ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 6, June 2016

Improved Accuracy of FOREX Intraday Trend Prediction through Text Mining of News Headlines using J48

Mrs. Remya Vijayan, Mrs M. A. Potey

***Abstract*— FOREX market is the largest financial market today. Short term predictions of FOREX based on text mining of news headlines is an ongoing research area of data mining. Investors or traders always need a tool to make decisions about their investments based on the current situations of the market. Our system is automating this process by capturing the information available regarding the FOREX market in the financial news headlines. Main objective of the proposed system is the improved FOREX intraday trend predictions from news headlines. Various existing systems utilize the KNN classifier or SVM for classification purposes but it suffers from time and memory efficiency. The proposed system uses the J48 classifier which is efficient in terms of time and memory. Results also shows that accuracy of the proposed system is increased by utilizing multilayer dimension reduction algorithm and j48 classifier.**

***Index Terms*—FOREX market, Text Mining, J48, SVM, Forex Intraday prediction, News headlines.**

I. INTRODUCTION

Today financial markets are the heart of economies. Forex market is one of the largest financial market where trillions of transactions are happening each day apart from weekends. It is important to forecast these market predictions for the betterment of the society. Financials markets run by means of information available. Traders or investors always keep an eye on latest news or reports to understand the current situation of the FOREX market. The proposed system uses news headlines and predicts the directional movements of the market.

Mainly there are two types of market predictions. They are long term and short term predictions. FOREX Intraday trend prediction is short term prediction of FOREX market based on the news headlines released in the past couple of hours. Based on the news headlines, investors or traders will get an idea about the present situation and this information is used to match with the past similar situations and assumption is that the market behave in the same way as it behaved in the similar situations in the past[1]. The msin objective of the proposed work is to improve the accuracy of prediction of FOREX intraday market based on the related financial news.

As the number of information sources increases due to internet, there is a huge availability of high dimensional data available in the internet. This can decrease the accuracy of predictions. The main motivation behind this work is the use of multilayer dimension reduction algorithm [1] to deal with the semantic redundancy as well as the sentiment integration. The multilayer dimension reduction algorithm consists of 3 layers. In each layer, it take care of dimensionality reduction in an incremental fashion.

The selection and processing of financial news for decision making is a challenging task. Another motivation towards this work is the use of news headlines instead of news article bodies. News headlines gives the summary of the news article and is straight to the point.

The existing system used SVM for classification whereas the proposed system uses J48 classifier which is time and memory efficient.

This paper is organized into 5 sections. Section II has review of related work. Section III gives the Proposed System. Section IV and V gives the results and conclusion respectively.

II. RELATED WORK

Market prediction through text mining is an emerging area To work. There are some theories existing while forecasting future prices of money markets. They are Efficient Market Hypothesis (EMH) [20] and Random Walk Theory [21]. EMH states that the price of a security reflects all of the information available and everyone has some degree of access to information. The market reacts to any given news and that it is impossible to outperform the market [10]. Random Walk Theory assumes that all public information is available to everyone. The predictions are considered ineffective where prices are determined randomly [10]. Paper [7] proposed a model for mining multiple time series data based on EMH.

From these theories, two types of analysis emerged. i.e., technical analysis and fundamental analysis. Fundamental analysis deals with unstructured textual data whereas technical analysis depends on historical and time-series data.

Manuscript received June 2016
***Mrs Remya Vijayan**, Dept. of Computer Engineering, D.Y. Patil College of Engineering, Akurdi, Pune.*
***Mrs. M. A. Potey,** Dept. of Computer Engineering, D.Y. Patil College of Engineering, Akurdi, Pune*

Paper [10] uses fundamental analysis of breaking financial news for stock market prediction.

A. Preprocessing Phase

In preprocessing phase, feature selection, dimensionality reduction and feature representation are the main steps.

1) Feature Selection

Feature selection improves the performance of the prediction and provides faster and more robust estimation of model parameters. One such method commonly used is called bag-of-words [7] [13] [14] which basically breaks the text into words and each word is considered as feature. Another Method is LDA i.e., Latent Dirichlet Allocation technique [9]. It is a topic model that automatically identifies topics that these documents contain. Charater-n-grams [8] is a continuous sequence of n items from a given sequence of text or speech.

