Interest Rate Prediction using Sentiment Analysis of News Information

Dr. Arun Timalsina[1], Bidhya Nandan Sharma[2], Everest K.C.[3], Sushant Kafle[4], Swapnil Sneham [5]

[1] *IOE, Central Campus*
[2] *IOE, Central Campus*
[3] *IOE, Central Campus*
[4] *IOE, Central Campus*
[5] *IOE, Central Campus*

Corresponding Email: t.arun@ioe.edu.np

Abstract: This paper presents an approach for interest rate prediction of banks utilizing non-financial parameters such as news data along with other numeric financial parameters. The approach relies on the design and development of an information retrieval system capable of mining online news directories and calculating sentiment scores of pre-indentified news events. The system utilizes cognitive map to contain the region of interest during mining by reflecting non-fuzzy, causal relationship between a selected set of news events and their individual relation on financial market. The accumulative sentiment scores, after causal propagation using cognitive map matrix, along with other numeric features are then supplied to the prediction model for the prediction of interest rates. This paper focuses on the integrated concept of cognitive map, sentiment analysis, prediction models and its effectiveness. The performance of the model was validated using K-fold cross validation technique. The method proved to be sufficiently accurate and generated some exciting results.

Keywords: Information retrieval; Sentiment Analysis; Cognitive Map; News Mining; Regression Analysis; Correlation Analysis

1. Introduction

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 which can be readily used for data mining applications 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 and that news needs to be collected and classified to be able to use them in data mining applications. There have been several attempts to predict interest rates using the time series model, neural networks model, the integrated model of neural networks and case-based reasoning [1, 2]. Meanwhile another approach was attempted in the prediction of the stock price index took into account non-numerical factors such as political and international events from newspaper information.

The system consists of mainly of four components: Information Retrieval, Sentiment Analysis, Knowledge Representation, and Prediction model. The general overview of each component of the system is described in this section. A detailed implementation description of every part is described in Section 2.

1.1 Information Retrieval

Information retrieval represents the process of mining the news data from the web and extracting required information for the data. Information extraction involves determining the polarity of the extracted news data. The news were extracted from ekantipur.com of eight years from 2002 to 2010 as it the only online site in Nepal that maintains repository of past news in English.

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 [3, 4], information retrieval from text, question-answering system [5], summarization of texts and many more [6].

The main concern of sentiment analysis in the system is to determine the polarity of the keywords, maintained on the knowledge base, from the extracted news data.

Polarity analysis of sentence means the problem of classifying a sentence 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 developed. Most existing techniques for polarity analysis sentiment classification are based on supervised learning, for example, n-gram features and other on machine learning methods (Naïve Bayes, Maximum Entropy classification, and SVM).

Keyword polarity analysis of sentences, on the other hand, has been a new topic of research and has had a slow start. Keyword polarity analysis means the problem of extracting sentiment of a keyword in a sentence. Different approaches were analyzed for the keyword sentiment analysis of the news data. The detailed overview is included in Section 1.2.

1.2 Approaches to Sentiment Analysis

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. The system is based on the later one where the concerned keyword is determined to be either increasing or decreasing. Following sentiment analysis techniques were implemented and tested.

1.2.1 n-gram Matching Technique. N-grams 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 [7].

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.

1.2.2 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 technique works by finding out such combination used to describe the keyword in the sentence, usually the nearest opinion phrase from the keyword. The sentiment of keyword is then determined by the sentiment of the respective opinion phrase. Extraction of opinion phrase makes it 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.

1.2.3 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 [5]. Breaking up complex sentence into simpler sentence allows easier analysis of sentences [8, 9]. The process is reliant upon the use of Syntactic Parser for dependency tree generation. Kernel sentences are generally represented by ternary expression such as:

<Subject, Verb, Target>

For a sentence:

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

Kernel Sentences would be:

<___, increase, economy>, <unemployment rate,
increasing, ___>

1.3 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 [10, 11]. Although Artificial Intelligence 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.

Cognitive Map was introduced by Axelrod. It 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.

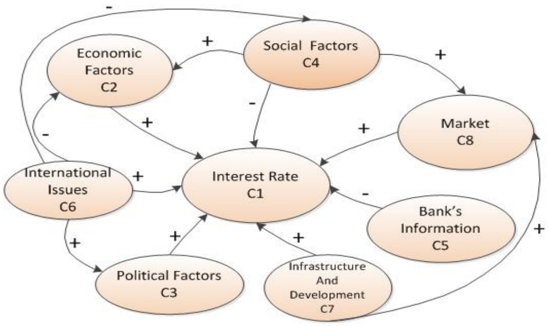


Figure 1: Cognitive Map of the System

1.4 Prediction Model

Prediction model is used to predict the final value of the predicted interest rate. For predication a model has to be chosen in such a way that it does a perfect modeling (neither the model has bias or variance). It should perfectly represent the nature of the data. Choosing a random algorithm and considering that to be the best model might not be a good idea. One has to select the best model by performing different test on a set of candidate algorithms. The candidate algorithms were Linear Regression, Multiple Regression, SVM, Decision Tree, Moving average and single exponential smoothing.

