ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

Survey on Content Digging of News-features for FOREX Market Forecast

Remya Vijayan[1], Ms. Soudamini S Pawar[2]

ME student, Dept. of Computer Engineering, D. Y. Patil College of Engineering, Akurdi, Pune, India[1]

Assistant Professor, **Dept.** of Computer Engineering, D. Y. Patil College of Engineering, Akurdi, Savitribai Phule Pune University, Pune, India[2]

ABSTRACT: Today's greatest economies of the world are business sector economies. This paper investigate how business news headlines from the various news sources can be used to predict very short-term currency exchange rate behaviors. The concept of the approach is the predictions generated from meaningful information extracted from the news headlines. With the conciseness and precise input of news headlines, improved predictions are expected in comparison to any other financial data input. The output of this approach is the forecast about currency exchange rates: a particular currency up, remains steady or goes down within a very short time-span of the next couple of hours. This work is an effort to put more emphasis on the content mining techniques furthermore, handle some particular viewpoints thereof that failed in previous works, in particular: the high dimensionality and ignoring estimation and semantics in managing literary language. In this paper, we survey on different techniques developed by the researchers for text mining and analysis, for the forex market prediction.

KEYWORDS: News mining, News semantic analysis, Market sentiment analysis, Market prediction, FOREX prediction, Text Mining.

I. INTRODUCTION

FOREX is cash exchanging business sector spread every-where throughout the world. The foreign trade business sector is the spot where monetary standards are exchanged. Monetary standards are imperative to the vast majority since monetary standards should be traded keeping in mind the end goal to lead outside exchange and business. Swapping scale forecast is the most difficult utilizations of time arrangement gauging. For turning into a decent merchant in outside trade market, they should know about the variables that are in charge of a money to acknowledge or devalue. FOREX markets are influenced by numerous instabilities and interrelated financial and political variables at both neighbourhood and worldwide levels.

Securities exchange anticipating is centred around accomplishing best results with least required information. Discovering the arrangement of variables for making precise expectations is a testing undertaking thus consistent securities exchange examination is extremely crucial. The securities exchanges developments are broke down and anticipated with a specific end goal to recover information that could control financial specialists on when to purchase and offer. It will like-wise help the financial specialist to profit through his interest inthe share trading system. News presents to us the most recent data about the stock market. Business and monetary news have a solid relationship with future stock execution. This news can be utilized to extricate assumptions and assessments which can be utilized to help market expectations. The greater part of the speculators settle on choice to either purchase or offer a specific stock in view of the feelings what's more, slants communicated in the news. Consequently there ought to be a few instruments to concentrate data and investigate in order to foresee FOREX market.The outside trade market is a decentralized business sector where the cash traded sums around one billion dollars every day. Banks and other market producers send their trade rates to specific organizations that disperse this data to endorsers. Trade rates really have two segments, the offer cost and the inquire cost. The offer cost is the most astounding cost at which the business sector producer is willing to purchase the cash, while the ask cost is the most reduced cost at which the business sector creator is willing to offer the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

coin to exchange a money, a merchant will call a business sector producer demand the present quote, and choose whether or not to place an exchange.

Another inspiration is the utilization of news features just, not news article bodies. This empowers short bits of content as opposed to long ones. News features are at times utilized and are straight to the point, less commotion as contrasted with general news. This paper is partitioned into distinctive areas. Segment A depicts FOREX Market Overview. Segment B portrays related work.

1.1 FOREX Market Overview

Forex works 24 hours a day, opening Sunday 5pm and shut-ting Friday 5pm. Most dynamic exchanging neighbourhood forex markets happen amid neighbourhood exchanging hours. This is the exchanging of the two monetary standards from two nations against one another. Agents decide the particular sets, who could conceivably offer a match for the coin combine that you need to exchange. A conversion scale is just the proportion of one coin esteemed against another, where the first money is alluded to as the base money and the second as the counter or cite money. There are two unmistakable parameters for coin exchanging - Buying and Selling. In the event that purchasing, a conversion scale determines the amount you need to pay in the counter or quote money to acquire one unit of the base coinIn the case of offering, the conversion standard determines the amount you get in the counter or cite money when offering one unit of the base cash. For instance, a well-known pair that is broadly exchanged is the EUR/USD. The EUR/USD is the European Dollar, additionally known as the EURO, and the USD, which is the US Dollar. Whenever the Euro gets to be worth more cash in dollars, the pair goes up, when it gets to be worth less cash in dollars, the pair drops in esteem. In the event that you anticipate that the Euro will pick up quality quicker than the US Dollar, you would purchase the pair In the event that you expect the US Dollar to pick up worth speedier than the Euro, you would offer this pair. A run of the mill money conversion standard is given as an offer cost what's more, an ask cost. The offer cost is constantly lower than the inquire cost. The offer cost speaks to what will be gotten in the cite coin when offering one unit of the base cash. The ask cost speaks to what must be paid in the quote coin to get one unit of the base cash.

