 63/CD ./.
```

ii. Stemmer

Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form—generally a written word form. Stemming program/tools are

commonly referred to as Stemmers.

Below are some examples of Stemmer outputs:

Word	Stemming
land transaction	land transact
investment	invest
industrialization	industri

Table 1: Stemmer Output

5.3.2.3 Kernel Sentence Extraction Technique

Kernel sentence extraction technique is an advanced technique that involves breaking a complex sentence into kernel sentences. Kernel sentences are the sentences with a single verb.[10] Breaking up complex sentence into simpler sentence allows easier analysis of sentences. The process reliant upon the use of Syntactic Parser for dependency tree generation. This tree is then appropriately broken down to create kernel sentences. Kernel sentences are generally represented by ternary expression such as:

```
<subject, verb, object>
```

For a sentence:

```
"Despite the increase in economy, unemployment rate is still
                        increasing."
```

Kernel Sentences would be:

```
<X,increase,economy> , <unemployment rate,increasing,X>
```

Here, X represents "empty".

i. Syntactic Parser

Parsing or syntactic analysis is the process of analyzing a string of symbols in natural language according to the rules of a formal grammar. Syntactic Parser is a tool for syntactic analysis of sentence.

For a sentence:

"Boeing is located in Seattle."

The output of a Syntactic Parser would be:

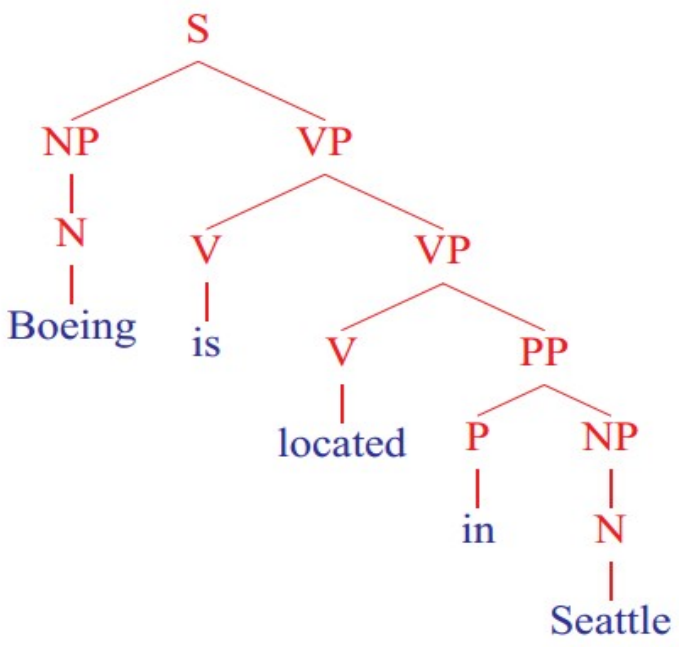

Fig 8: Parser Output Tree

As the result shows, a syntactic parser provides information about the part-of-speech of the words in the sentence, the phrases in the sentences such as NP, VP , PP in the figure and it also shows useful relationships in the sentence.

5.4 Prediction Models

5.4.1 Moving Average

In statistics, a moving average (rolling average or running average) is a calculation to analyze data points by creating a series of averages of different subsets of the full data set. It is also called a moving mean (MM) or rolling mean and is a type of finite impulse response filter. Variations include: simple, and cumulative, or weighted forms (described below).

Given a series of numbers and a fixed subset size, the first element of the moving average is obtained by taking the average of the initial fixed subset of the number series. Then the subset is modified by "shifting forward"; that is, excluding the first number of the series and including the next number following the original subset in the series. This creates a new subset of numbers, which is averaged. This process is repeated over the entire data series. The plot line connecting all the (fixed) averages is the moving average. A moving average is a set of numbers, each of which is the average of the corresponding subset of a larger set of datum points. A moving average may also use unequal weights for each datum value in the subset to emphasize particular values in the subset.

A moving average is commonly used with time series data to smooth out short-term fluctuations and highlight longer-term trends or cycles. The threshold between short-term and long-term depends on the application, and the parameters of the moving average will be set accordingly. For example, it is often used in technical analysis of financial data, like stock prices, returns or trading volumes. It is also used in economics to examine gross domestic product, employment or other macroeconomic time series. Mathematically, a moving average is a type of convolution and so it can be viewed as an example of a low-pass filter used in signal processing. When used with non-time series data, a moving average filters higher frequency components without any specific connection to time, although typically some kind of ordering is implied. Viewed simplistically it can be regarded as smoothing the data.

Let us take an example to illustrate how moving average works.

```
8,1,3,2,9,1,5
```

Now let us use a window of length N=3 to and compute the average using that window.

```
Ma1 = (8+1+3)/3 = 4

Ma2 = (1+3+2)/3 = 2

Ma3 = (3+2+9)/3 = 4.66

Ma4 = (2+9+1)/3 = 4

Ma5= (9+1+5)/3 = 5
```

In general, `Ma = (Pm + Pm-1 …..+ Pm-(n-1) )/n .....(1)`

Since the model assumes a constant underlying mean, the forecast for any number of periods in the future is the same as the estimate of the parameter Ma.

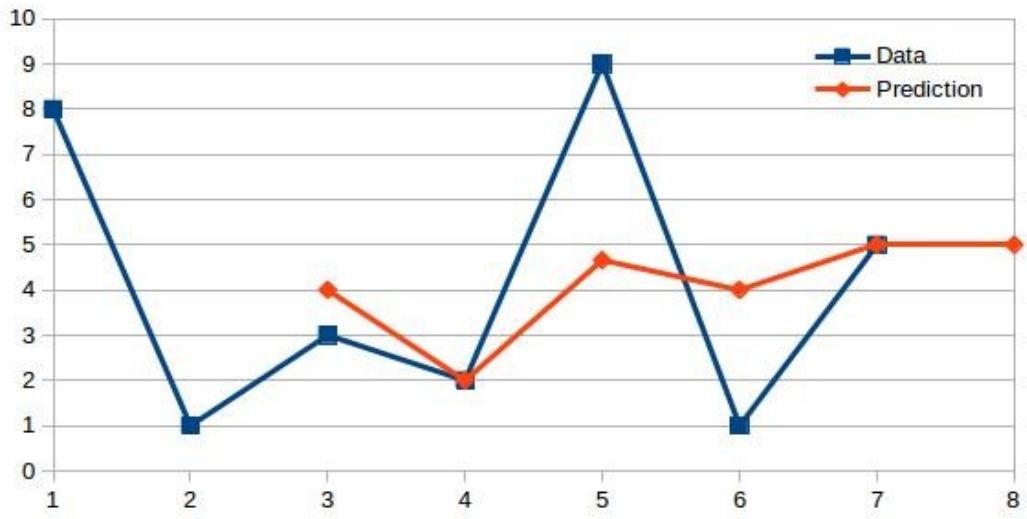

Fig 9: Sample Prediction using Moving Average

5.4.2 Single Exponential Smoothing

Exponential smoothing is a technique that can be applied to time series data, either to produce smoothed data for presentation, or to make forecasts. The time series data themselves are a sequence of observations. The observed phenomenon may be an essentially random process, or it may be an orderly, but noisy, process. Whereas in the simple moving average the past observations are weighted equally, exponential

smoothing assigns exponentially decreasing weights over time.

Exponential smoothing is commonly applied to financial market and economic data, but it can be used with any discrete set of repeated measurements. The simplest form of exponential smoothing should be used only for data without any systematic trend or seasonal components.

The raw data sequence is often represented by {xt} beginning at time t=0, and the output of the exponential smoothing algorithm is commonly written as {st}, which may be regarded as a best estimate of what the next value of x will be. When the sequence of observations begins at time t = 0, the simplest form of exponential smoothing is given by the formulae.

```
S0 = X0  .................................(2)

St  = alpha * Xt-1 + (1-alpha)* St  t>0.....(3)
```

where alpha is smoothing factor. 0<alpha<1

5.4.3 Linear Regression

In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables denoted X. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. (This term should be distinguished from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.)

In linear regression, data are modeled using linear predictor functions, and unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, linear regression refers to a model in which the conditional mean of y given the value of X is an affine function of X. Less commonly, linear regression could refer to a model in which the median, or some other quantile of the conditional distribution of y given X is expressed as a linear function of X. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of y given X, rather than on the joint probability distribution of y and X, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares loss function as in ridge regression. Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

Let us formulate the model that represents a simple linear regression. The equation is given below.

$$Y = mX + c \quad \text{.....(4)}$$

Y is the output value and X is the input value where as m is slope and c is intercept but in statistics these are termed differently.

- Y is called the regressand, endogenous variable, response variable, measured variable, criterion variable, or dependent variable
- X are called regressors, exogenous variables, explanatory variables, covariates, input variables, predictor variables or independent variables
- m is called as parameter vector
- c is called as noise,error term or disturbance term

5.4.5 Multiple Regression

Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables[11]. The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable). The variables we are using to predict the value of the dependent variable are called the independent variables (or

sometimes, the predictor, explanatory or regressor variables).

A model of the following form is said to be multiple regression.

```
Y = c + a1 X1 +a2X1 + …... + an Xn .............(5)
```

During training a regression model the X given above is a vector of training examples. It is n by m dimensional vector where n is the number of rows(training examples) and m is number of columns which is number of features. Y is n by 1 dimension vector. It column vector and is output corresponding to each of the training set in X. a is the parameter vector of 1 by m size. It is row vector and m is the total number of parameter and is equal to number of features. C is a scalar value. So as vector the above equation can be written as:

```
[Y]  = [a]*[X] ..............................(6)
```

Unlike simple linear regression where there is only one feature input one output, in multiple regression the output depends upon various independent inputs .i.e {X} and thus the trained coefficients are equals the number of independent inputs and is challenging problem to find them. In learning, the model is given inputs and outputs and it is expected to find [a] and c.

5.4.6 SVM (Support Vector Machine)

In machine learning, support vector machines (SVMs, also support vector networks are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

More formally, a support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it was proposed that the original finite-dimensional space be mapped into a much higher-dimensional space, presumably making the separation easier in that space. To keep the computational load reasonable, the mappings used by SVM schemes are designed to ensure that dot products may be computed easily in terms of the variables in the original space, by defining them in terms of a kernel function k(x,y) selected to suit the problem. The hyperplanes in the higher-dimensional space are defined as the set of points whose dot product with a vector in that space is constant. The vectors defining the hyperplanes can be chosen to be linear combinations with parameters αi of images of feature vectors that occur in the data base. With this choice of a hyperplane, the points x in the feature space that are mapped into the hyperplane are defined by the relation.

$$\sum_i \alpha_i k(x_i, x) = \text{constant.}$$

If k(x,y) becomes small as y grows further away from x, each term in the sum measures the degree of closeness of the test point x o the corresponding data base point xi .In this way, the sum of kernels above can be used to measure the relative nearness of each test point to the data points originating in one or the other of the sets to be discriminated. Note the fact that the set of points x mapped into any hyperplane can be quite convoluted as a result, allowing much more complex discrimination between sets which are not convex at all in the original space.

Classifying data is a common task in machine learning. Suppose some given data points each belong to one of two classes, and the goal is to decide which class a new data point will be in. In the case of support vector machines, a data point is viewed as a p-dimensional vector (a list of p numbers), and we want to know whether we can

separate such points with a (p − 1)-dimensional hyperplane. This is called a linear classifier. There are many hyperplanes that might classify the data. One reasonable choice as the best hyperplane is the one that represents the largest separation, or margin, between the two classes. So we choose the hyperplane so that the distance from it to the nearest data point on each side is maximized. If such a hyperplane exists, it is known as the maximum-margin hyperplane and the linear classifier it defines is known as a maximum margin classifier; or equivalently, the perceptron of optimal stability.

5.4.7 Decision Tree

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is one of the predictive modeling approaches used in statistics, data mining and machine learning. More descriptive names for such tree models are classification trees or regression trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making. This page deals with decision trees in data mining.

Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown on the right. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

A decision tree is a simple representation for classifying examples. Decision tree learning is one of the most successful techniques for supervised classification learning. For this section, assume that all of the features have finite discrete domains, and there is a single target feature called the classification. Each element of the domain of the classification is called a class. A decision tree or a classification tree is

a tree in which each internal (non-leaf) node is labeled with an input feature. The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes.

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions. This process of top-down induction of decision trees (TDIDT) is an example of a greedy algorithm, and it is by far the most common strategy for learning decision trees from data.

In data mining, decision trees can be described also as the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data.

Data comes in records of the form:

$$(\mathbf{x}, Y) = (x_1, x_2, x_3, ..., x_k, Y)$$

The dependent variable, Y, is the target variable that we are trying to understand, classify or generalize. The vector x is composed of the input variables, x1, x2, x3 etc., that are used for that task.

Decision trees used in data mining are of two main types:

- Classification tree analysis is when the predicted outcome is the class to which the data belongs.
- Regression tree analysis is when the predicted outcome can be considered a real number (e.g. the price of a house, or a patient's length of stay in a hospital,interest rate prediction).

ID3 is one of the mostly used decision tree algorithm. The ID3 algorithm begins with the original set S as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set S and calculates the entropy H(S) (or information gain GI(S)) of that attribute. It then selects the attribute which has the

smallest entropy (or largest information gain) value. The set S is then split by the selected attribute (e.g. age < 50, 50 <= age < 100, age >= 100) to produce subsets of the data. The algorithm continues to recurse on each subset, considering only attributes never selected before.

Recursion on a subset may stop in one of these cases:

- every element in the subset belongs to the same class (+ or -), then the node is turned into a leaf and labeled with the class of the examples
- there are no more attributes to be selected, but the examples still do not belong to the same class (some are + and some are -), then the node is turned into a leaf and labeled with the most common class of the examples in the subset
- there are no examples in the subset, this happens when no example in the parent set was found to be matching a specific value of the selected attribute, for example if there was no example with age >= 100. Then a leaf is created, and labeled with the most common class of the examples in the parent set.

Throughout the algorithm, the decision tree is constructed with each non-terminal node representing the selected attribute on which the data was split, and terminal nodes representing the class label of the final subset of this branch.

5.4.8 Correlation

For the prediction of interest rate many independent parameters or features are used (GDP, GNI, Gold price,petrol price, FOREX, sentiment value). The dependent variable in this case is interest rate and depends on the above mentioned features. Correlation helps to find out the degree of dependence of interest rate on those independent variables. This helps to understand which feature is effecting the change in interest rate the most.

In statistics, dependence is any statistical relationship between two random variables or two sets of data. Correlation refers to any of a broad class of statistical relationships involving dependence.

Familiar examples of dependent phenomena include the correlation between the physical statures of parents and their offspring, and the correlation between the demand for a product and its price. Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. For example, an electrical utility may produce less power on a mild day based on the correlation between electricity demand and weather. In this example there is a causal relationship, because extreme weather causes people to use more electricity for heating or cooling; however, statistical dependence is not sufficient to demonstrate the presence of such a causal relationship (i.e., correlation does not imply causation).

Formally, dependence refers to any situation in which random variables do not satisfy a mathematical condition of probabilistic independence. In loose usage, correlation can refer to any departure of two or more random variables from independence, but technically it refers to any of several more specialized types of relationship between mean values. There are several correlation coefficients, often denoted ρ or r, measuring the degree of correlation. The most common of these is the Pearson correlation coefficient, which is sensitive only to a linear relationship between two variables (which may exist even if one is a nonlinear function of the other). Other correlation coefficients have been developed to be more robust than the Pearson correlation – that is, more sensitive to nonlinear relationships.

Correlation between dependent and independent variable will give a value between 0 and 1 and shows how strongly the dependent variable is correlated with independent one. The value 0 tells there is no correlation whereas 1 represents strong correlation.

Chapter 6

6. TECHNICAL BACKGROUND

6.1 Web Scrapper

Scrapy framework was used to crawl news from **ekantipur.com**. To scrape the data using scrapy first the site has to be picked. Then the data, we want to scrape was defined through Scrapy Items as:

```
import scrapy

class TorrentItem(scrapy.Item):
    url = scrapy.Field()
    name = scrapy.Field()
    description = scrapy.Field()
    size = scrapy.Field()
```

Next,we wrote a spider that defines the start URL, such as,
http://www.mininova.org/today
The rules for following links and the rules for extracting the data from pages.

The URL format was like http://www.mininova.org/tor/NUMBER where NUMBER is an integer. We'll use that to construct the regular expression for the links to follow: /tor/\d+.

XPath was used for selecting the data to extract from the web page HTML source.

In the HTML source page it can be seen that the file name is contained inside a <h1> tag:

```
<h1>Darwin - The Evolution Of An Exhibition</h1>
```

An XPath expression to extract the name could be:

```
//h1/text()
```

And the description is contained inside a tag with id="description":

```
<h2>Description:</h2>

<div id="description">
Short documentary made for Plymouth City Museum and Art Gallery
regarding the setup of an exhibit about Charles Darwin in conjunction
with the 200th anniversary of his birth.
```

An XPath expression to select the description could be:

```
//div[@id='description']
```

Finally, the file size is contained in the second <p> tag inside the tag with id=specifications:

```
<div id="specifications">

<p>
<strong>Category:</strong>
<a href="/cat/4">Movies</a> &gt; <a href="/sub/35">Documentary</a>
</p>


<p>
<strong>Total size:</strong>
150.62 megabyte</p>
```

An XPath expression to select the file size could be:

```
//div[@id='specifications']/p[2]/text()[2]
```

Finally, here's the spider code:

```
from scrapy.contrib.spiders import CrawlSpider, Rule
from scrapy.contrib.linkextractors import LinkExtractor
class MininovaSpider(CrawlSpider):

    name = 'mininova'
    allowed_domains = ['mininova.org']
    start_urls = ['http://www.mininova.org/today']
    rules = [Rule(LinkExtractor(allow=['/tor/\d+']),
```

```
'parse_torrent')]

    def parse_torrent(self, response):
        torrent = TorrentItem()
        torrent['url'] = response.url
        torrent['name'] = response.xpath("//h1/text()").extract()
        torrent['description'] =
response.xpath("//div[@id='description']").extract()
        torrent['size'] = response.xpath("//div[@id='info-
left']/p[2]/text()[2]").extract()
        return torrent
```

Finally, we'll run the spider to crawl the site and output the file scraped_data.json with the scraped data in JSON format:

```
scrapy crawl mininova -o scraped_data.json
```

6.2 Sentiment Analysis Implementation

6.2.1 POS Tagger

For the purpose of sentiment analysis, a sentence is first fed into a POS tagger which outputs a POS tagged sentences. The POS tagger can be implemented in multiple ways[12]. According to the need in performance and speed, the POS tagger can be chosen. Some of the tagger we used in our system were:

6.2.2 Regexp Tagger

Regexp Tagger works by assigning the tags based in regular expressions. The regular expressions are often describes the morphological structure of words. The rules/regexp in the tagger are processed in order, the first one that matches is applied.

The regexp pattern used in our system is:

```
[(r'^-?[0-9]+(.[0-9]+)?$', 'CD'), # cardinal numbers
(r'.*able$', 'JJ'), # adjectives
```

```
(r'.*ness$', 'NN'), # nouns formed from adjectives
(r'.*ly$', 'RB'), # adverbs
(r'.*ing$', 'VBG'), # gerunds
(r'.*ed$', 'VBD'), # past tense verbs
(r'^[A-Z].*s$', 'NNPS'), # plural proper nouns
(r'.*s$', 'NNS'), # plural nouns
(r'^[A-Z].*$', 'NNP'), # singular proper nouns
(r'.*', 'NN')] # singular nouns (default)
```

The tagging is very straightforward and fast. The tagger is then initialized using the patterns, as:

```
tagger = nltk.RegexpTagger(patterns)
print tagger.tag("…")
```

6.2.3 N-Gram Tagger

N-gram Tagger takes into accounts the grammatical context of words. It works by finding out the most likely tag for each word, given the preceding tag. In a training dataset with sentences and tags as following:

```
"They are content"
Tag: are/VBP content/JJ

"The most important part is content"
Tag: is/VBP content/NN
```

word "content" would more likely be JJ(adjective) when the preceding tag is a VBP(verb) and is likely to be a NN(noun) when followed by VBP(verb).

Then is the test data:

```
"He is content " with is/VBP content/?
Tag: is/VBP content/NN
```

6.2.4 Brill Tagger

Brill tagger is a transformation-based (supervised) learning tagger. The Brill tagger uses the initial POS tagger to produce initial part of speech tags, then corrects those POS tags based on Brill transformational rules. These rules are learned by training the Brill tagger with rules templates. If the word is known, it first assigns the most frequent tag, or if the word is unknown, it naively assigns the tag "noun" to it. Applying over and over these rules, changing the incorrect tags, a quite high accuracy is achieved.

Example: BrillTagger in nltk

```
from nltk.tag.brill import *

templates =
[nltk.tag.brill.SymmetricProximateTokensTemplate(ProximateTagsR
ule, (1,1)),
SymmetricProximateTokensTemplate(ProximateTagsRule, (2,2)),
SymmetricProximateTokensTemplate(ProximateTagsRule, (1,2)),
SymmetricProximateTokensTemplate(ProximateTagsRule, (1,3)),
SymmetricProximateTokensTemplate(ProximateWordsRule, (1,1)),
SymmetricProximateTokensTemplate(ProximateWordsRule, (2,2)),
SymmetricProximateTokensTemplate(ProximateWordsRule, (1,2)),
SymmetricProximateTokensTemplate(ProximateWordsRule, (1,3)),
ProximateTokensTemplate(ProximateTagsRule, (-1, -1), (1,1)),
ProximateTokensTemplate(ProximateWordsRule, (-1, -1), (1,1)),
]

trainer =
nltk.FastBrillTaggerTrainer(initial_tagger=unigramDefault,
templates=templates, trace=3,
deterministic=True)

tagger = trainer.train(training, max_rules=10)
```

Snapshot of rules generated from this training data:

```
113 167 54 1 | NN -> VB if the tag of the preceding word is
'TO'
76 78 2 0 | VBP -> VB if the tag of words i-3...i-1 is 'MD'
72 72 0 0 | NN -> VB if the tag of the preceding word is 'MD'
60 69 9 0 | VBD -> VBN if the tag of words i-2...i-1 is 'VBZ'
59 63 4 5 | IN -> WDT if the text of the following word is
'*T*-1'
58 58 0 0 | VBP -> VB if the tag of the preceding word is 'TO'
56 56 0 1 | POS -> VBZ if the tag of words i-2...i-1 is 'PRP'
54 57 3 0 | VBD -> VBN if the tag of words i-2...i-1 is 'VBP'
52 58 6 0 | VBD -> VBN if the tag of words i-2...i-1 is 'VBD'
48 63 15 2 | VB -> VBP if the tag of the preceding word is
'NNS'
```

6.2.5 Evaluation of Taggers

Each taggers were trained and tested on the 5% of Penn Treebank in NLTK with 90% on training and 10% on testing.

```
training = nltk.corpus.treebank.tagged_sents()[:3522]
test = nltk.corpus.treebank.tagged_sents()[3522:]
```

Upon training each taggers above about with these data, we found the accuracy details as below:

SN.	**Tagger**	**Accuracy**
1.	Regex tagger	34.57 %
2.	Unigram/Regex Tagger	**92.78%**
3.	Bigram/Unigram/Regex Tagger	**93.74%**
4.	Brill with bigram/unigram/regex	**93.98%**

Table 2: Evaluation of Taggers

* % accuracy : number of correct result / total

* Tagger1/Tagger2 : Tagger1 with backoff Tagger2

With the following accuracy results and cross validation with performance speed, Unigram/Regex Tagger was used in our system.

6.3 Porter Stemmer

The Porter stemming algorithm (or 'Porter stemmer') is a process for removing the commoner morphological and inflexional endings from words in English. Its main use in our system is for term normalization process, that is usually done when setting up Information Retrieval systems. It stems using a set of rules, or transformations, applied in a succession of steps without recursion.

Algorithm of Porter Stemmer [13]:

```
Step 1: Gets rid of plurals and -ed or -ing suffixes

Step 2: Turns terminal y to i when there is another vowel in
the stem

Step 3: Maps double suffixes to single ones: -ization,
-ational, etc.

Step 4: Deals with suffixes, -full, -ness etc.

Step 5: Takes off -ant, -ence, etc.

Step 6: Removes a final -e
```

6.4 Keyword Polarity Analysis

6.4.1 n-gram Matching Technique

n-gram Matching techniques starts by initializing all the normalized keywords and opinion words from the database. The sentence feed to the matching technique, if passes the relevancy test, is then searched for any opinion words. This is the matching process. The opinion word is then evaluated for polarity and then assigned to the keyword.

Below is the algorithm used for n-gram in our system:

```
Step 1: Pre-process the Sentence.

      Step 1.1: involves tokenization, lemmatization(if
      necessary), stemming.

Step 2: Check if Keyword Present .

      Step 2.1: check for all the normalized keyword in the
            database.

      Step 2.2: check for all the possible synonyms of a
      keyword.

Step 3: Check if Keyword is increasing/decreasing in the
      context.

      Step 3.1: Create a Neg-Map of the sentence, regions of
      negatives in the                sentence.

      Step 3.2: Look for opinion Words in the sentence.
      Opinion Words are collected as Opinion-Base.

      Step 3.3: Use opinion word to determine the polarity of
the keyword.

      Step 3.4: Use the Neg-Map to re-evaluate the polarity
      of the keyword.
```

6.4.2 Opinion Phrase Extraction Techniques

Opinion phrase extraction technique works in similar way as the n-gram matching technique, with the exception that the range of search is limited (in matching process) and instead of a opinion word, opinion phrases are extracted.

Below is the algorithm for Opinion phrase extraction used in our system:

```
Step 1: Pre-process the sentence

      Step 1.1: involves tokenization, POS tagging, stemming,
      negative tag assigning.

Step 2: Check if Keyword Present

      Step 2.1: check for all the normalized keyword in the
      database.

      Step 2.2: check for all the possible synonyms.

Step 3: Check if Keyword is increasing/decreasing in the context

      Step 3.1: Extract Opinion Phrase, both right and left of the
      keyword.

      Step 3.2: Use distance metric to select Opinion Phrase for the
keyword.

      Step 3.3: Analyze the phrase for Opinion/Polarity.

      Step 3.3: Find out the orientation of Opinion in the phrase
      using negative tag.
```

6.4.3 Kernel Sentence Extraction Techniques

The kernel sentence extraction technique requires a syntactic parser for a dependency tree of a sentence. Given a dependency tree, Kernel sentence extraction works by following the steps given below[14]:

Sentence:

```
      "The book inspired Bob and boosted his confidence when he  was
at an all time low."
```

Dependency Tree:

```
(ROOT
  (S
    (NP (DT The) (NN book))
    (VP
      (VP (VBD inspired)
        (NP (NNP Bob)))
      (CC and)
      (VP (VBD boosted)
        (NP (PRP$ his) (NN confidence))
        (SBAR
          (WHADVP (WRB when))
          (S
            (NP (PRP he))
            (VP (VBD was)
              (PP (IN at)
                (NP (DT an) (DT all) (NN time)))
              (ADJP (JJ low)))))))
    (. .)))
```

i. Pronoun Resolution

The sentence must be looked for any pronouns used. Pronouns can be detected by the use of tag “PRP” with the word. Once the pronoun is detected following algorithm should be used to resolve it:

```
Step 1: Parse tree traversed starting at pronouns.

      - traversal moves up and to the left

Step 2: First noun phrase on traversal path replaces pronouns

      - noun phrase must contain proper nouns or traversal
      continues

      - noun phrase must agree in number with pronoun
```

This will get rid on any pronouns in the sentence, therefore removing any ambiguity offered by pronouns in the later stages during sentence decomposition.

The output of the sentence from this stage would be:

```
"The book inspired Bob and boosted Bob's confidence when Bob
was at an all time low."
```

ii. Sentence Fragmentation

After the pronoun resolution process, the sentence are now fragmented. The fragmentation follows the following algorithm:

```
Step 1: New sentences created from compound structures

      - verb phrases linked by correlative conjunctions

      - main noun phrase used in both sentences

Step 2: Fragments created from independent and subordinate
clauses

      - identified by co-coordinating and subordinate
      conjunctions
```

The output sentences from this stage are the kernel sentences. The output looks like:

```
{
      "The book inspired Bob”,
       ”The book boosted Bob's confidence”,
      ”Bob was at an all time low."
}
```

iii. Kernel Sentence Evaluation

The kernel sentence extracted are then evaluated for keyword polarity. First the sub-sentences is represented as a T-Expression, as:

```
{
      <book,inspired,Bob>,
      <book, boosted, Bob's confidence>,
      <Bob,was at an all time low,X>
}
```

Then the T-expressions are analyzed for polarity using the following steps:

```
Step 1: Check for the keyword in expressions

Step 2: Analyze the "verb phrase" in the expression for
polarity

Step 3: Assign the obtained polarity to the keyword
```

6.4.4 Evaluation of Techniques

After the successful implementation of all the above-mentioned techniques and algorithm, the best algorithm satisfying the constraints and the performance demanded by our system was chosen. For this particular task, several datasets of different sources were collected. The datasets from news, blog and twitter feeds were gathered and manually labeled. The labeled data was feed to each of the polarity analyzing techniques and the results were evaluated.

Following were the data used for the evaluation:

Data	**Number**
News	50 Sentences
Blog	50 Sentences
Twitter	500 Tweets

Table 3: Evaluation Dataset

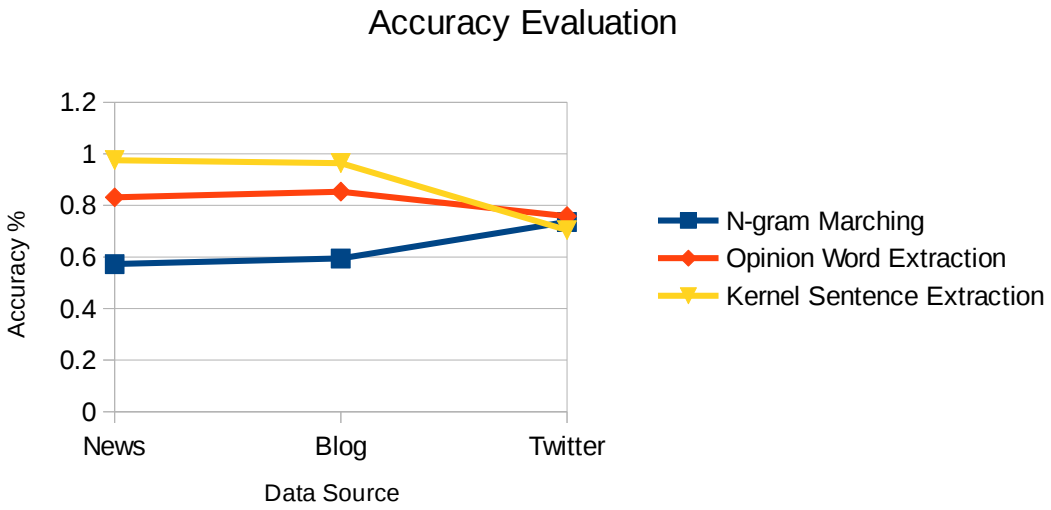

Fig 10: Accuracy Evaluation of Polarity Analyzers

This evaluation was cross referenced with the speed of performance to select the best algorithm for the analysis.

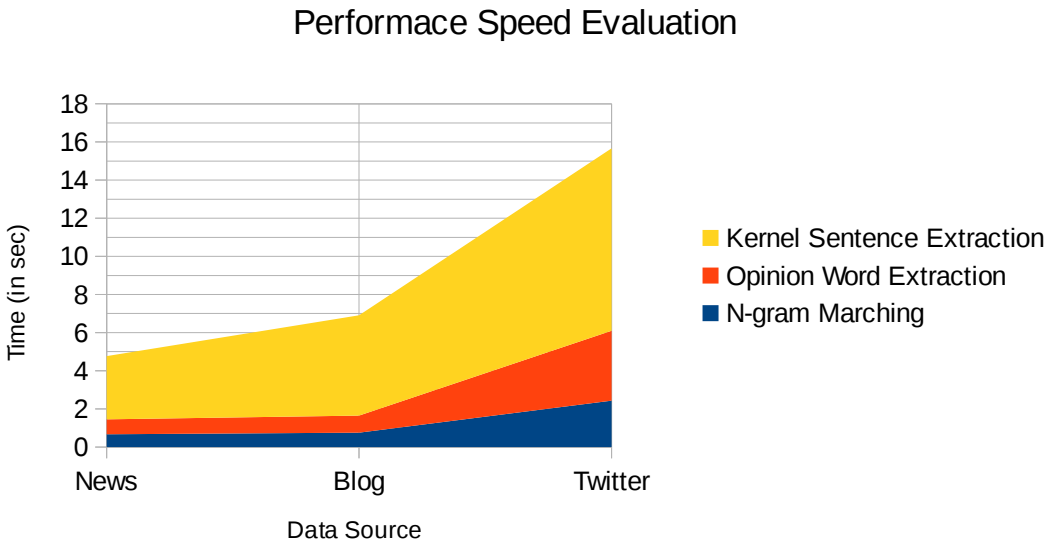

Fig 11: Performance Speed Evaluation of Polarity Analyzers

From the analysis, we found that Opinion Word Extraction technique suits best for our system in consideration to speed and accuracy.

6.5 Model Selection

Out of various model considered for prediction,we only selected a few of them for final use. The selection was based on K-fold cross validation. But before explaining what we did in K-fold cross validation, let us see the models we used.

1. The first model was a simple linear regression which used interest rate as output (y) and input(x) was time label represented as 1,2,3.....n. The equation can be represented as y=mx+c

2. The second model considered was multiple regression model. The model can be represented as

   $y = c + a_1*PP + a_2*GP + a_3 * FO + a_4 *GDP + a_5 *GNI + a_5 *S \quad ...(7)$

   where, PP = petrol price

   GP = gold price

   FO = for-ex

   GDP = gross domestic product

   GNI = gross national income

   S = sentiment value

   To be termed as simple regression model in below figures.

3. The third model was extended form of the above model using squared power of each of the features .i.e(PP,GP,FO,GDP,GNI,S).

   This can be represented as below.

   $y = c + \sum (a_i * F_i) + \sum (b_i * F_i^2) \quad .....(8)$

   where i changes from 1 to N(number of features)

   a_i and b_i are the coefficients

   F_i are the features

   To be termed as Quadratic model in below figures.

4. The fourth model was extended form of the above model using cubic power of each of the features .i.e(PP,GP,FO,GDP,GNI,S). This can be represented as below.

$$y = c + \sum (a_i * F_i) + \sum (b_i * F_i^2) + \sum (c_i * F_i^3) \quad \ldots\ldots\ldots (9)$$

where I changes from 1 to N(number of features)

a_i, b_i and c_i are the coefficients

F_i are the features

To be termed as Cubic model in below figures.

5. This model was adjusted so as to check if there are any sinusoids and cosines component governing the data change pattern. The model can represented as below.

$$y = c + \sum (a_i * \sin(F_i)) + \sum (b_i * \cos(F_i)) \quad \ldots\ldots\ldots (10)$$

where

a_i and b_i are the coefficients , F_i are the features

To be termed as Sinusoidal model in below figures.

6. The another model was SVM.The svm model of sklearn was used ,which is a python machine learning library. Just for representation, the following is the equation.

$$y = svm(f1, f2, f3, \ldots fn) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (11)$$

where f's are features and y is output(interest rate)

7. Decision tree was another model used. We used the decision tree of sklearn.

After having explained these candidate models, let see how K-fold cross validation method can be used to select the best model out of these models. K-fold cross validation is a method of validating different models. In this approach, the training data is divided into k-folds. In the implementation, it was 10-fold validation. In this approach,once the data is divided into 10-folds, 9 out of 10 folds in used for training and 1 fold is used for testing. So there becomes a told of 10 training sets and 10 test sets after running 10 folds. All of the above models were cross validated against k-fold cv and mean squared error was calculated for each model in each fold. Then graph was plotted using each model with k-fold cross validation error, to see which model has got the highest accuracy.

The graph is given below:

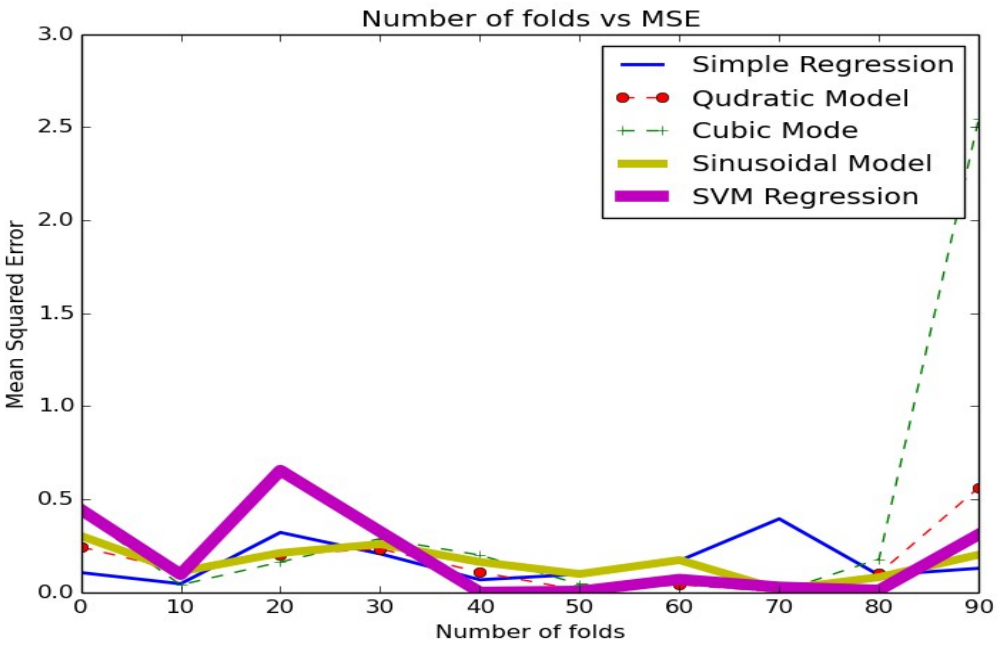

Fig 12: Number of folds vs MSE

From the above graph, it can be seen that Quadratic model, sinusoidal model and simple regression model has got the minimum error over different folds. SVM being the one with the maximum error but has least error at end fold. The model to be selected is Quadratic model and Simple Regression model.

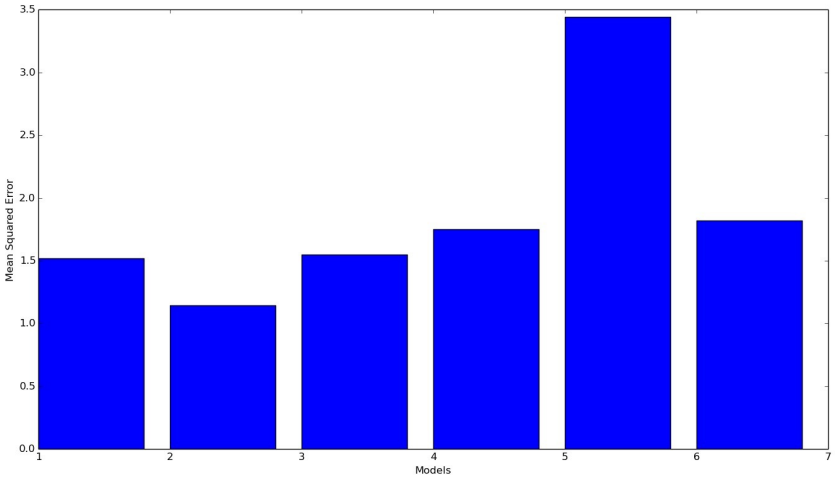

Fig 13: MSE of each model in bar graph

1 = Simple Regression Model 2 = Quadratic model 3 = Cubic Model

4 = Sinusoidal Model 5 = SVM 6 = Decision Tree

The numerical values of MSE error of each of the above given model is in below table:

Models	MSE	Remarks
Simple Regression Model	1.52	(Selected)
Quadratic Model	1.14	Minimum Error(Selected)
Cubic Model	1.55	
Sinusoidal Model	1.75	
SVM	3.44	Maximum Error
Decision Tree	1.63	(selected)

Table 4: MSE error of different models

6.5.1 Model Evaluation

The above selected models can be evaluated by different indicators. Some of the error indicators considered for model evaluation are mean squared error, mean absolute error and root mean squared error. These error values for each of the selected models are tabulated below.

Models	MAE	MSE	RMSE	Remarks
Simple Regression Model	0.23771134657	0.07984018511	0.28256005576	
Quadratic Powered Regression Model	0.22511224790	0.07273068114	0.26968626428 9	
Decision Tree	4.93929177394 e-12	7.40740865189 e-22	2.72165549839 e-11	Minimum Error

Table 5: Error Comparison of Models

The different error indicators are tabulated above. It can be seen that Decision tree gives the minimum error for all of the indicators. But the problem with Decision tree may be overfitting. Overfitting is condition of a model where it correctly covers all the data points given during training but fails to predict for new data. Simple regression model has got error more that in Decision Tree and it has even more error than in Quadratic model. Quadratic model has less error than simple regression model and slightly more error than Decision tree, so in this respect it can be viewed as model with neither a great bias or variance.

6.5.2 Prediction Model

For prediction all of the selected models can be used. The graph of the error these models produce will be given below for further explanation. The model having the equation of form :

$$y = c + \sum (a_i * F_i) + \sum (b_i * F_i^2)$$

is found to be the best suited for our purpose. The model as it is being called is Quadratic model. Decision tree on the other hand can be also handy as it is not much of certain that is has any sorts of overfitting nature. Simple regression model can also used as good model as it is equally good as quadratic model. In this scenario, all of these models will be used for prediction.

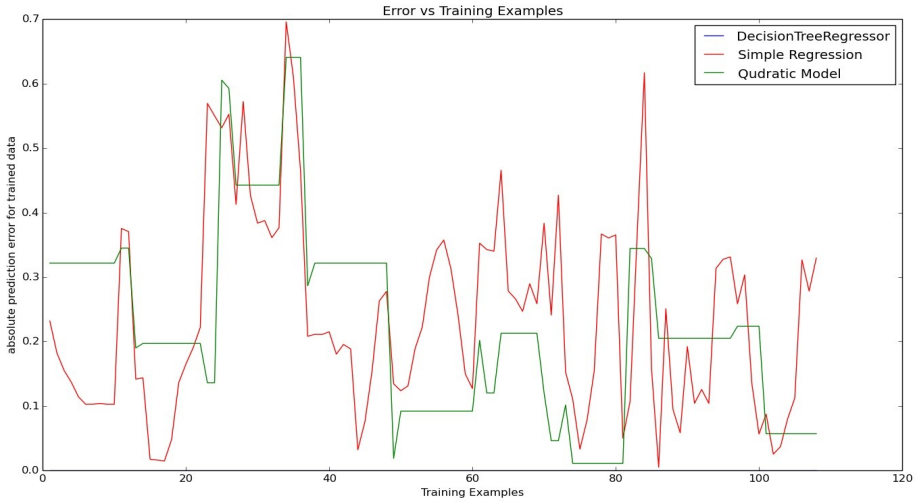

Fig 14: Error Vs Training Examples Graph

The error line of the decision tree is not visible as the value tends to zero for each and every training example. The error of simple regression is much more noisy and fluctuating than that of Quadratic model. The error of quadratic model tends to stay low as much as possible at the end data point. But this graph cannot be used for evaluation of models, as it is just the absolute difference between actual values and predicted values of trained data. New data will be tested and their result will be given in the results section.

Chapter 7

7. SYSTEM ANALYSIS

7.1 Requirement Specification

Requirement specification provides a detailed picture of functional and non-functional requirement of the system. The Software Requirement Specification provides the description of the purpose and environment for the software development. Requirement will be obtained in consideration to our target-user groups.

7.1.1 High Level Requirement

The system shall be capable of predicting interest rate of banks with sufficient accuracy thus helping clients with financial decisions. The system shall crawl the news data, obtain sentiment data about relevant keywords (intangible measurement) as our latent features. The system shall also obtain and visualize various different relationships between various financial parameters and their ultimate effect to the interest rate. The system will have two views: Administrator View and Client View. The administrator is responsible for managing and maintaining the server side and its data, while Client View the results of analysis carrier out from the administrator side.

7.1.2 Functional Requirement

Various functional requirements of our projects are listed below:

i. News Mining

The system shall mine the daily news from the news directory openly available in Nepal. The news shall be mined using the information from our knowledge base. The information is used to filter out data irrelevant or invaluable for our system.

ii. Sentiment Analysis

Sentiment Analysis shall be used to extract information from the mined data. These

information shall provide for the missing latent features required by the system, that are unavailable from the numeric financial data. Sentiment Analysis shall also collect major relevant news events for future analysis and visualizations.

iii. Forecasting

The system shall also forecast interest rate based on the previous interest rate trends, also accounting the sentiment strength obtained from news data. Forecasting shall attract client who require help in making crucial financial decision. It has also help understanding the nature of interest rate in the banks of Nepal.

iv. Correlation

With all the influential data collected and obtained, the system shall also be able to correlate them for deeper analysis.

v. Visualization

The system shall provide interactive visualization for current and future interest rate of various banks, and also various factors affecting it using different graphical tools.

vi. Interface

The system shall have two main interfaces:

1. Web Interface and
2. Mobile Interface.