2. Overall System

The proposed system for the project is discussed in the following sub-topics.

2.1 System Overview

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 primary knowledge base for the system.

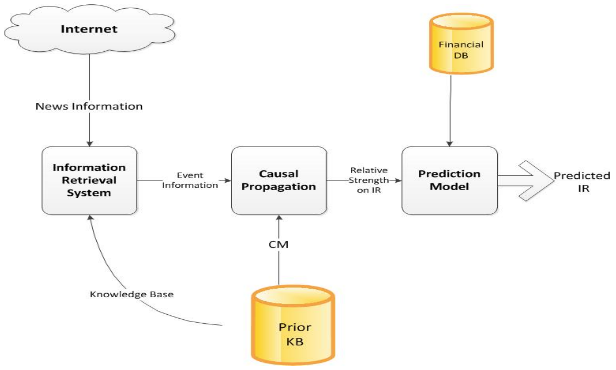


Figure 2: System Block Diagram

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.

2.2 Information Retrieval 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 the database in a daily basis.

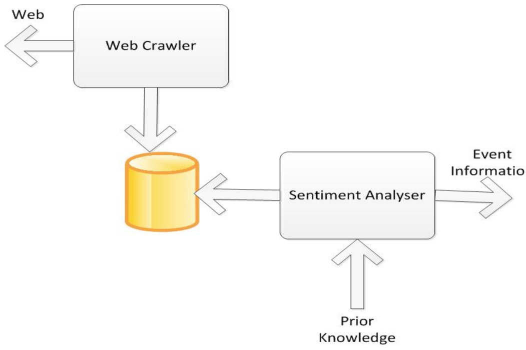


Figure 3: Information Retrieval System

This process is characterized as web scraping in the system. 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 Prior Knowledge Base. Sentiment Analyzer makes use of "Opinion Extraction Technique" for keyword polarity analysis. This returns a vector of event information where each vector element represents an event.

2.3 Prior Knowledge Base

Prior Knowledge Base represents the heart of the system. It represents the knowledge discovered by human intelligences, thus delineating it as an 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.

2.4 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.

2.5 Prediction Model

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

$$y = c + \sum (a_i * F_i) + \sum (b_i * F_i^2) \qquad .... \ (1)$$

The prediction model uses financial data along with the relative strength as features for its model. The models train its self with the financial data.

3. Dataset

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

3.1 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

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.

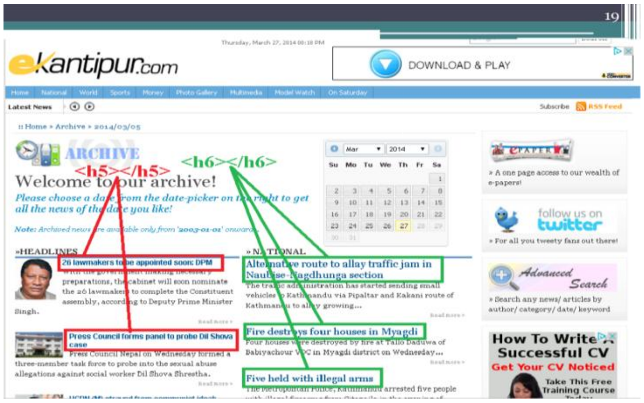
index: The index, starting at 1, which identifies which table or list as defined in the HTML source should be returned.

Example:

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

3.2 News Data

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



Thursday, March 27, 2014 08:18 PM

:: Home » Archive » 2014/03/05

ARCHIVE

Welcome to our archive!

Please choose a date from the date-picker on the right to get all the news of the date you like!

Note: Archived news are available only from '2003-01-01' onwards.

»HEADLINES

26 lawmakers to be appointed soon: DPM

With the government making necessary preparations, the cabinet will soon nominate the 26 lawmakers to complete the Constituent assembly, according to Deputy Prime Minister Singh.

Press Council forms panel to probe Dil Shova case

Press Council Nepal on Wednesday formed a three-member task force to probe into the sexual abuse allegations against social worker Dil Shova Shrestha.

» NATIONAL

Alternative route to allay traffic jam in Naubise-Nagdhunga section

The traffic administration has started sending small vehicles to Kathmandu via Pipaltar and Kakani route of Kathmandu to allay growing...