2) Dimensionality Reduction

In some situations we need to decrease the dimension of the data to a size which can be easily handled, keeping as much of the original information as possible. Paper [1] [8] uses a nominal repetition limit and reduces the terms by choosing the ones reaching a number of occurrences. Another common approach is the use of predefined dictionary [11] to put them in the place of a category name or value. Feature stemming, conversion to lower case letters, removal of punctuation, numbers, stop words and web page addresses are some commonly used techniques. Paper [1] uses a multilayer dimension reduction algorithm for the reduction of dimensions. Parallel rare term vector replacement [15] is another method used. Rare terms are replaced by vectors of common terms.

3) Feature Representation

The features are represented in binary, tf-idf, sentiment score etc. The model utilized in this study [19] to represent the document in high dimensional space. It represents each document as binary vectors where each element is a word from a vocabulary. The elements will have a value of 1 if the corresponding word is available inside the record or have an estimation of zero generally. A weight can be associated with each element to reflect their relative significance. The concepts are weighted by the conventional multiplicative combination of term frequency (TF) and the inverse document frequency (IDF), so that terms occur more often in a document and/or rarer in other documents will be given a higher weight.

4) Machine Learning Algorithms

Paper [5] puts an accentuation on the kind of utilized dataset, the structure of pre-preparing and the type of machine learning algorithm used in categorizing the available systems. In machine learning, feature learning or representation learning is a set of techniques that learn a feature: a transformation of raw data input to a representation that can be effectively exploited in machine learning tasks. There are many machine learning algorithms such as linear Regression, Logistic Regression, C4.5 Decision Tree [12], SVM, Naive Bayes[6], KNN, K-Means, Random Forest, Dimensionality Reduction Algorithms, Gradient Boost and Adaboost which can be applied to almost any data problem.

In paper [1], it is additionally sometimes referred to as the 'Multi-layer algorithm' in short. It manages the core aspects of how features are selected for a feature-vector to be fed into the machine learning algorithm with the end goal of its training (model creation) and prediction execution. For model creation and prediction, they used SVM. Paper [4] studies various ways of predicting fluctuations in the foreign market. Paper [2] utilizes twitter information for stock market predictions. In [3], the author uses textual news information for predicting intraday stock returns. Paper [17] put emphasis on the need of sentiment analysis on social media for financial market prediction.

III. PROPOSED SYSTEM

The proposed system is divided into various phases.