The web interface shall be targeted to demonstrate our system's work view. It shall be portal to let our clients know of your analytic works and prediction models. The mobile interface shall be a secondary interface targeted to a more general people to present our system as a useful tool to easily learn about current interest rates in different banks, user specific functioning such as grouping of banks, notifications and even general interest rate prediction.

7.1.3 Non Functional Requirement

The non-functional requirements of our system are encompassed in the following points:

i. Performance

The system shall have a quick, accurate and reliable results. Evaluation of sentiment of daily news data shall not take no more 3-4 seconds. Predictions should be generated in real time as well.

ii. Capacity and Scalability

The system shall be able to accommodate daily news scraped by our scraper. In case of addition of new news sources, the system should be able to scale accordingly.

iii. Availability

The system shall be available to client at any time. Both the Web and Mobile platform should be up and responsive when ever required.

iv. Recovery

In case of malfunctioning or unavailability of server during crawling, the crawler should be able to recover and prevent data loss or redundancy.

7.2 Feasibility Assessment

7.2.1 Operational Feasibility

Our project is related to data analysis, so it relies mainly in the availability of the data. Without enough availability of data, it is almost impossible to carry out this project. The collection of data is the most challenging part in our application. As for the scope within our Major Project, this application uses raw data available in unstructured format from websites of banks and **ekantipur.com**. The data that we need is mainly the financial data such as FOREX, Gold Price, Petrol Price, GDP, GNI and Interest Rate, and news data. But, these data are hard to collect and more difficult to convince the banking personnel about the nature of data required and sometimes it becomes so difficult that they reject to share even these publicly available data. But, we were able to collect data from various websites of different banks and financial institutions.

Thus, the project is operationally feasible to implement though it is hard to collect the required source data.

7.2.2 Technical Feasibility

The technical feasibility of the project deals with the availability and its actual implement ability
with the existing tools and techniques available in the software market world. The following are
the notable points about the technical feasibility of the project:

- For scraping the numeric data from tables in website, we used Google Spreadsheet and to scrape news data from **ekantipur.com** we used Scrapy, a python based web-scraping framework.
- For maintaining the target tables of the data, we can use the MySQL database.
- For predicting interest rate we used scikit in Python programming language.
- For web-app we have used Django, a python based web-framework and for mobile app we have used Android SDK.

With all these perspectives taken into consideration, the project is technically feasible to implement.

7.2.3 Economic Feasibility

The economic feasibility of the project includes its economic appropriateness with respect to its presented output. If a project provides results of lower significance than the invested budgets, then that project can't be economically viable. In case of our project, we have only used Free and Open Source Software. Similarly, the programming language and libraries we have used are also Free and Open Source. The major cost of the project are cost of data collection and the product development. We also need a server to crawl and store data and deploy our web app. This cost can be covered by selling advertising spaces in our webpage. As this application doesn't have any specific external hardware requirements, so it is not going to be expensive for production.

The potential user of our application are normal people it is, thus, economically feasible for development providing useful financial insights to the clients.

Chapter 8

8. SYSTEM DESIGN

8.1 Overview

The overview of the system can be obtained by the following system diagram. It shows the major components of the system and how they interact to produce results in the system.

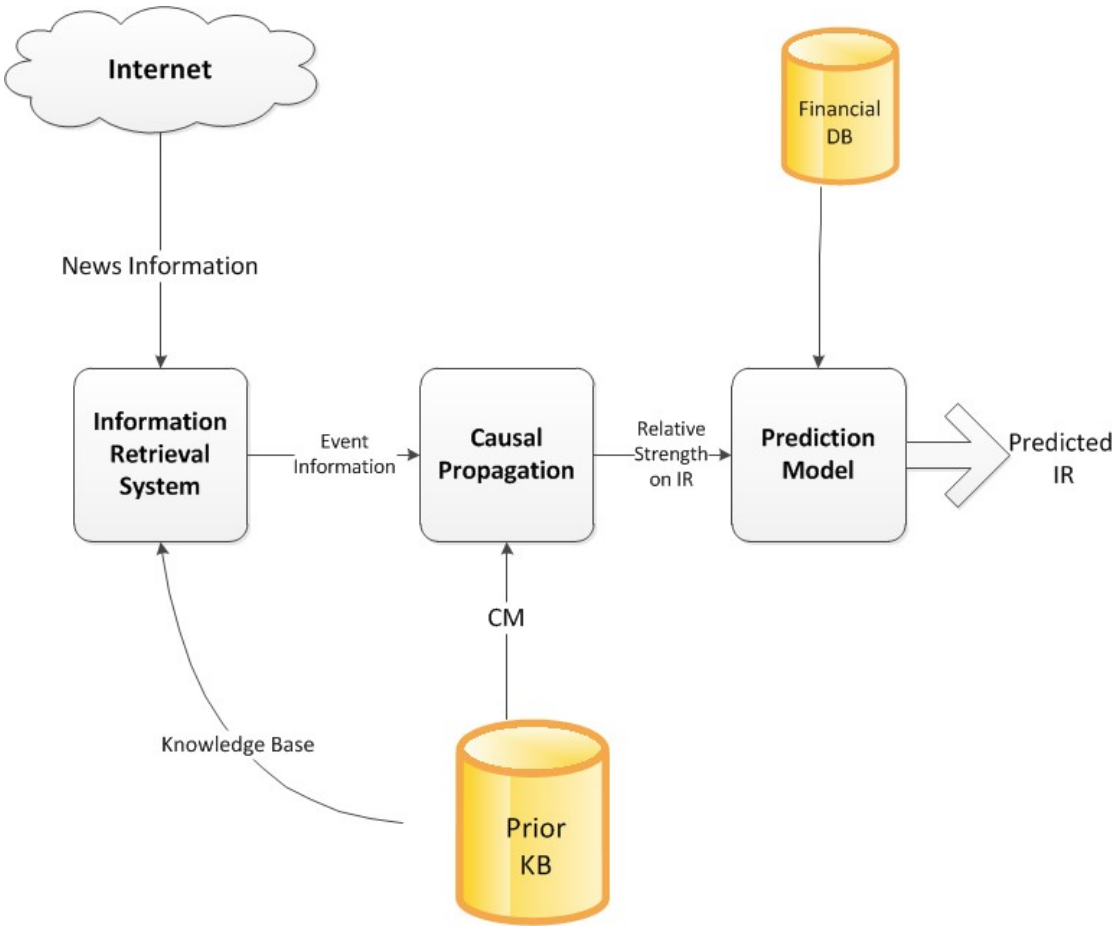

Fig 15: System Block Diagram

8.2 Description of System

8.2.1 Information Retrieval System

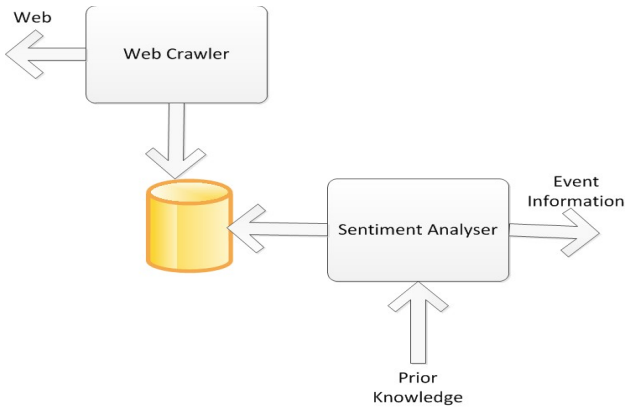

Fig 16: Components of IR System

Information Retrieval System is responsible for extracting news data and evaluating event information from the web. The news data available in the web are extracted and stored in our database in a daily basis. This process is characterized as web scraping in our system. The scraper has been scheduled a crone job in our server that allows the scraper to run on a daily basis. We have used MySQL to store our data.

The stored news data is the accessed by the Sentiment Analyzer. Sentiment Analyzer is responsible for filtering the data, detecting relevant events in the data and finally determining the polarity of the events; events are characterized by keywords in the Prior Knowledge. Prior Knowledge is obtained from our Prior Knowledge Base. Sentiment Analyzer makes use of "Opinion Extraction Technique" for keyword polarity analysis. This returns a vectors of event information where each vector element represents a event.

8.2.2 Prior Knowledge Base

Prior Knowledge Base represents the heart of the system. It represents the knowledge discovered by human intelligences, thus presenting itself (delineating) as a expert of the domain. Prior Knowledge Base is created upon discussion and observation and contains the valuable information about events and their causal relationships within

themselves and the Interest Rate. Prior Knowledge Base not only has the event information and relation, but also search patterns required to detect relevant events for the Sentiment Analyzer.

The ER diagram of your Prior Knowledge Base:

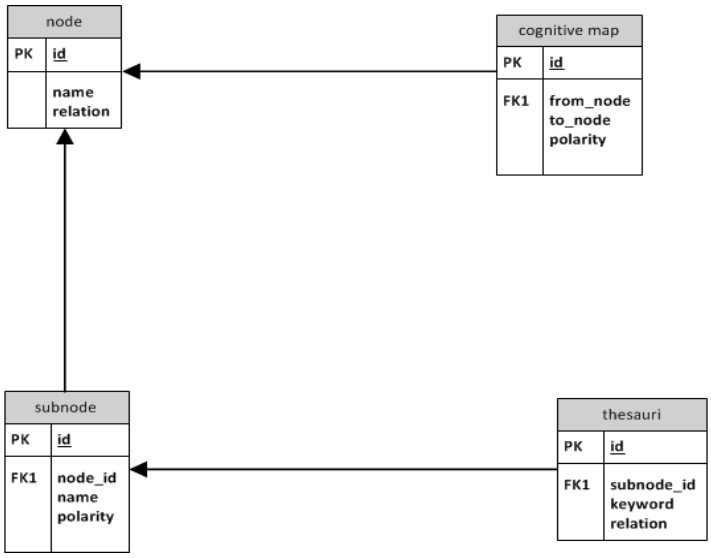

Fig 17: ERD of Prior Knowledge base

8.2.3 Causal Propagation

The event information obtained from IR system needs to be propagated to represent causal inference form other events. The cognitive map represents this information. The causal matrix is thus obtained from the cognitive map and is then used for causal propagation.

This causally propagated even information, represents the ultimate even information which is then subjected to obtain the relative strength. Relative Strength represents a value signifying the overall effect of event information vector to the Interest Rate. If the Relative Strength is greater than 0.5, it signifies that then overall effect of events is positive towards the Interest Rate, and if it is lesser than 0.5 it signifies otherwise.

8.2.4 Prediction Model

The prediction model uses the Quadradic Regression model characterized by the equation

$$y = c + \sum (a_i * F_i) + \sum (b_i * F_i^2)$$

The prediction model uses financial data along with the relative strength as features for its model. The models trains its self with the financial data, relative strength and the past interest rate.

8.3 Component Diagram

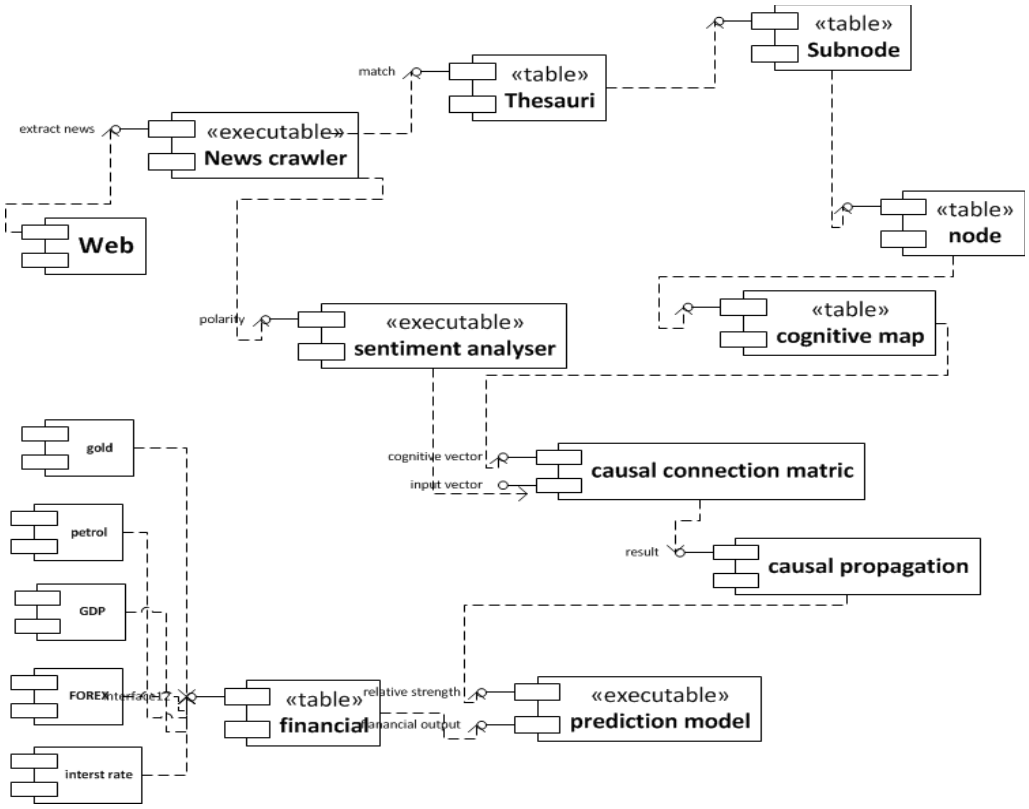

Fig 18: Component Diagram of the system

The diagram above shows the most basic components of the system. The three tables "node", "sub node" and "thesauri" are related to each other. The table "node" represents all the relevant events affecting the interest rate. The "sub node" represents a more atomic category. The "sub node" is derived from the "node". The "thesauri" contains the list of keywords corresponding to an event in the "sub node". The IR system utilizes this "thesauri" for extraction and Cognitive Map is created out of the information from the "node" table.

8.4 Activity Diagram

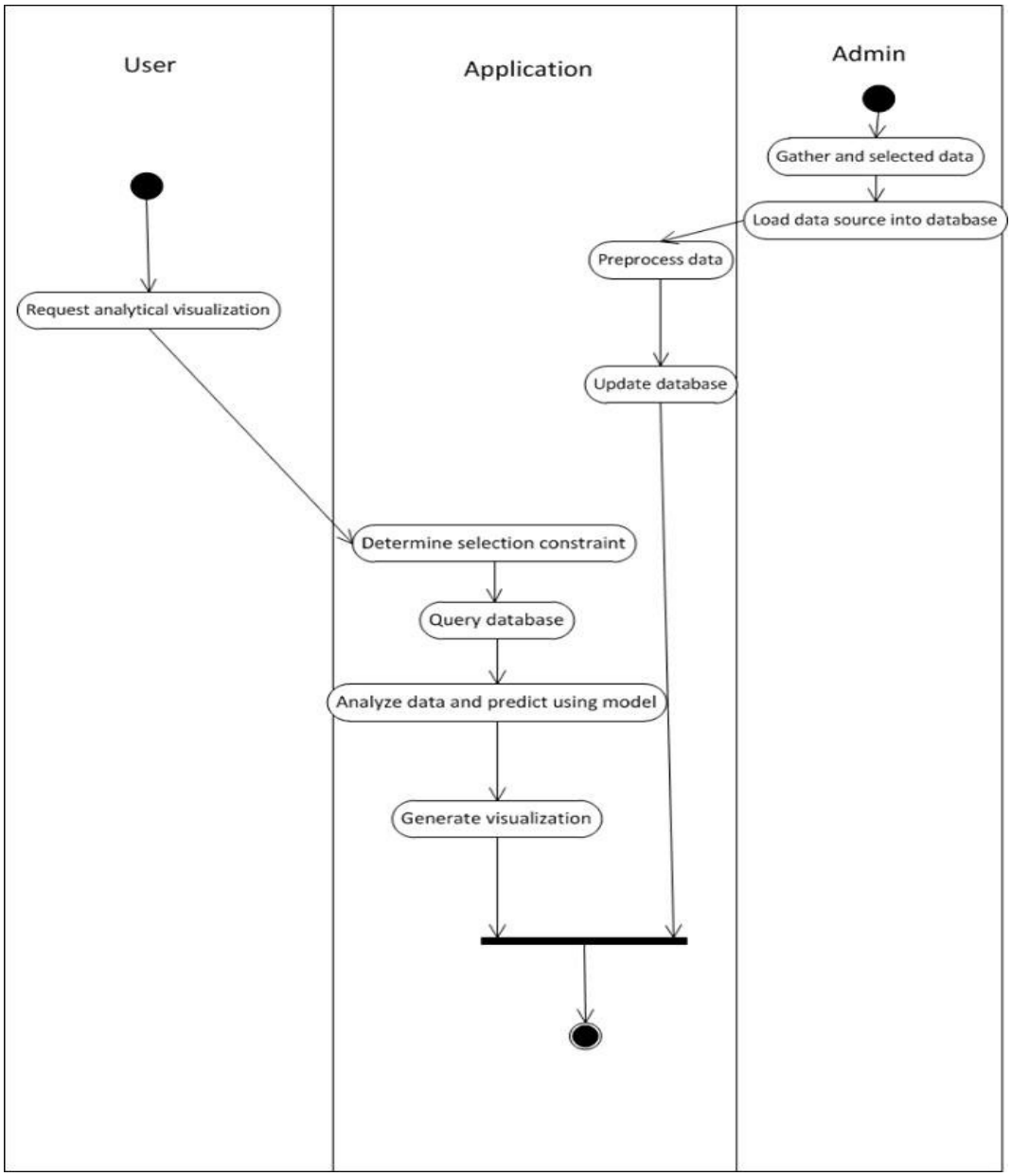

Fig 19: Activity Diagram of the system

The diagram above show the activity diagram of the system. It shows how three main activities, “User”, “Application” and “Admin”, interact. The “User” initiates action requesting visualization and results, the “Application” responds it by generating the visualization.

8.5 Use Case Diagram

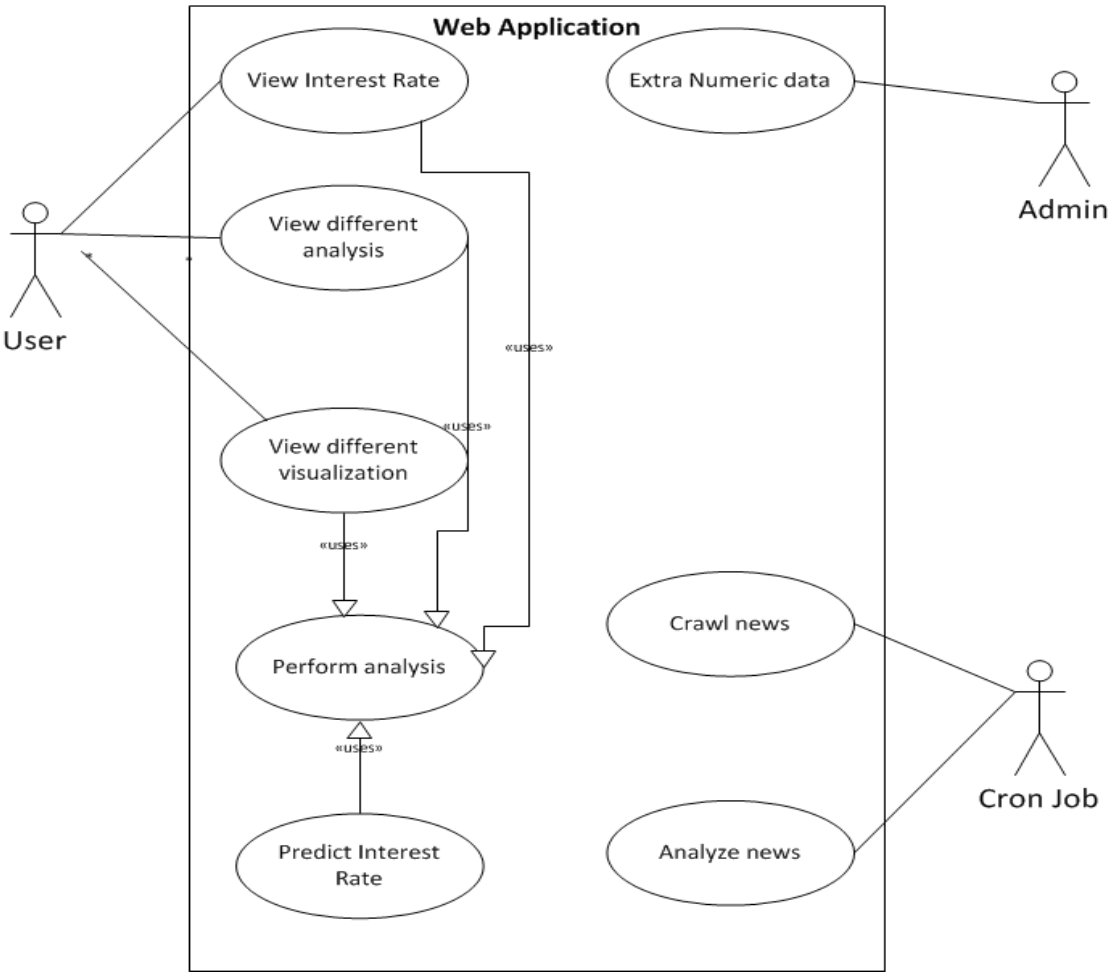

Fig 20: Use case diagram of the system

The above use case diagram show three main actors, “user”, ”admin” and the “cron job”. These are the three main actors interacting the system. The “user” analysis and view the results. The “admin” prepares data and the “cron job” initiates the crawler for daily news crawl and also triggers the IR system in action.

Chapter 9

9. IMPLEMENTATION

9.1 Overall System Work-flow

The overall system workflow is defined in two section below:

9.1.1 Flow chart

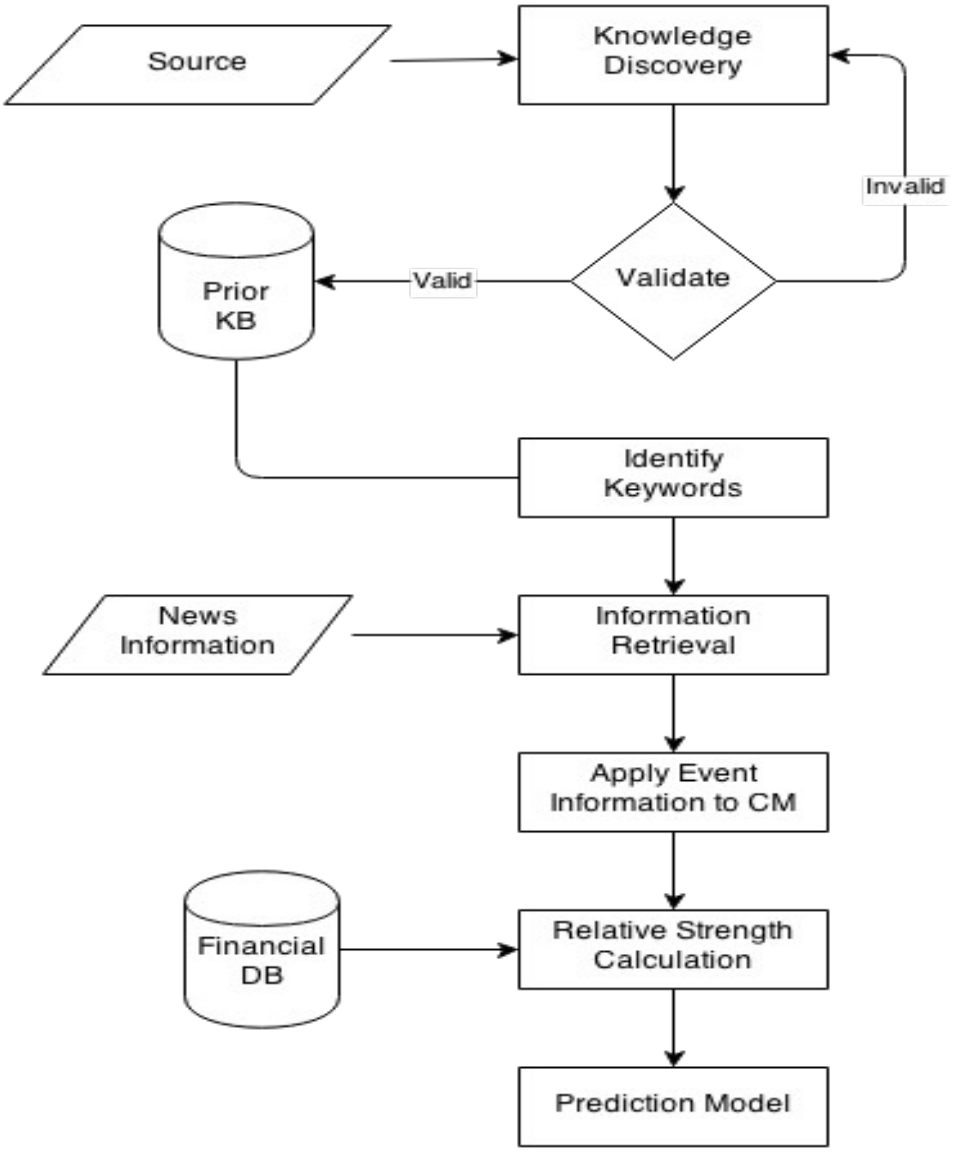

Fig 21: Flow Chart Diagram of the system

9.1.2 Workflow

The system begins with the description of knowledge in the Prior Knowledge Base. This process is iterative, where the knowledge to be represented in the Prior Knowledge Base must be chosen carefully. This knowledge represents the brain of the system, upon which the system rests. Upon discussion with the mentors, and few other personnel for the prior knowledge base, a list of events affecting the interest rate was identified. Continuing with deeper analysis, the intra-component relations were identified. This would be our primary knowledge base. The knowledge base was further refined by adding appropriate synonyms for the keyword to increase the chances of being noticed during the IR process. Prior knowledge is built by using CMs of specific domains as its primary source of solving problems in that domain.

The IR System is used to retrieve news information on the Internet by drawing on prior knowledge. The results of the retrieved information are applied to CMs. Knowledge Application Systems apply the retrieved event information to CMs and perform the causal propagation with a causal connection matrix. The final result of the causal propagation is input into a prediction model as positive or negative information along with other financial variables.

9.2 Data Collection Implementation

Two different types of data were collected for this project viz. Numeric Data and News Data.

9.2.1 Numeric Data

Google Spreadsheet was used to collect Numeric Data from website. In the spreadsheet IMPORTHTML formula was used to import the table into excel-sheet.