Fire destroys four houses in Myagdi

Four houses were destroyed by fire at Tallo Daduwa of Babiyachour VDC in Myagdi district on Wednesday...

Five held with illegal arms

The Metropolitan Police, Kathmandu arrested five people

» A one page access to our wealth of e-papers!

» For all you tweety fans out there!

» Search any news/ articles by author/ category/ date/ keyword

Figure 4: Screenshot of ekantipur.com

4. Evaluation

Both the sentiment analysis techniques and prediction models were evaluated for the system.

4.1 Keyword Sentiment Extraction

The best keyword sentiment analysis algorithm satisfying the constraints and the performance demanded by the system was chosen upon evaluation using several datasets. The datasets from news, blog and twitter feeds were gathered and manually labeled. The labeled data was fed to each of the polarity analyzing techniques and the results were evaluated.

The data used for the evaluation are shown below.

Table 1: Evaluation Dataset

Data	Number
News	50 Sentences
Blog	50 Sentences
Twitter	500 Tweets

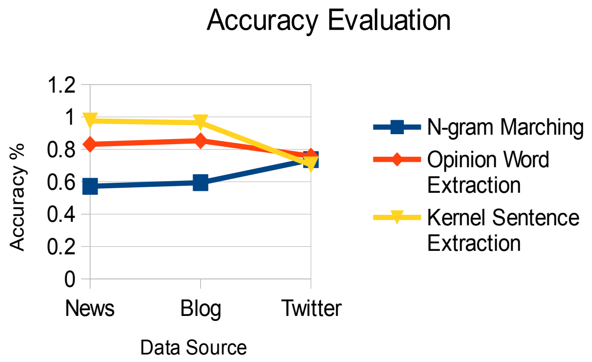


Figure 5: Accuracy Evaluation of Polarity Analyzers

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

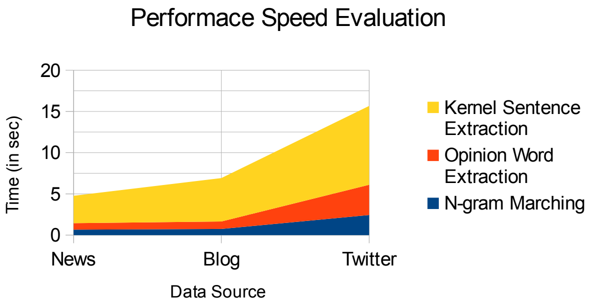


Figure 6: Performance Speed Evaluation of Polarity Analyzers

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

4.2 Model Learning and Predictions

SVM, Decision tree and different forms of multi-variants linear regressions were chosen as candidate models for learning. Details on the variants of linear regression used in the project are described below in points.

i. Simple Model

$y = c + a_1{*}PP + a_2{*}GP + a_3 * FO + a_4 *GDP + a_5 *GNI + a_5 *S$ .... (2)

where,

y = interest rate

PP = petrol price

GP = gold price

FO = Foreign Exchange

GDP = Gross Domestic Product

GNI = Gross National Income

S = Sentiment score

F_i ε (PP, GP, FO, GDP, GNI, S)

ii. Quadratic Model

$y = c + \sum (a_i * F_i) + \sum (b_i * F_i^2)$ .... (3)

where i changes from 1 to N(number of features) a_i and b_i are the coefficients, Fi are the features.

iii. Cubic Model

$y = c + \sum (a_i * F_i) + \sum (b_i * F_i^2) + \sum (c_i * F_i^3)$ .... (4)

where i changes from 1 to N(number of features) a_i ,b_i and c_i are the coefficients, Fi are the features.

iv. Sinusoidal Model

$y = c + \sum (a_i *\sin(F_i)) + \sum (b_i * \cos(F_i))$ ...... (5)

where a_i and b_i are the coefficients, F_i are the features.

v. SVM

Sklearn is a machine learning library in python. SVM was used in the system and it was available in sklearn.

vi. Decision Tree

Decision Tree Regression was also used with the help of sklearn library.

4.3 Cross Validation

In order to select the best model for the purpose, k-fold cross validation was done. The result is shown below.

Table 2: Results of Cross Validation

Models	MSE	Remarks
Simple Regression Model	1.52	
Quadratic Model	1.14	Min error(Selected)
Cubic Model	1.55	
Sinusoidal Model	1.75	
SVM	3.44	Max error
Decision Tree	1.63	(selected)

The graph of error each of the models generated during each folds of cross validation is given below.