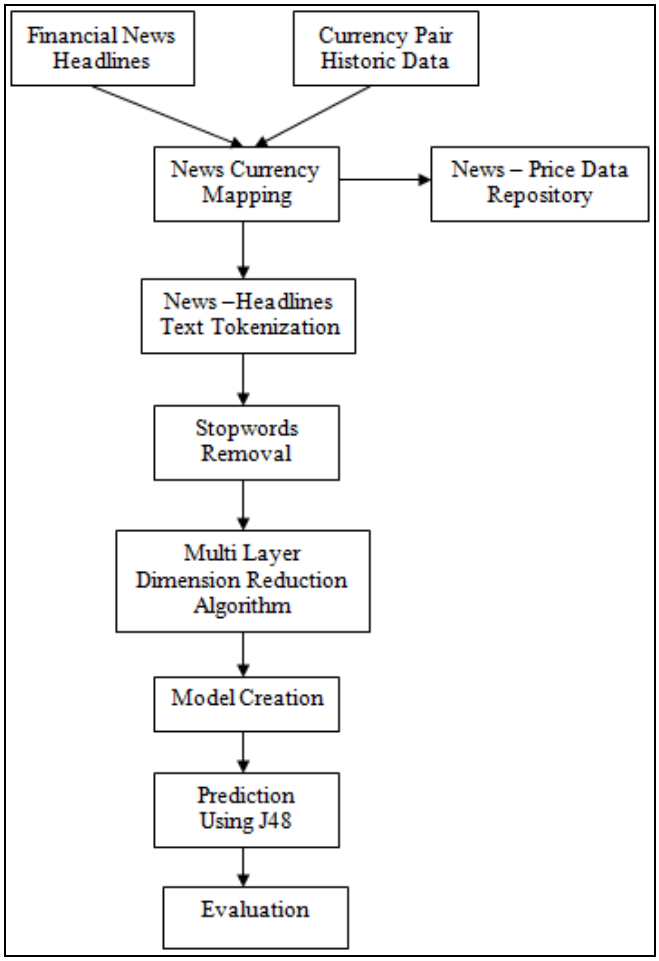

Figure 1. Proposed System Architecture

A. Input Data Selection

In data retrieval phase, the system input two dataset. First dataset is news dataset in which content news and time are retrieved. The second dataset is the currency dataset in which the rate of currency pair along with time is provided.

B. News Currency Mapping

News-headlines are grouped as per the time interval and label is assigned as per the currency pair rate historic data.

C. Text Tokenization

In this phase the system first tokenize the text of the news headlines which are grouped based on time interval.

D. Stopword Removal

The stopwords are identified and removed.

E. Multilayer Dimension Reduction Algorithm

1) Semantic Abstraction via Heuristic-Hypernym Feature Selection.

Feature Selection is done using Heuristic-Hypernym method. Hypernyms are found for each word from WordNet dictionary. It makes simpler feature space by reducing the words that are used as features. This phase addresses the problem of semantic abstraction and coreferences.

2) Sentiment Integration via TF-IDF*SumScore weighting.

The scaling by TF-IDF is important due to the amount of repetitions of a word in a file or document and the corpus is significant for a weighting scheme to be specific to a context. The SentiWordNet SumScore that is a new score defined here. SentiWordNet is a dictionary of sentiment values that have a Positivity Score (PosScore) and/or a Negativity Score between 0 and 1 for each WordNet entry.

3) Synchronous Targeted Feature-Reduction

It contains target window determination, target feature reduction and target feature vector creation. It is used to solve the problem of no effective feature-reduction. Further the number of features are reduced and target feature vector is generated.

F. Prediction of Label

Here model is generated and used to guess the label for the targeted record. For prediction we use J48 classifier which takes less time and memory than the other classifiers.

G. Evaluation

Precision and recall is used for evaluation purposes. Let

True class A (TA) = correctly classified into class A,
False class A (FA) = incorrectly classified into class A,
True class B (TB) = correctly classified into class B,
False class B (FB) = incorrectly classified into class B.
Precision = TA / (TA + FA)
Recall = TA / (TA + FB)

IV. ALGORITHM

A. Synchronous Targeted Label Prediction (STLP) [1] algorithm

STLP includes Synchronous Target Feature Reduction (STFR). STFR optimizes feature-reduction by minimizing the features to the lowest that is required for the prediction purpose and creates a new model for every prediction. It creates new models synchronously at the time as the prediction needs to happen. STLP takes TF-IDF*SumScore Feature Matrix (M) as input and the output is the label of prediction (LABEL) i.e., Up or Down movements of the FOREX market.

B. J48 algorithm

J48 is the open source implementation of C4.5 algorithm in Weka. J48 classifier is the classification algorithm used for detecting the novel and multi-novel class. For the problem of classification, the methodology of decision tree is used. C4.5 algorithm deals with continuous attributes and missing data. Decision maker selects decision tree since it is simple and easy to learn [18] [16].

1. Input: Training data DS
2. Output: Decision Tree DT
3. DSTBUILD (*DS)
4. {
5. DT= ϕ;
6. DT=Generate root node and label with splitting attribute;
7. DT=Add arch to root node for each splitting predicate and label;
8. DS=By applying split predicate to DS database is created;
9. If stopping point reached for this path, then;
10. DTr =generate leaf node and label with the appropriate class;
11. DTr =DSTBUILD(*DS);
12. Else
13. DTr =DSTBUILD (DS);
14. DT=add DTr to arc;
15. }

V. DATASETS

System uses two types of datasets as inputs. The News-headlines datasets consists of the timestamped news headlines, captured from economic times. The second dataset used is the currency pair dataset which depicts the rate of

currency pair along with the time. INR/USD currency pair is used for the experimental purposes.