```
IMPORTHTML

Imports data from a table or list within an HTML page.

Syntax: IMPORTHTML(url, query, index)

url: The URL of the page to examine, including protocol (e.g.
```

```
http://).

query: Either "list" or "table" depending on what type of
structure contains the desired data.

index: The index, starting at 1, which identifies which table
or list as defined in the HTML source should be returned.
```

Example:

```
IMPORTHTML("http://en.wikipedia.org/wiki/Demographics_of_India"
,"table", 4)
```

9.2.2 News Data

Scrapy, a python based web and screen scraping tool was used to collect data from ekantipur.com.

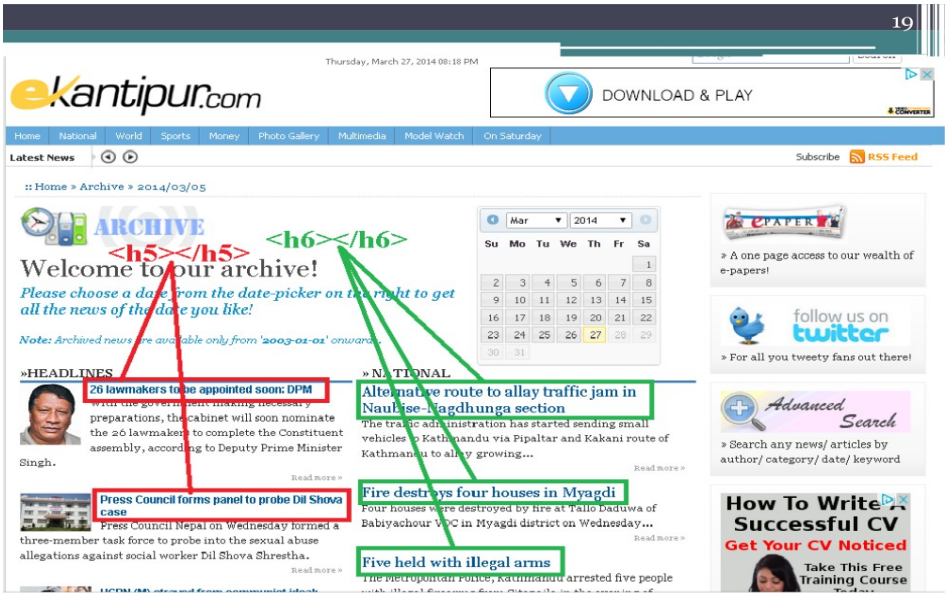

Fig 22: Screenshot of ekantipur.com

Following steps were followed to scrape the data from the website:

1. A site was picked

2. The data we wanted to scrape was defined using Scrapy Items.

3. A spider was then written to extract the data. It is written in spider.py file that resides inside the spider folder.

4. Finally the spider was run to extract the data.

Fig 23: Screenshot of page containing news

The news heading were in between header <h5></h5> and header <h6></h6>. The title of the news was extracted from the header. So, the links that were between <h5></h5> and <h6></h6> were traversed. When the links were traversed, we reached to a page as shown below.

In this page containing news, the main news was between the tag

```
<div class="main_wrapper_left"></div>
```

The main news was extracted from here and then stored to database.

title	news	date
Over dozens of protestors detained, two seriously injured	KATHMANDU, FEB 01 - Police intervention in a protest rally organized by five political parties against "regression" has left over scores of demonstrators injured including two sustaining serious injuries. The five agitating political parties as per their scheduled program displayed black flags as part of protest in the capital Saturday afternoon. The demonstrators chanting slogans rallied through Ratnapark, Bhotahiti, Naradevi and converged to a mass meeting at Bhotahitichowk. Addressing the mass meeting, Leelamani Pokhrel of People's Front Nepal said," we have come out without caring our lives by wrapping shrouds on our heads so the movement against regression will not be stopped." Shortly after Pokhrel's address, police intervened in the meeting and took hold of five protesters including Pokhrel himself. Then the police baton charged the protestors indiscriminately to disperse the mob of protestors. Meanwhile, two students were left seriously injured in the scuffle between the police and students. According to Gagan Thapa, General Secretary of Nepal Students Union (NSU) police without any reason charged batons at the students and broke a leg of Lakshman Dhakal and a hand of Chudamani Sapkota both of them being the cadres of NSU. Likewise, police detained over 18 students from a similar program in Banepa Saturday. (snn)Posted on: 2004-02-01 01:12 AddThis Button BEGIN	2004/02/01
Nepal-India border talks inconclusive	KATHMANDU, FEB 01 - The Nepal-India talks on border management ended inconclusively Saturday after both the sides failed to come to any agreement on the modality of the treaty on the extradition of wanted persons. Sources said the talks were dogged due to the refusal by the Nepali side to agree to the Indian request to extradite any third country individuals to India. In fact, negotiations on the extradition and mutual legal assistance framework failed to make any headway much like what happened nearly two years ago. However, both the sides have expressed commitment to chase the agenda further by holding talks on the issue within the next six months. Other highlights of the meeting include that both the sides reiterated their commitment to not to allow their territories to be used by terrorists who may be harming the interests of the other. Asked as to what delays the treaty, he said that differences still exist, although he denied to spell out the differences. â€œIt is a process and is going on,â€ K S Ramasuban, leader of Indian delegation, said. However Joint Secretary at the Ministry of Home, Umesh Mainali, said the treaty could not be signed because the issues involved are yet to be debated by both the sides. (snn)Posted on: 2004-02-02 03:07	2004/02/01

Fig 24: Screenshot of news data in table

9.3 Implementation of Knowledge Base

Our system is designed to search and retrieve news information from the Internet utilizing the cognitive map and prior knowledge base. The proper sentiment analysis of the extracted keywords defines how the news information affects interest rate which serves as the insight on the interest rate movements. The values of sentiment analysis are applied to the prediction model in order to achieve more accurate interest rate prediction.

The prior knowledge base in our system is built using Cognitive Map. Cognitive Map is built mainly in two phases.

In the first phase we defined the concept nodes, whose movements could change the interest rate values and whose information could be found via news data. Different theories like Classical Theory, Loanable fund theory, Liquidity preference etc. were reviewed to determine the relevant nodes of CM. News information were also mined in order to determine basic categories of the information that is found in news information and to find what information could be extracted whose affect is significant to the interest rate. In addition to this, news was also mined in order to find the frequent keyword list.

In the second phase the causal relationship between these nodes were found by the knowledge of domain experts. The polarity of each node was found by analyzing how each node affects the other and finally the interest rate. There could be three types of polarity:

i. Either a node affects another node positively (+1) i.e. increase in value of one node increases the value of other node.

ii. Or, a node affects another node negatively (-1) i.e. increase in value of node decreases the value of another node.

iii. Else a node has no affect on the other (0) i.e. the increase or decrease in the value of one does not affect the value of the other node.

Only polarity was used between the nodes, no any weight information was used in order to avoid the biasness.

9.3.1 Cognitive Map of the System

Final nodes of the cognitive map derived through the above process are:

1) Interest Rate
2) Economic Factor
3) Political Factor
4) Social Factor
5) Bank's Information
6) International issues
7) Infrastructure and Development
8) Market

The concept nodes and the causal relationship among them is shown in the **Fig 2:

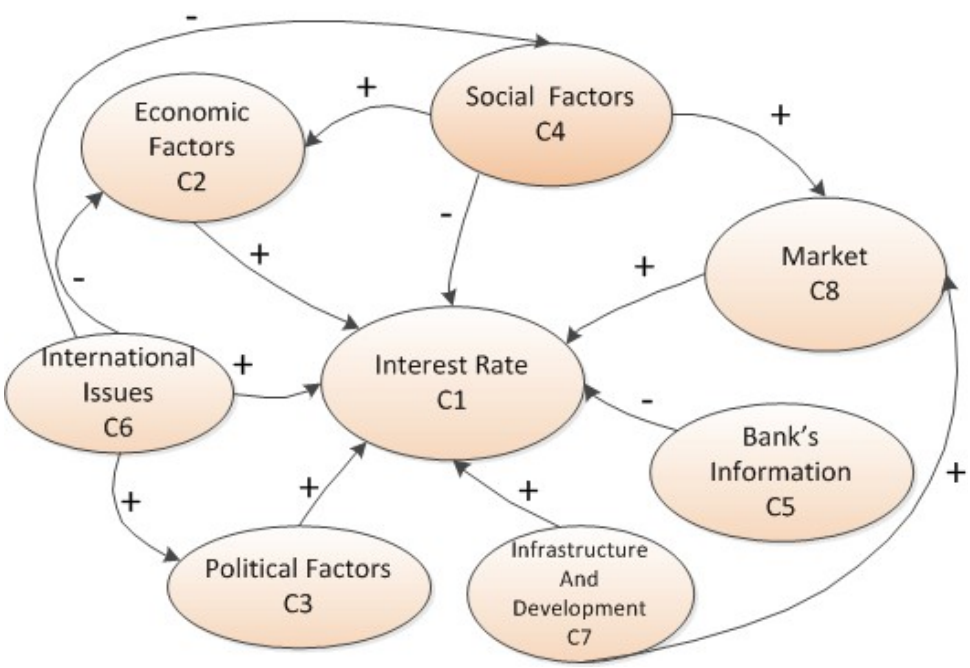

Fig 25: Cognitive Map of the system

These eight nodes form the initial and main layer of the cognitive map. The second layer is the division of the initial layer(first layer node) into more atomic nodes. For example, social factor is further divided into natural disaster, unemployment, diseases, education, agriculture etc. Each node of the first layer has its own division into simpler form which makes up the second layer. Division of each node is shown below:

Economic Factors	Political Factors	Social Factors	Bank's Information
Transaction	Political Disputes	Agriculture	Rivalry between banks
Remittance	Legal Issues	Education	Spread rate of banks
Subsidy	Political Awareness	Health	Policy of NRB
Currency	Policy for IR	Alternative energy	Bank net profit
Depression	Budget	Tourism	
Prosperity		Natural resources	
Recession		Disaster	
Demand of Funding		Epidemics	
Supply of Funding		Smuggling	
Inflation		Black Market	
Deflation		Corruption	
		Crime	
		Trafficking	
		Unemployment	
		Income	
		Availability of natural resources	

International Issues	Infrastructure and Development	Market
International Disputes	Construction	Price of product
Boundary Disputes	Urbanization	Consumption
Foreign aid	Industrialization	Land Transaction
Indo-Nepal relationship	Living Standard	Read Estate
China Nepal relationship		

Table 6: Second Layer of CM

Thus, the second layer was form with list of simpler divisions from the first layer node. The third and the final layer consist of all the possible relevant synonym list of the keywords from second layer. This synonym list forms the thesauri of our system. For example the sub-node unemployment is broken down into a set of synonyms like {unemployed, employed, job, vacancy, etc.} This thesauri is compared to the extracted news text in order to find information according to the meaning of the concept node and we regard those text as having the same meaning as those of the concept node.

9.3.2 Causal Connection Matrix

We consider 8-by-8 causal connection matrix M such that there is one row and one column for each node of the Cognitive Map as shown in Figure 2. Our Causal Connection Matrix looks as shown in Figure 3 below:

	C1	C2	C3	C4	C5	C6	C7	C8
C1	0	0	0	0	0	0	0	0
C2	1	0	0	0	0	0	0	0
C3	1	0	0	0	0	0	0	0
C4	-1	1	0	0	0	0	0	1
C5	-1	0	0	0	0	0	0	0
C6	-1	-1	1	-1	0	0	0	0
C7	1	0	0	0	0	0	0	1
C8	1	0	0	0	0	0	0	0

Fig 26: Causal Connection Matrix

With this causal connection matrix, we can apply to the causal connection propagation, event information which is gathered by the information retrieval system. Let us consider the information retrieval result news information of a certain date has the input vector D, where D is [1 0 0 -1 0 0 0 0]. This vector signifies that in the news information retrieved, C1 and C8 node are being positively affected (or, C1 and C8 node are increasing), C4 node is decreasing and other nodes, C2, C3, C5, C6, C7 is not being affected.

The input vector D is multiplied by the causal connection matrix to obtain the output vector matrix [1 -1 0 0 0 0 0 -1].

Finally the positive and negative event information are converted into the relative

strength of effects on interest rate. If the relative strength, $Ek_t = Pk_t /(Pk_t + Nk_t)$ is over 0.5 then it can be stated as the positive effect on the interest rate is stronger than the negative interest rate. If the relative strength is under 0.5 then the negative effect on the interest rate is stronger than the positive effect. This relative strength is put into the neural network to find meaningful results.

9.4 Implementation of Prediction Model

For prediction three models were selected namely Simple Regression Model, Quadratic Model and Decision Tree Regressors. In our application, there will be an option for user to allow to select any of the above mentioned model. Then the model will predict the value and graphical representation will be show. The overall procedure of implementation of prediction model is explained in the next paragraph.

The features i.e GDP, GNI, Gold Price, Petrol Price, FOREX along with sentiment strength were collected and saved in a central repository. These values were retrieved and preprocessed so as to remove any noise, irregularity and fill missing values. Once preprocessing was done, it was stored into database. Before we can use a model to predict interest rate, it has to be trained using past data. The candidate models were trained with past 100 data out of available 108 data. Then we did k-fold cross validation so as to chose the best model out of the candidate models and we found that Simple Regression model, Quadratic model and Decision Tree Regressors were the good prediction models. Once the k-fold cross validation is done, we trained each of the selected algorithm to use for prediction. As mentioned in results we used each of these models to predict interest rate. In this process we used 100 data our of available to train and used remaining 8 to test. This test showed that decision tree gives the best fit of the nature of data and Simple Regression Model and Quadratic Model also did reasonable job on the data we trained. But one testing it found that, Decision tree is the suitable model of prediction. SVM with rbf kernel also performed well for prediction.

Hence we will be using Decision Tree and SVM with rbf kernel as main models of prediction.

9.5 Implementation of Interface

9.5.1 Implementation of Web Interface

The web interface is the primary interface of our system that is used for visualizing and representing our system's output. The web interface runs on a django-backend server utilizing JavaScript along with different visualization libraries. AJAX was used to communicate data between Client – Server for visualization.

9.5.2 Implementation of Mobile Interface

The mobile interface is used as a secondary interface of our system. Mobile Interface was presented as tool for interest rate information and analysis to a more general users. The mobile interface was developed using android development tool-kit for android OS devices.

The ER diagram of the database used our mobile system is:

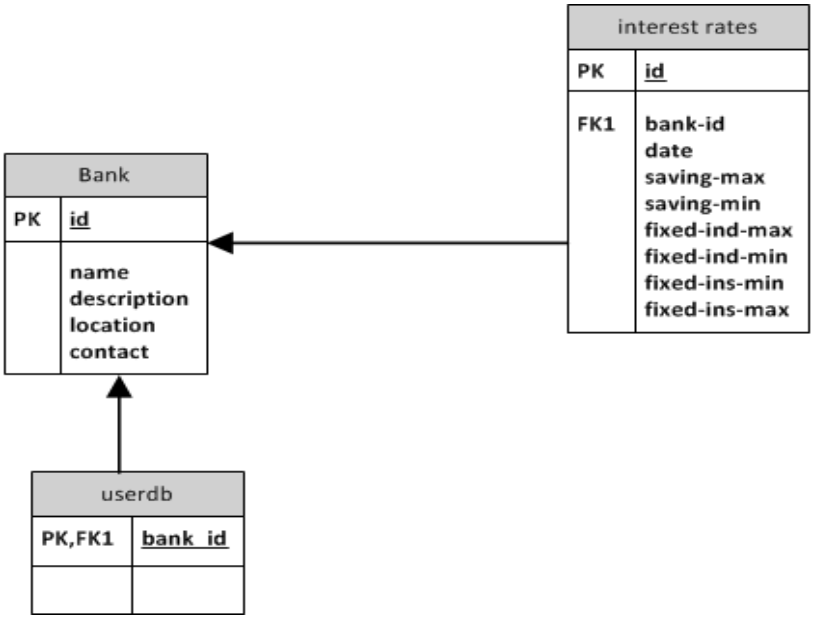

Fig 27: ERD of mobile database

The data exchange with the main server is done via API call to the main web-server. The server returns the data to the interface in JSON format. The JSON data is parsed and is used to update the mobile database. The data interaction occurs only for

consistency. The consistency of the databases is checked every time during the application startup.

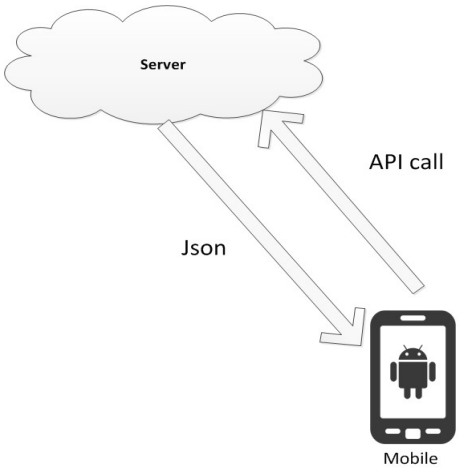

Fig 28: Communication of Server and Mobile Device

The JSON format specification for the communication can be represented as below:

```
{
      'id':001_ ,
      'bank':_ ,
      'description':_ ,
      'info':_ ,
      'location':_ ,
      'contact':_ ,
      'interest rates':[
      { 'id':_, 'bank id':_, 'date':_ },
      { 'id':_, 'bank id':_, 'date':_ },
      ...
      { 'id':_, 'bank id':_, 'date':_ },
      ]
}
```

Chapter 10

10. TOOLS AND TECHNOLOGIES

10.1 Python

Python was used as our main programming language. Several Python based libraries were used to complete our project.

Python is a widely used general-purpose, high-level programming language. Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of code than would be possible in languages such as C. The language provides constructs intended to enable clear programs on both a small and large scale. Python supports multiple programming paradigms, including object-oriented, imperative and functional programming or procedural styles. It features a dynamic type system and automatic memory management and has a large and comprehensive standard library.

10.2 Django

Django was used to develop our web app.

Django is a free and open source web application framework, written in Python, which follows the model–view–controller architectural pattern. It is maintained by the Django Software Foundation (DSF), an independent organization established as a non-profit.

Django's primary goal is to ease the creation of complex, database-driven websites. Django emphasizes reusability and "pluggability" of components, rapid development, and the principle of don't repeat yourself. Python is used throughout, even for settings, files, and data models. Django also provides an optional administrative create, read, update and delete interface that is generated dynamically through introspection and configured via admin models.

Some well known sites that use Django include_Pinterest, Instagram, Mozilla, The Washington Times, Disqus, and the Public Broadcasting Service.

10.3 Scrapy

Scrapy, a scraping framework was used to scrape the news from **ekantipur.com.**

Scrapy is a fast high-level screen scraping and web crawling framework, used to crawl websites and extract structured data from their pages. It can be used for a wide range of purposes, from data mining to monitoring and automated testing.

10.4 NLTK Library

NLTK(Natural Language ToolKit) is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.

This library was used for processing and analyzing the raw news data and extracting the strength of keyword that affects the interest rate.

10.6 scikit-learn

scikit-learn is an_open source machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, logistic regression, naive Bayes, random forests, gradient boosting, *k*-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

This libaray was used to create and train our prediction model.

10.7 Android SDK

The Android_software development kit (SDK) includes a comprehensive set of development tools. These include a debugger, libraries, a handset emulator based on QEMU, documentation, sample code, and tutorials. Currently supported development platforms include computers running Linux (any modern desktop Linux distribution), Mac OS X 10.5.8 or later, and Windows XP or later. For the moment one can also develop Android software on Android itself by using the AIDE - Android IDE - Java,

C++ app and the Java editor app. The officially supported integrated development environment (IDE) is Eclipse using the Android Development Tools (ADT) Plugin, though IntelliJ IDEA IDE (all editions) fully supports Android development out of the box, and NetBeans IDE also supports Android development via a plugin. Additionally, developers may use any text editor to edit Java and XML files, then use command line tools (Java Development Kit and Apache Ant are required) to create, build and debug Android applications as well as control attached Android devices (e.g., triggering a reboot, installing software package(s) remotely).
Android SDK was used to develop mobile app for Android Devices.

10.8 MySQL

MySQL is the world's second most widely used open-source_relational database management system (RDBMS). MySQL is a popular choice of database for use in web applications, and is a central component of the widely used LAMP open source web application software stack (and other 'AMP' stacks). LAMP is an acronym for "Linux, Apache, MySQL, Perl/PHP/Python." Free-software-open source projects that require a full-featured database management system often use MySQL.
Our database system was based on MySQL database.

10.9 HIghcharts

Highcharts is a charting library written in pure JavaScript, offering an easy way of adding interactive charts to your website or web application. Highcharts currently supports line, spline, area, areaspline, column, bar, pie, scatter, angular gauges, area range, area spline range, column range, bubble, box plot, error bars, funnel, waterfall and polar chart types. Highcharts was used to create interactive visualizations of data.

10.10 D3.js

D3.js (or just D3 for Data-Driven Documents) is a JavaScript library that uses digital data to drive the creation and control of dynamic and interactive graphical forms which run in web browsers. It is a tool for data visualization in W3C-compliant

computing, making use of the widely implemented Scalable Vector Graphics (SVG), JavaScript, HTML5, and Cascading Style Sheets (CSS3) standards. It is the successor to the earlier Protovis framework. In contrast to many other libraries, D3 allows great control over the final visual result. D3.js was used to create much advanced visualizations.

10.11 Git

Git is a distributed revision control and source code management (SCM) system with an emphasis on speed, data integrity, and support for distributed, non-linear workflows. Git was initially designed and developed by Linus Torvalds for Linux kernel development in 2005, and has since become the most widely adopted version control system for software development.

As with most other distributed revision control systems, and unlike most client–server systems, every Git working directory is a full-fledged repository with complete history and full version-tracking capabilities, independent of network access or a central server. Like the Linux kernel, Git is free software distributed under the terms of the GNU General Public License version 2.

Git was used to collabrate between us, the project members and to control version in the server.

Chapter 11

11. RESULTS

11.1 Prediction

11.1.2 Prediction Using Decision Tree

The first model used for prediction was decision tree. Out of available data, a few of them were used for testing and rest were used to train the model. The graph of prediction using Decision Tree Regressors is given below.

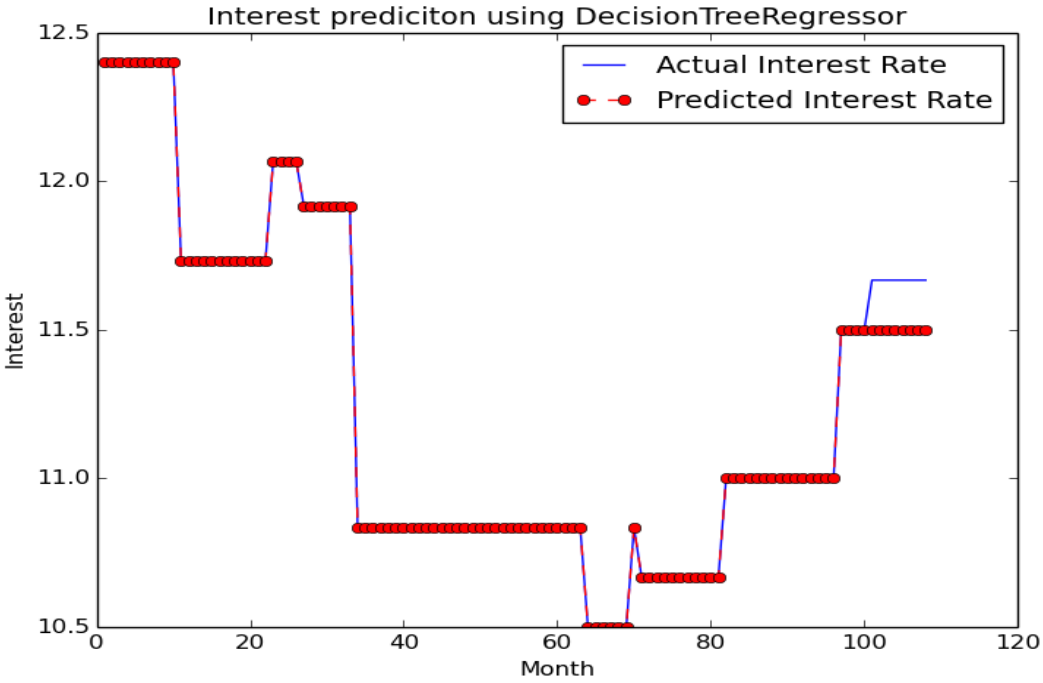

Fig 29: Interest rate prediction using Decision tree

The interest rate after year 2010-4 tends to stay at 11.5 as predicted by decision tree regressors whereas the true interest rate after year 2010-4 tends to stay at 11.66 and thus has an error of 1.37 %.

11.1.2 Prediction Using Quadratic Model

The quadratic model being one of the selected model for prediction was also used. The graph of prediction performed by the Quadratic Model is given below.

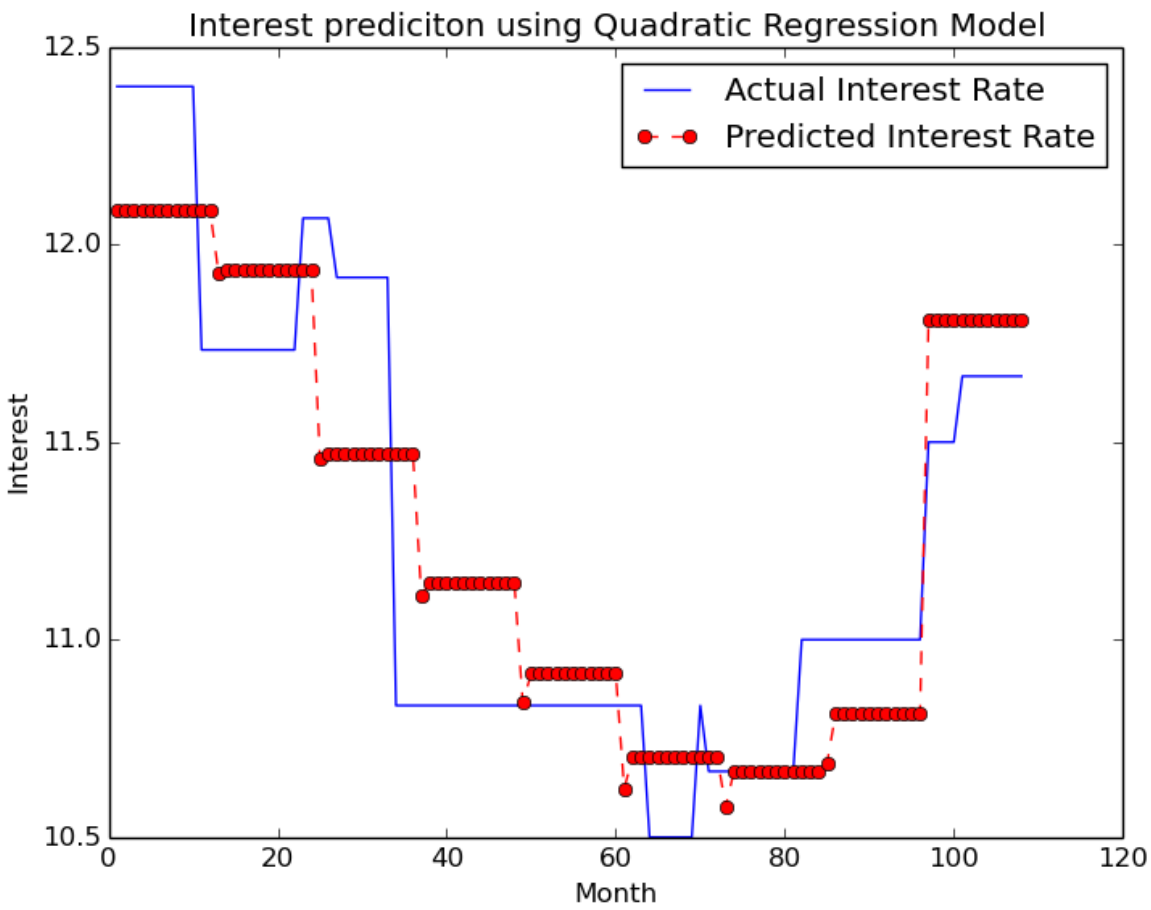

Fig 30: Prediction of interest rate using Quadratic model