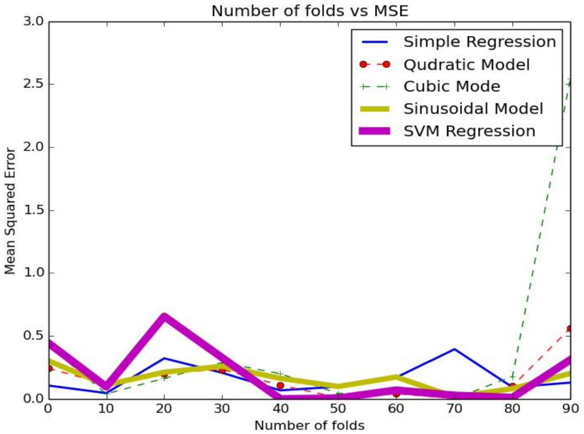


Figure 7: MSE vs Number of folds of each models

4.4 Selection and Training

As k-fold cross validation suggested, quadratic model was selected for the purpose. Decision Tree was also selected as it is less likely to face over-fitting or under-fitting problem. The table below shows the errors the selected models generated during training.

Table 3: Evaluation Dataset

Models	MAE	MSE	RMSE	Remarks
Quadratic Powered Regression Model	0.225	0.072	0.269	
Decision Tree	4.9e-12	7.4e-22	2.7e-11	Min Error

5. Results and Prediction

Quadratic and Decision Tree models were used for prediction. The values they predicted for next four months are in the table below.

Table 4: Evaluation Dataset

Month	Decision Tree	Quadratic
September	10.833	11.94
October	10.833	12.10
November	10.833	12.36
December	10.833	12.77

The prediction results were visualized as:

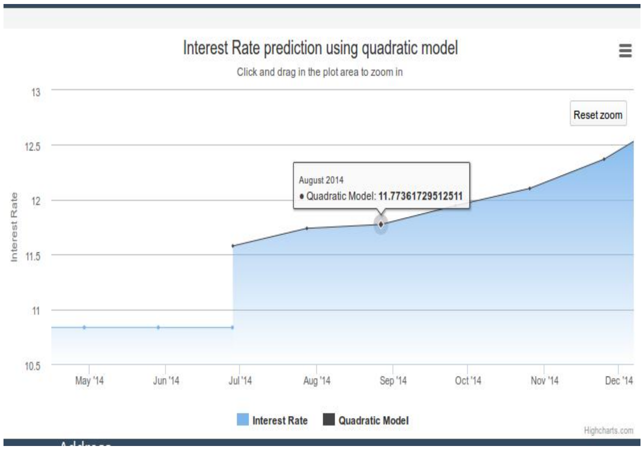


Figure 8: Prediction using quadratic Model

The predicted value using Quadratic model for month of August was very close compared to the actual value 12% as mentioned in http://www.rbb.com.np/interest_rates.php

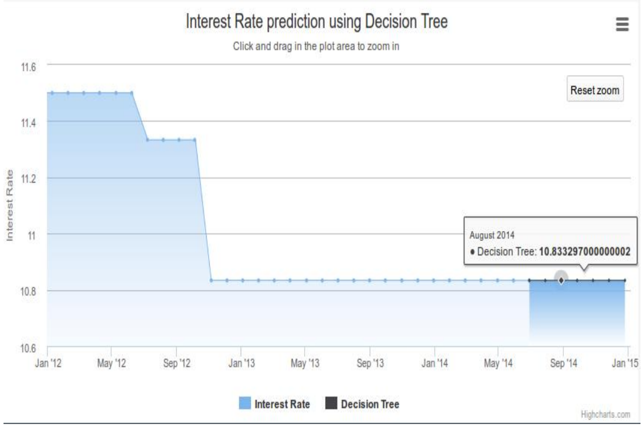


Figure 9: Prediction using Decision Tree Regression

The value predicted by Decision Tree proved sufficiently accurate as well.

6. Conclusion and Future Work

This system attempts to sketch a framework for predicting financial terms with the use of news sentiments and other numerical features. The use of cognitive maps to represent the knowledge and scope of the system and its effectiveness in Information Retrieval provided insightful results. These collaborative techniques performed well in the empirical tests also proved the importance of news sentiments in the system for the prediction.

The keyword sentiment analysis technique also proved impressive as it was successful in extracting the news event sentiments with sufficient accuracy. The results on recent news mining (August 17, 2014 to August 23,

2014) revealed significant news composition results such as news on disaster in Nepal (flood of 2014, Aug), also news on epidemics (the attack of Ebola virus, 2014).

Besides, the study also shows how CMs could be used for knowledge representation and usage.

The system would perform even better if the CMs were refined and made more atomic, incorporating more features. The sentiment analysis is also an area of improvement. Other financial prediction models can be explored for performance.

References

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

[2]: 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

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

[4]: 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

[5]: 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

[6]: 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

[7]: 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

[8]: 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

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

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

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