VI. Result

The experimental findings show that J48 is time as well as memory efficient as compared to SVM classifier. The J48 take less time than the SVM for classification and improves the performance.

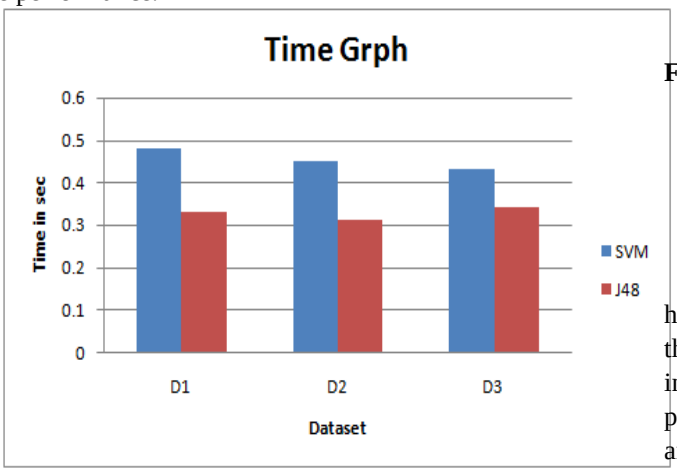

Figure 2. Time Comparison of SVM and J48

The memory comparison between SVM and J48 in byte can also be evaluated. The J48 takes less memory than the SVM for classification and minimize the memory overhead.

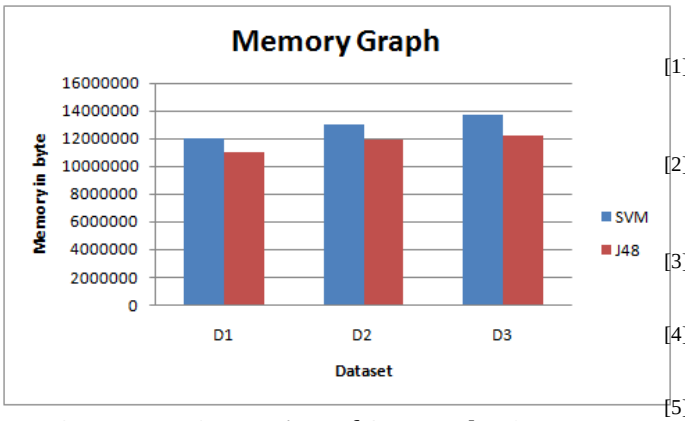

Figure 3. Memory Comparison of SVM and J48

The graph below shows the accuracy comparison between SVM and J48 in percentage. The J48 have high accuracy than the SVM for classification and improves the performance of the system.

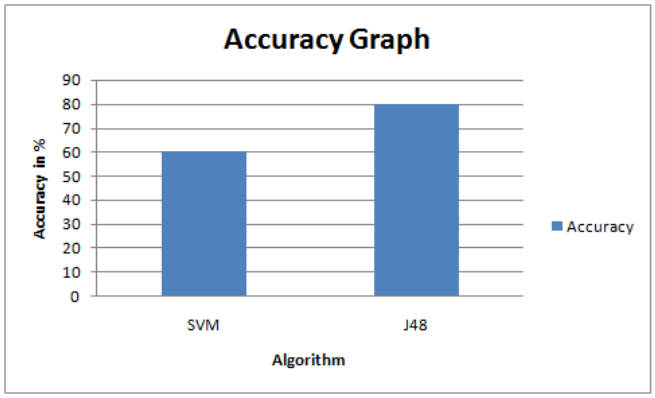

Figure 4. Accuracy Graph

VII. Conclusion And Future Scope

FOREX market prediction from unstructured text of news headlines is a challenging task. The proposed system predicts the intraday movements of a currency pair from the information available in the financial news-headlines. The proposed framework used J48 algorithm for model creation and label prediction. J48 algorithm is memory and time efficient and improves overall accuracy of the system. In future, the system can be used for paid softwares which predicts market behavior. The proposed techniques can be applied in other contexts like movie reviews, other financial market types.