While using this model for the predication of interest rate for date after 2010-4, it is found that the model predicts a value of 11.8 % whereas the true value interest rate at the time being was 11.66 and thus there is an error of about 1.72%.

11.1.3 Prediction Using Simple Regression model

The Simple Regression model fitted for multiple features can also be used for the prediction. The graph given below shows how well the model represents the dataset.

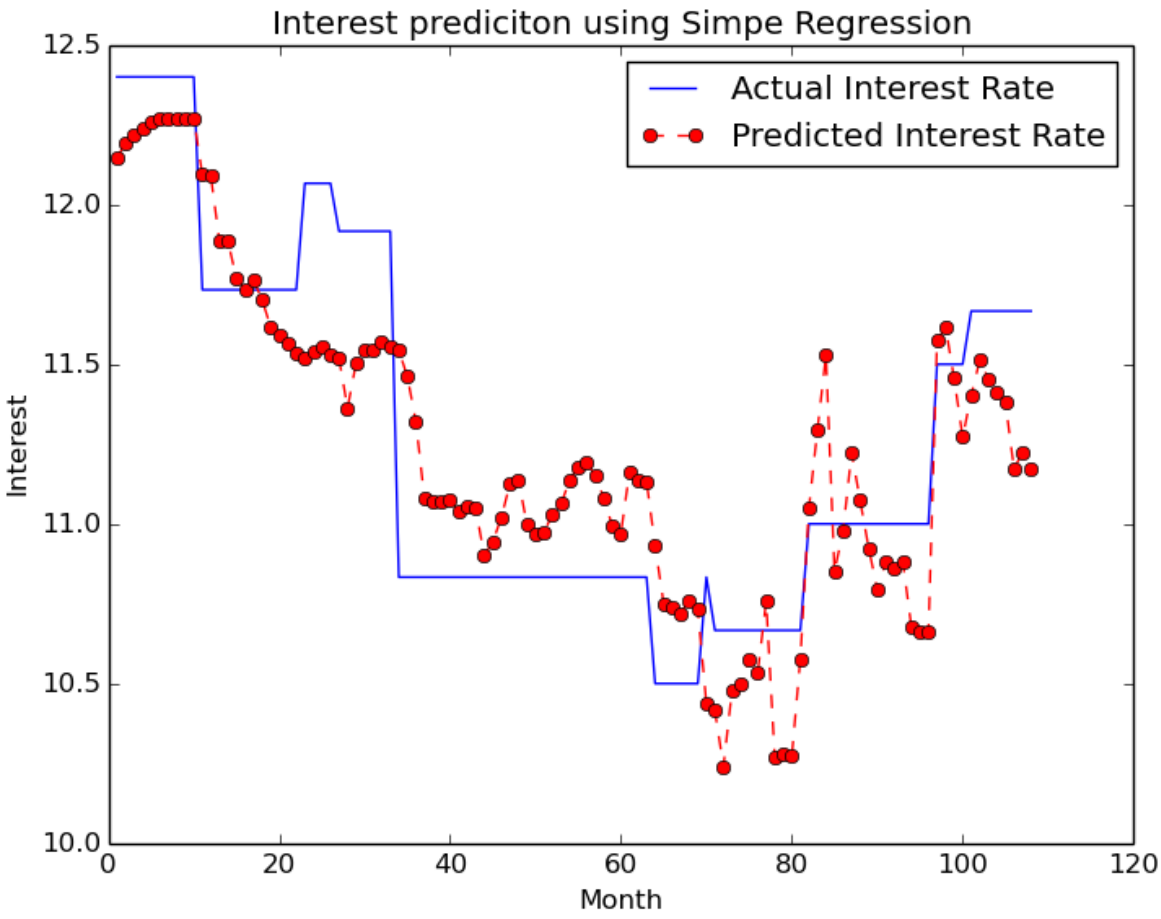

Fig 31: Interest rate prediction using regression model

This model seems to be good approximation of the trend of interest rate which is considered to be depending upon various features like GDP, GNI, petrol price, gold price, FOREX and sentiment value. The true value of interest rate after data 2010-4 tends to stay near 11.66 and the models shows values ranging from 11.5 % to 11.2%. The prediction on long run seems to be accurate as the interest rate tends to decrease over time due to various factors but still there are some hidden fluctuations as well. This model has an error of about 3.08%.

11.1.4 Prediction Using Moving average

Moving average can be used to predict future values as it assumes a constant moving average in future. The variable that can be changed in moving average method in window length. Window length is the number of data points which is averaged and slided through each and every point to calculate corresponding moving average value.

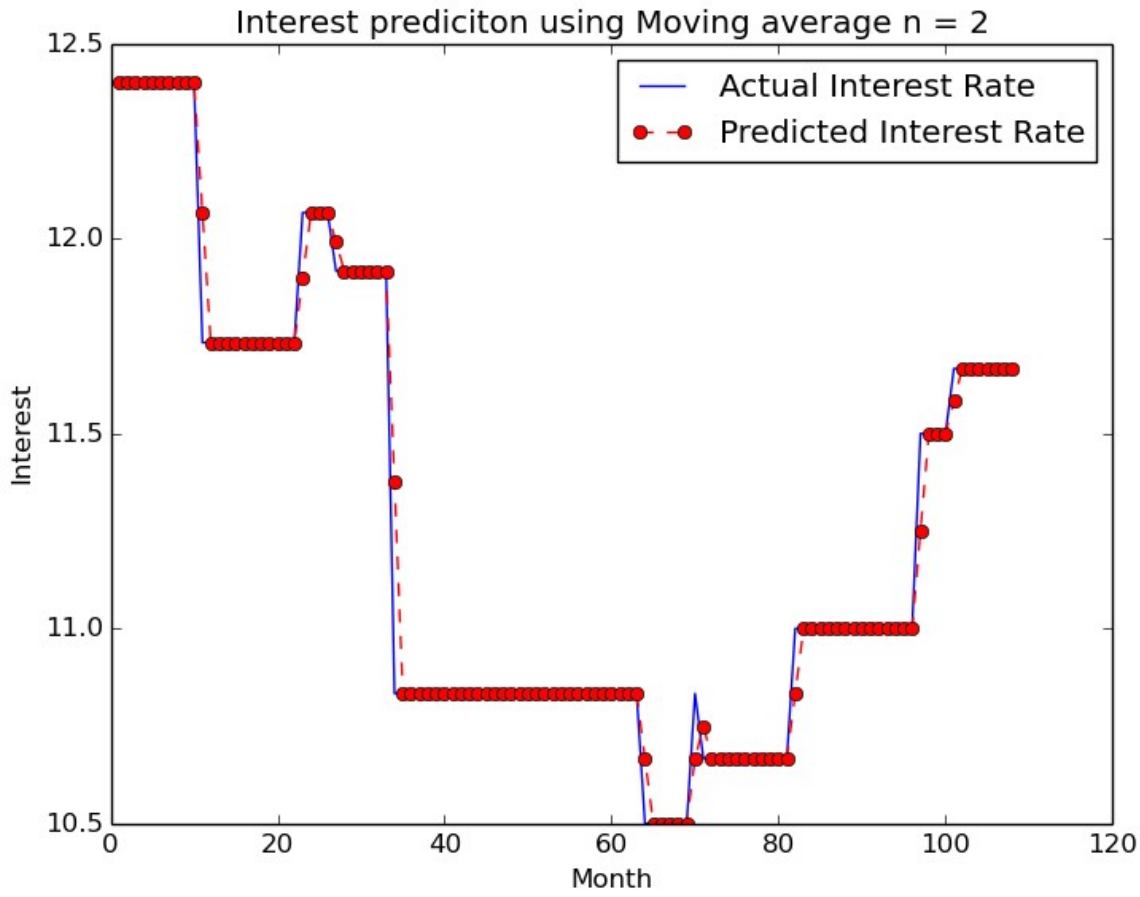

Fig 32: Prediction of interest rate using moving average with n =2

This model has not sufficient property to predict interest rate into more than one time frame in future. Even if it is done, there will be almost no variation and the trend of the interest rate will remain the same or retrace the moving average path. The prediction of interest rate after year 2010-4 is between 11.5 to 11.66 which is very much near to actual values.

11.1.5 Prediction Using Exponential Smoothing

Exponential Smoothing is another technique that can predict a value a head into future by smoothing the data considering the nearest data point has more effect than the data points at large distance. Using exponential weights, a similar approach as moving average is done to achieve exponential smoothed values.

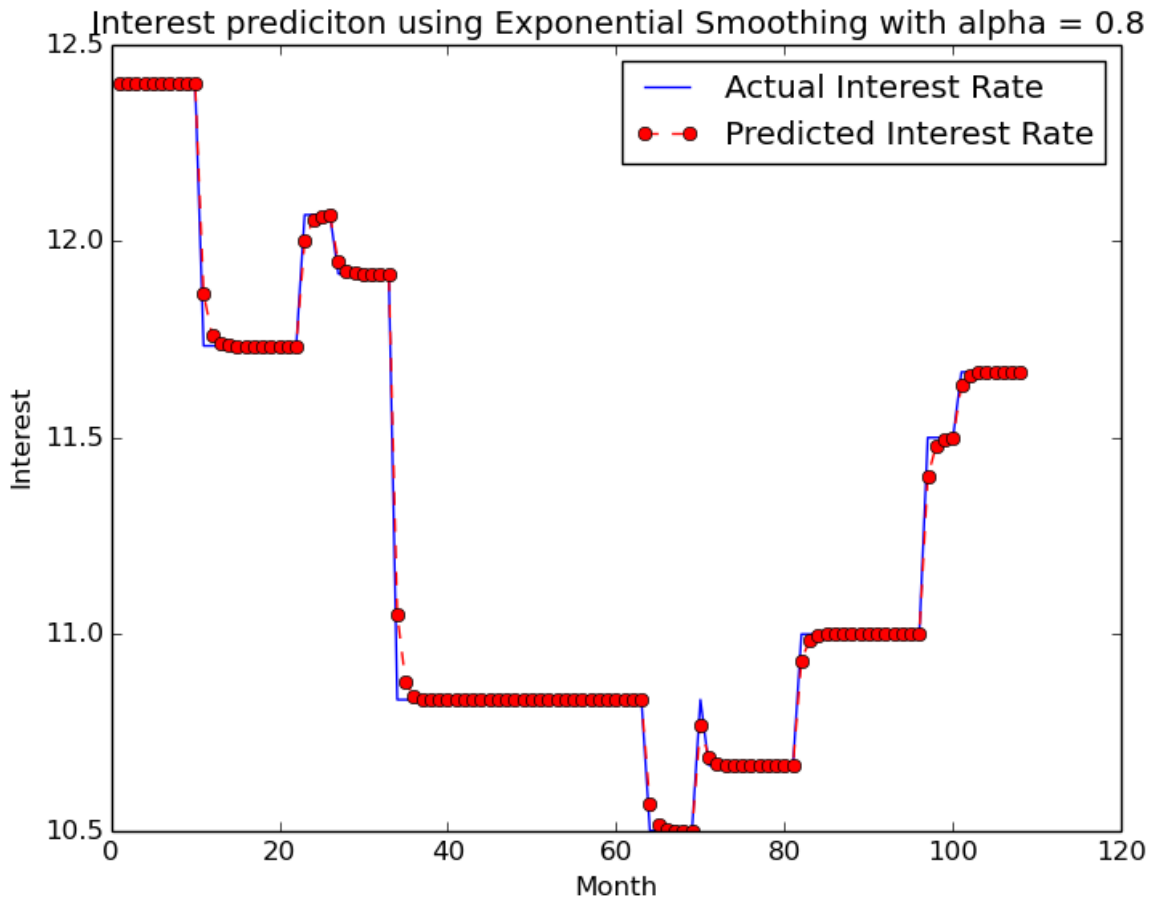

Fig 33: Interest Rate predication using exponential smoothing

This model almost overlaps with actual data points and predicts well. The value of alpha which is smoothing factor is chosen to 0.8 despite it can range between 0 and 1. The above graph shows the interest rate after time of 2010-4 remains constant at 11.66.

11.2 Visualization

11.2.1 Variation of Moving Average with Window Length N

With the change in window length the nature of moving average method varies. So the effect of change in n on moving average is visualized below. These graphs shows prediction model using moving average for n =2,3,4

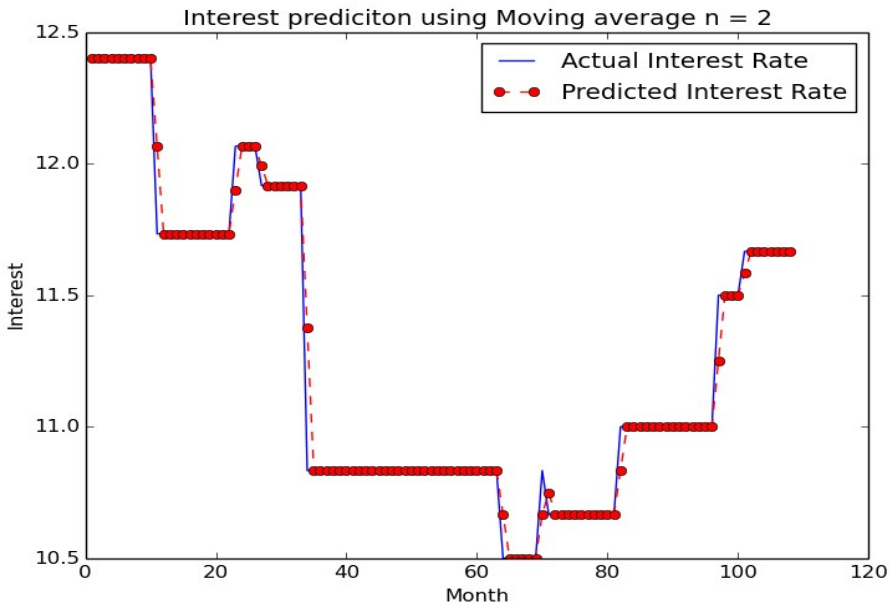

Fig 34: Moving average when window length is N = 2

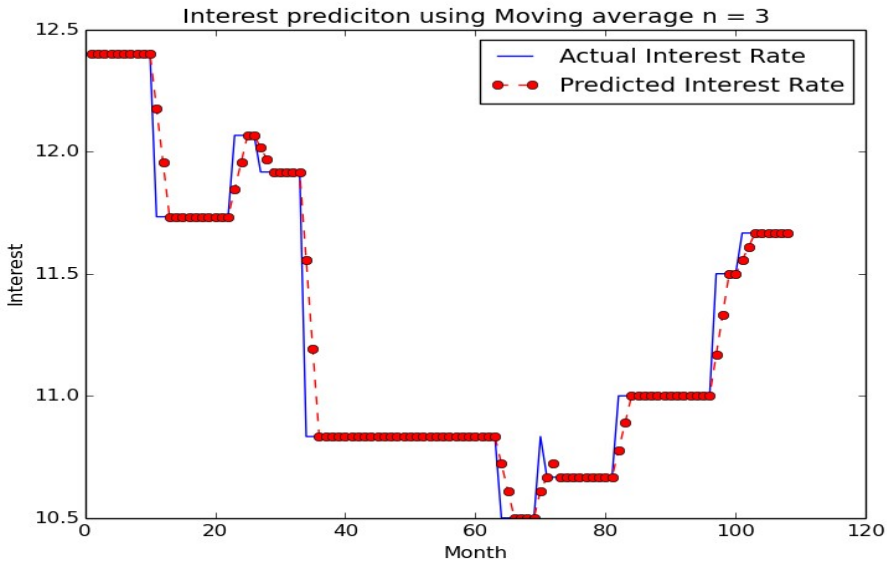

Fig 35: Moving average when window length is N = 3

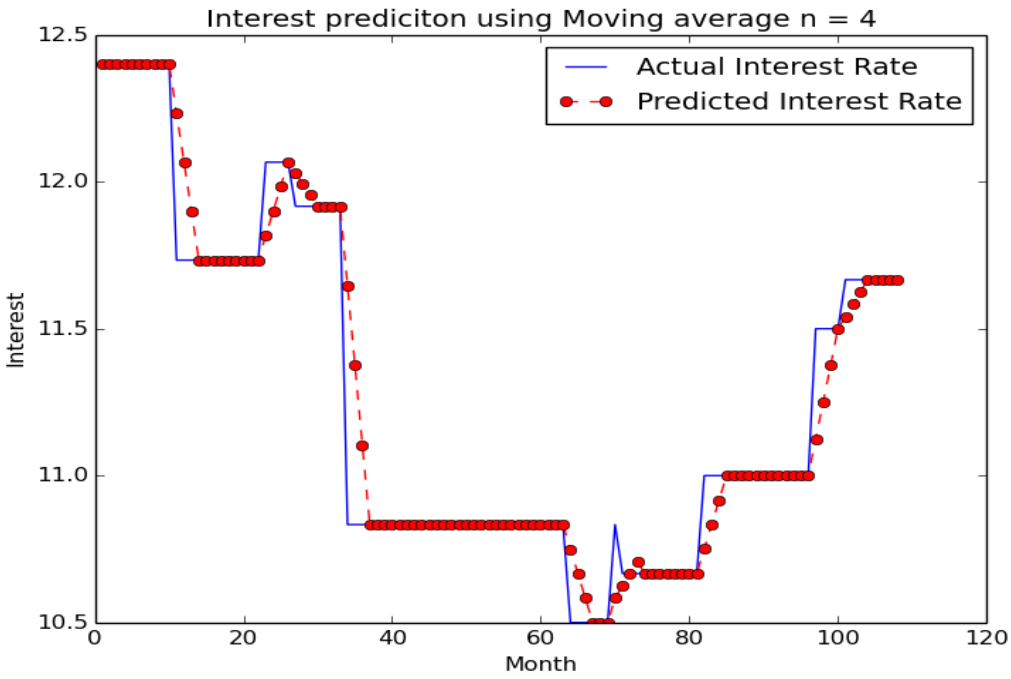

Fig 36: Moving average when window length is N = 4

From the above three figures it can be seen that there is not much difference when a window size is chosen to be 3 or 4. Window length of 3 and 4 has exactly same effect with some noise that is acceptable. Window length of 2 seems to have some overfitting nature as it overlaps the true data points and tends to stay the same as training set. But since this method is used only to predict the data one time ahead in future, it is more acceptable as noise has lesser effect.

11.2.2 Choosing Value of Alpha in Exponential Smoothing

Alpha also known as smoothing factor is one of the factor that has to be chosen carefully. The value of alpha is to be chosen so as it falls between 0 and 1. In order to get the best value of alpha for the purpose of predication, a graph is drawn of error vs alpha. The value of alpha is iterated between 0 and 1 with an increment of 0.1 and for each value of alpha, exponential smoothing is done and corresponding error is determined.

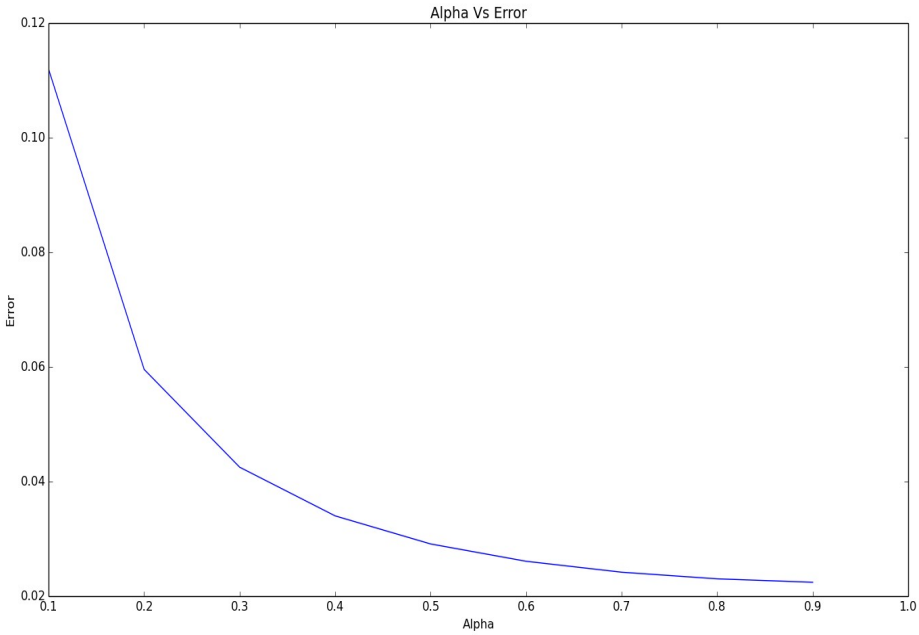

Fig 37: Variation of error with change in Alpha

The error is maximum when alpha is chosen to be 0.1. As value of alpha increase, the error is decreased gradually and seems to attain saturation stage after alpha is 0.8. Exponential smoothing is used to predict interest rate with alpha =0.1 and alpha =0.8.

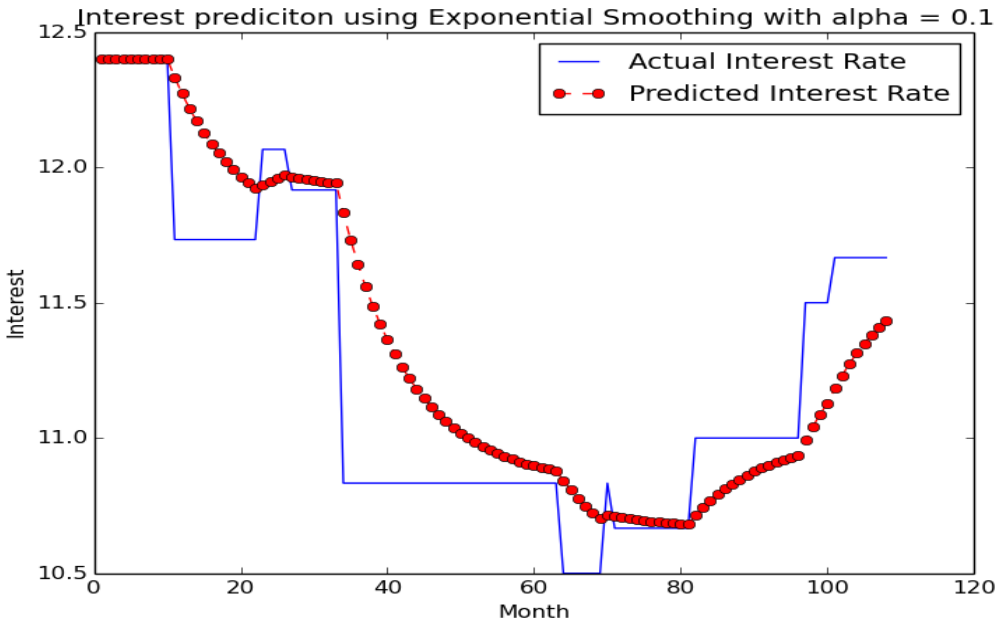

Fig 38: Exponential smoothing with value of alpha = 0.1

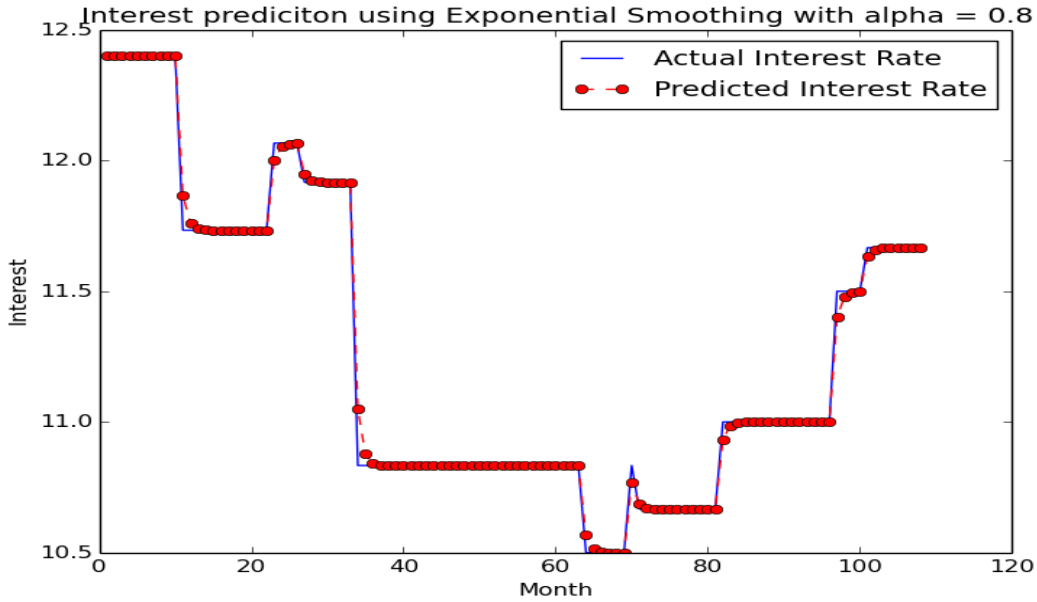

Fig 39: Exponential smoothing with value of alpha = 0.8

Using alpha=0.1 introduces more error than alpha =0.8. So alpha =0.8 will be used for the predication.

11.2.3 Correlation

In this approach all the features were plotted into a single graph to see how closely the features are related with interest. From the graph given below, it is seen that FOREX and interest rate are very closely related.

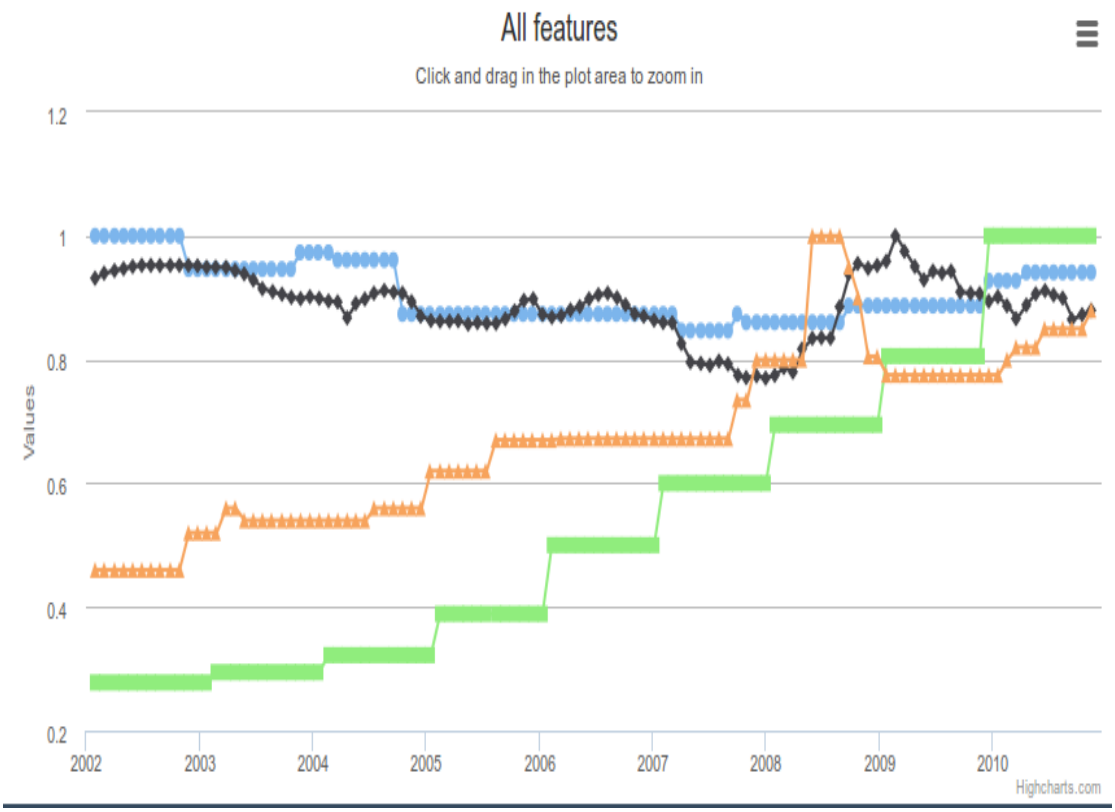

Fig 40: Correlation graph of all features and interest rate

11.2.4 Visualization of Relevant News Data

The figure given below shows different category of news related to the propose of project. The news collected is analyzed to find the strength of different key words stored in cognitive map and thus used as feature to train and test the model.

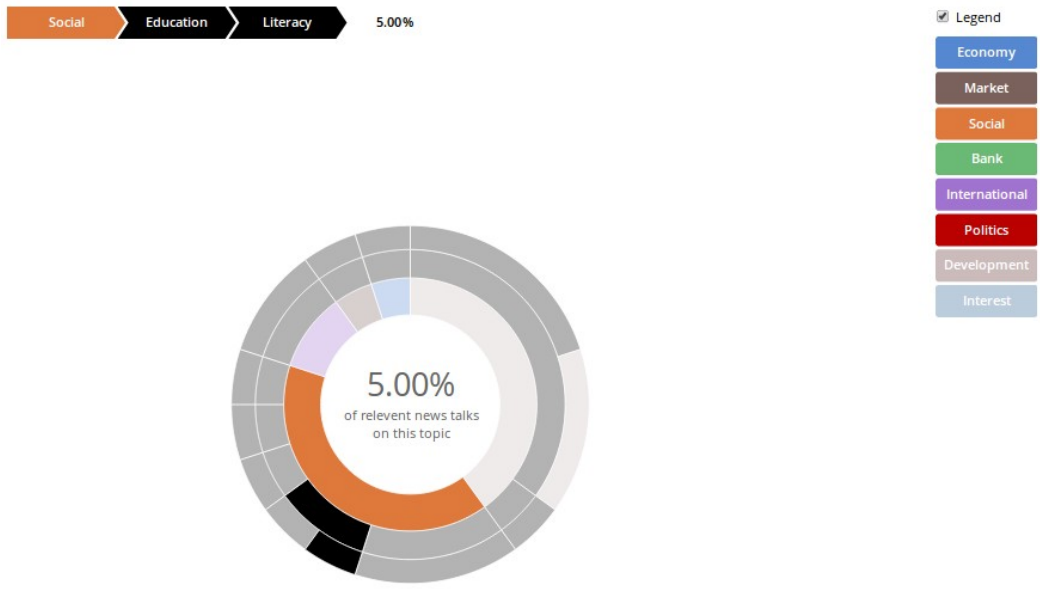

Fig 41: Relevant News Visualization

The graph shows, 5 % of the news related to literary is collected. News for other keywords are being collected for further processing.

11.2.5 Trend of Relative Strength to Interest Rate

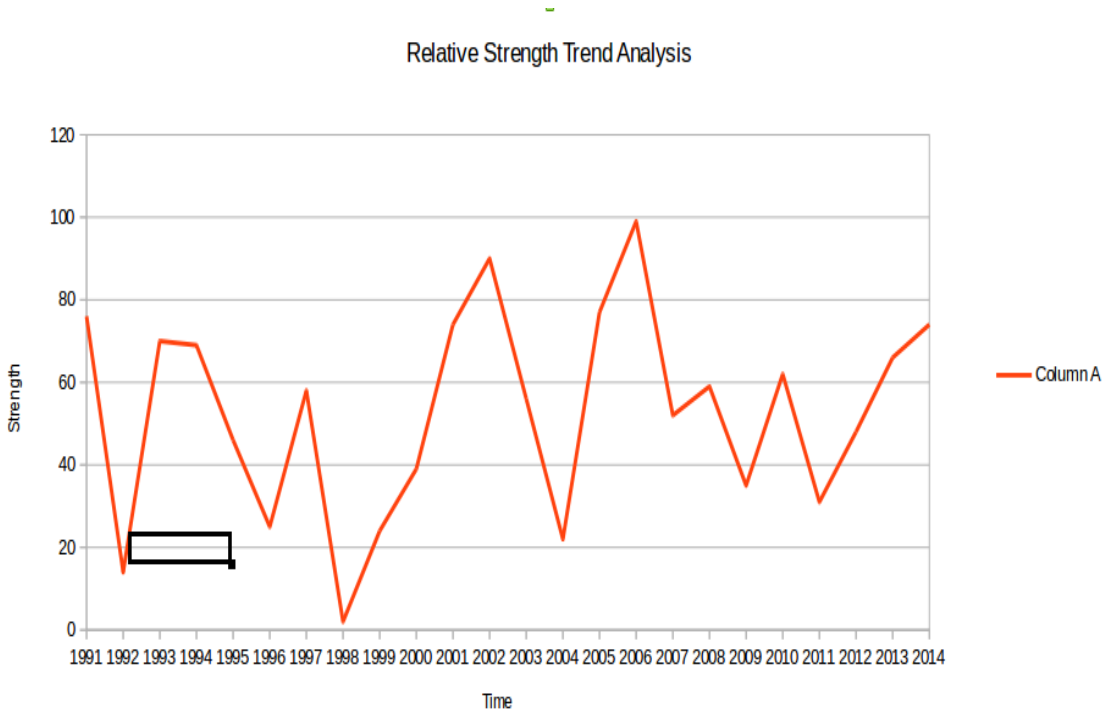

Fig 42: Relative Strength Trend Analysis

The news collected was analyzed using different sentiment analysis algorithms to extract their relative strength and their impact on interest rate. The news trend having value greater than 50% shown in the graph above has positive impact on interest rate where as those trend having value less that 50% has negative impact.

11.2.6 Cognitive Map

Cognitive map is a graphical representation of casual events that has effect on interest rate. The graph shows the snapshot of cognitive map used in the project.

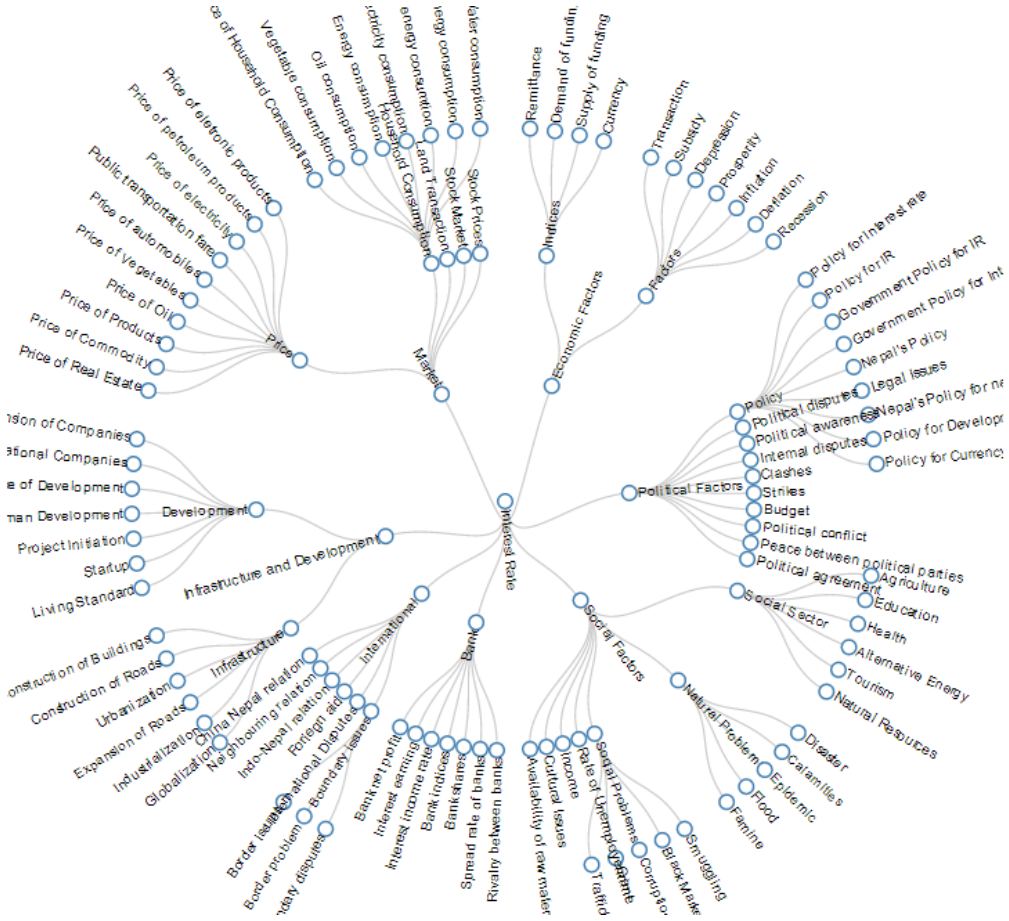

Fig 43: Tree Visualization of Cognitive Map of System

11.3 Data Analysis

Prediction of interest rate using different features involved use of different prediction models. These models take feature vectors as input and gives interest rate as output. Analysis of data involves current trend and future trend of interest rate as predicted by our model. Besides, analysis also includes explanation of factors that are affecting interest rate.

Analysis of different graphs shows that interest rate is gradually decreasing. The reason behind the gradual decrease in interest rate is different form of economic growth that are taking place in the country. This inference can be supported by the trend of GDP and GNI which are increasing. GDP and GNI are indicators of economic development of a country and since there is economic growth more money flow occurs in the market. The cognitive map suggests that with the economic growth interest rate decrease and this is very much observable.