REFERENCES

[1] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, D. C. Ling Ngo, "Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment", Expert Systems with Applications 42 (2015) 306–324.

[2] Oliveira, Nuno, Paulo Cortez, and Nelson Areal. "Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from Twitter." Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics. ACM, 2013.

[3] Geva, Tomer, and Jacob Zahavi. "Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news." Decision support systems 57 (2014): 212-223.

[4] Cross, D. W., Chris J. Hinde, and Martin D. Sykora. "Predicting fluctuations in foreign exchange rates." Computational Intelligence (UKCI), 2013 13th UK Workshop on. IEEE, 2013.

[5] Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., & Ngo, D. C. L, Text mining for market prediction: A systematic review. Expert Systems with Applications, 2014.

[6] Yu, Yang, Wenjing Duan, and Qing Cao. "The impact of social and conventional media on rm equity value: A sentiment analysis approach. " Decision Support Systems 55.4 (2013): 919-926.

[7] Fung, Gabriel Pui Cheong, Jerey Xu Yu, and Wai Lam. "Stock prediction: Integrating text mining approach using real-time news." Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on. IEEE, 2003.

[8] Butler, Matthew, and Vlado Keselj. "Financial forecasting using character n-gram analysis and readability scores of annual reports." Advances in artificial intelligence. Springer Berlin Heidelberg, 2009. 39-51.

[9] Mahajan, Anuj, Lipika Dey, and S K Mirajul Haque. "Mining financial news for major events and their impacts on the market. " Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on. Vol. 1. IEEE, 2008.

[10] Schumaker, Robert P., and Hsinchun Chen. "Textual analysis of stock market prediction using breaking financial news: The AZFin text system." ACM Transactions on Information Systems (TOIS) 27.2 (2009): 12.

[11] Li, Feng. "The information content of forward-looking statements in corporate filings—a naïve Bayesian machine learning approach." Journal of Accounting Research 48.5 (2010): 1049-1102.
[12] Vu, Tien-Thanh, and et al. "An experiment in integrating sentiment features for tech stock prediction in twitter." (2012): 23-38.
[13] Peramunetilleke, Desh, and Raymond K. Wong. "Currency exchange rate forecasting from news headlines." Australian Computer Science Communications 24.2 (2002): 131-139.
[14] Groth, Sven S., and Jan Muntermann. "An intraday market risk management approach based on textual analysis." Decision Support Systems 50.4 (2011): 680-691.
[15] Berka, Tobias, and Marian Vajteršic. "Dimensionality reduction for information retrieval using vector replacement of rare terms." Data Mining for Service. Springer Berlin Heidelberg, 2014. 41-60.
[16] Selvanayaki, M., et al. "Supervised Learning Approach for Predicting the Quality of Cotton Using WEKA." Information Processing and Management. Springer Berlin Heidelberg, 2010. 382-384.
[17] Nguyen, Thien Hai, Kiyoaki Shirai, and Julien Velcin. "Sentiment analysis on social media for stock movement prediction." Expert Systems with Applications 42.24 (2015): 9603-9611.
[18] Dr. Neeraj Bhargava, Manish Mathuria, "Decision Tree Analysis on J48 algorithm for Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013.
[19] Zhai, Y., Hsu, A., & Halgamuge, S. K. (2007). Combining news and technical indicators in daily stock price trends prediction. In Proceedings of the fourth international symposium on neural networks: advances in neural networks, Part III (pp. 1087-1096). Nanjing, China: Springer-Verlag.
[20] Fama, E., The behavior of Stock Market Prices, in Graduate School of Business, 1964, University of Chicago.
[21] Malkiel, B.G., A Random Walk Down Wall Street. 1973, New York: W.W.Norton & Company Ltd.

.