The future trend of interest rate as suggested by our implementation model is predicted to stay constant or slightly decrease to stay near 11%. The predication shows there will be only slight decrement in interest rate over time. Since there is no noticeably large growth in economy and other factors suggested by cognitive map, interest rate will not suffer a great variation as well. Depending upon other factors and current trend it is much likely that interest rate will slightly decrease over time but the decrement won't be that large.

Interest rate is not only dependent on the feature like GDN, GNI, FOREX, gold price and petrol price as we considered for model. But it also depends on other various hidden variables responsible for random nature of interest rate change. The hidden variables are maintained in cognitive map and these are searched in news to get the strength. These strength also effect the interest rate. For instance, if there is unexpected growth in remittance, then it likely that interest rate will decrease.

Chapter 12

12. CONCLUSION AND FUTURE ENCHANCEMENTS

12.1 Conclusion

In this project Interest Rate of Bank was predicted for the context of Nepal. Numeric data collected from different websites and strength retrieved from analysis of news data were used to train a model and predict Interest Rate. The Interest Rate was successfully predicted with sufficient accuracy.

Web application has been developed in which user can see different data visualizations and the predicted interest rate. An Android application has also been developed that can be used to view visualization and keep updated about interest rate.

12.2 Future Enhancements

As per now we have thought of following future enhancements:”

- Design better UI/UX for both the web application and android application to make it more interactive and user friendly.
- Use news data from more than one news source.
- This application might be commercialized by including extra features like ability to compare between different banks and banking plans, EMI Calculator etc.
- We are planning to release API so that other financial institutions and financial newspapers will be able to access our system and utilize different visualizations and analysis.

BIBLIOGRAPHY AND REFERENCES

[1]: Hong, T. and Han, I., (2002), Knowledge Based Datamining of News Information on the Internet using Cognitive Maps and Neural Network, Science Direct, Retrieved From:http://www.sciencedirect.com/science/article/pii/S0957417402000222

[2]: David Enke, Manfred Grauer and Nijat Mehdiyev, (2011), Stock Market Prediction with Multiple Regression, Fuzzy Type-2 Clustering and Neural Networks, ScienceDirect, Retrieved From:www.sciencedirect.com/science/article/pii/S1877050911005035

[3.]: David Enke and Nijat Mehdiyev, (2013), Type-2 Fuzzy Clustering and a Type-2 Fuzzy Inference Neural Network for the prediction of Short-term Interest Rates, ScienceDirect, Retrieved From:www.sciencedirect.com/science/article/pii/S187705091301048X

[4]: Alon Y. Halevy, Jayant Madhavan, (2003), Corpus-Based Knowledge Representation, cs.washington.edu, Retrieved From:https://homes.cs.washington.edu/~alon/files/ijcai03.pdf

[5]: , (2010), Dow Theory, en.wikipedia.org, Retrieved From:http://en.wikipedia.org/wiki/Dow_theory

[6]: Hong, T. and Han, I., (2002), Knowledge Based Datamining of News Information on the Internet using Cognitive Maps and Neural Network, Science Direct, Retrieved From:http://www.sciencedirect.com/science/article/pii/S0957417402000222

[7]: B. Chaib-Draa and J. Desharnais, (1998), A Relational Model of Congnitive Map, AP, Retrieved From:www.mariapinto.es/cibersbstracts/Ariculos/IJHCS-98.pdf

[8]: V.S. Jagtapa and Karishma Pawar, (2013), Polarity Analysis of Sentence, IJSET, Retrieved From:http://ijset.com/ijset/publication/v2s3/paper11.pdf

[9]: Hong, T. and Han, I., (2002), Knowledge Based Datamining of News Information on the Internet using Cognitive Maps and Neural Network, Science Direct, Retrieved From:http://www.sciencedirect.com/science/article/pii/S0957417402000222

[10]: Boris Katz, (1997), From Sentence Proceesing to Information Access on the World Wide Web, AAAIPress.org, Retrieved From:http://aaaipress.org/Papers/Symposia/Spring/1997/SS-97-02/SS97-02-010.pdf

[11]: Hayes, Stephen, (2014), Multiple Regression Analysis using SPSS statistics, Laerd Statistics, Retrieved From:https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php

[12]: Tim Armstrong, (2011), NLTK Part-of-Speech Tagging, eecis.udel.edu, Retrieved From:http://www.eecis.udel.edu/~trnka/CISC889-11S/lectures/armstrong-nltk-tagging.pdf

[13]: Daniel Waegel, (2011), The Porter Stemmer, eeci.udel.edu, Retrieved From:http://www.eecis.udel.edu/~trnka/CISC889-11S/lectures/dan-porters.pdf

[14]: Adam Stepinski, (2005), Automated Event Coding Using Machine Learning Techniques, National Science Foundation, Retrieved From:www.cs.rice.edu/~devika/conflict/papers/poster5.pdf

[15]: Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu and Wayne Niblack(2003),Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques, IBM Almaden Research Center, Retrieved From: http://oucsace.cs.ohiou.edu/~razvan/papers/icdm2003.pdf

[16]: Shlomo Argamon and Moshe Koppel(2013),A SYSTEMIC FUNCTIONAL APPROACH TO AUTOMATED AUTHORSHIP ANALYSIS,lingcog.iit.edu,Retrieved From: http://lingcog.iit.edu/wp-content/papercite-data/pdf/argamon-law-policy-2013.pdf

[17]: Apoorv Agarwal, Fadi Biadsy and Kathleen R. Mckeown(2009), Contextual Phrase-Level Polarity Analysis using Lexical Affect Scoring and Syntactic N-grams, Retrieved From: http://www.aclweb.org/anthology/E09-1004

[18]: Minqing Hu and Bing Liu(2012),Mining and Summarizing Customer Reviews, Retrieved From: http://www.inf.ed.ac.uk/teaching/courses/dme/studpres/MiningSummarizingCustomerReviews.pdf

[19]: Pang, Bo and Lee, Lillian(2004), A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, ACL Web, Retrieved From: http://www.aclweb.org/anthology/P04-1035

[20]: Giorgio Olimpo(2011),Knowledge flows and graphic knowledge representations,ScienceDirect, Retrieved From: http://www.sciencedirect.com/science/article/pii/B9781843346463500058

[21]: Daekook Kang,Yongtae Park(2014),Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach, ScienceDirect, Retrieved From: http://www.sciencedirect.com/science/article/pii/S0957417413006027

[22]: Guglielmo Trentin(2011),Graphic knowledge representation as a tool for fostering knowledge flow in informal learning processes, ScienceDirect, Retrieved From: http://www.sciencedirect.com/science/article/pii/B978184334646350006X

APPENDIX

APPENDIX-A

The following table represents the total tag-set used in the system based on Penn Treebank Tagsets.

SN.	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential there
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun

19. PRP$ Possessive pronoun
20. RB Adverb
21. RBR Adverb, comparative
22. RBS Adverb, superlative
23. RP Particle
24. SYM Symbol
25. TO to
26. UH Interjection
27. VB Verb, base form
28. VBD Verb, past tense
29. VBG Verb, gerund or present participle
30. VBN Verb, past participle
31. VBP Verb, non-3rd person singular present
32. VBZ Verb, 3rd person singular present
33. WDT Wh-determiner
34. WP Wh-pronoun
35. WP$ Possessive wh-pronoun
36. WRB Wh-adverb