The distinction between the offer and the ask cost is alluded to as the spread. Most monetary standards are exchanged specifically against the US Dollar. The business sector rates that are communicated for such coin sets are called direct rates. For some coin combines, the US Dollar is not the base cash but rather the counter or quote coin. The business sector rates that are communicated for such money sets are called roundabout rates.One money is exchanged against any money other than the USD, the business sector rate for this cash pair is known as a cross rate. Cross rate is the swapping scale between two monetary standards not including the US Dollar. In spite of the fact that the US dollar rates don't show up in the last cross rate, they are generally utilized as a part of the estimation thus must be known. Exchanging between two non-US Dollar monetary standards as a rule happens by first exchanging one against the US Dollar and after that exchanging the US Dollar against the second non-US Dollar.

1.3 Related Work

In this chapter we present a brief overview of Information Extraction, which is an area of natural language processing that deals with finding factual information in free text. Due to high volatility and complexity of market data, there exists some difficulty in predicting currency exchange rates. Time series methods have their limitations for multidimensional time series with mutual non-linear dependencies.

In general the predictive measures are classified into technical and fundamental analyses. Most of the research has been done on the technical analysis approaches in the past. Fundamental data is more challenging to use as input especially when it is unstructured. Market prediction through text mining is a promising area to work. Feature selection, dimensionality reduction and feature representation play an important role in pre-processing step of predictive text mining.

1) Feature Selection: Variable and feature selection have become the focus of much research in areas of application for which datasets with tens or hundreds of thousands of variables are available. Feature selection is used to improve

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

the performance of the prediction and to provide faster and more robust estimation of model parameters. Two discriminative feature selection techniques are identified. They are SVM-based technique and a bagging decision tree technique. SVM-based Selection Technique is a feature selection technique which is an extension of the SVM-RFE algorithm. It finds the best feature subset which minimizes SVM classification error. Bagging trees are excellent tools for feature selection. For each feature, the out-of-bag mean squared error after removingthat feature is averaged over all trees. This is repeated for each feature to come up with the best features list. The most widely recognized system is the so called"Bag of-words" which is basically separating the content into its words and considering each of them as a component. On the other hand, Vu et al. (2012) effectively enhance results by building a Named Entity Recognition (NER) framework to recognize whether a Tweet contains named elements identified with their focused on organizations in light of a direct Conditional Random Fields (CRF) mode. Multilayer Dimensionality Reduction Algorithm [3] is used for feature representation and reduction technique.

The most important method is so called bag-of-words [7] which is essentially breaking the text into words and considering each of them as a feature. Another interesting technique is LDA i.e., Latent Dirichlet allocation technique [19]. It is a topic model that automatically discovers topics that these documents contain. Character-n-grams [14] is a continuous sequence of n items from a given sequence of n items from a given sequence of text or speech.

2) Dimensionality Reduction Techniques: In some situations it is necessary to reduce the dimension of the data to a manageable size, keeping as much of the original information as possible, and then feed the reduced-dimension data into the system. A given processing system is only effective with vector data of not more than a certain dimension, so data of higher dimension must be reduced before being fed into the system. The main underlying text-mining challenge is the high dimensionality of the feature-space. Paper [14] uses a minimum occurrence limit and reduces the terms by selecting the ones reaching a number of occurrences. Another common approaches is the use of predefined dictionary [13] to replace them with a category name or value. Features stemming, conversion to lower case letters, punctuation removal and removal of numbers, stop words and web page addresses are some commonly used techniques. Paper [9] produces a multi-layer dimension reduction algorithm to respond to this need. Another detailed method for dimensionality reduction of text is based on parallel rare term vector replacement [20]. Rare terms are replaced by vectors of common terms.

a) Multilayer Dimension Reduction Algorithm: The algorithm tackles a different root cause of the problem at each layer. The first layer is termed the Semantic Abstraction Layer and addresses the problem of co-reference in text mining that is contributing to sparsity. Co-reference occurs when two or more words in a text corpus refer to the same concept. This work produces a custom approach by the name of Heuristic-Hypernyms Modeling which creates a way to recognize words with the same parent- word to be regarded as one entity. As a result, prediction accuracy increases significantly at this layer which is attributed to appropriate noise-reduction from the feature-space. The second layer is termed Sentiment Integration Layer, which integrates sentiment analysis capability into the algorithm by proposing a sentiment weight by the name of SumScore that reflects investor's sentiment. This layer reduces the dimensions by eliminating those that are of zero value in terms of sentiment and thereby improves prediction accuracy. The third layer encompasses a dynamic model creation algorithm, termed Synchronous Targeted Feature Reduction (STFR). It is suitable for the challenge at hand whereby the mining of a stream of text is concerned. It updates the models with the most recent information available and, more importantly, it ensures that the dimensions are reduced to a number that is many times smaller. The algorithm and each of its layers are extensively evaluated using real market data and news content across multiple years and have proven to be solid and superior to any other comparable solution. On top of a well-rounded multifaceted algorithm, this work contributes a much needed research framework for this context with a test-bed of data that must make future research endeavours more convenient. The produced algorithm is scalable and its modular design allows improvement in each of its layers [9].

3) Feature Representation: Once features are identified, each feature is represented by a numeric value which can be fed to the next phase i.e., machine learning. The numeric value acts like a score or a weight. The most basic technique is a Boolean or a binary representation [11] [13]. Another efficient technique used is TF-IDF [12]. The TF-IDF value increases proportionally to the number of times a word appears in the document. TF-CDF i.e., Term Frequency-Category Discrimination is based on category frequency and is more effective than TF-IDF [12].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

4) Machine Learning algorithms: After the pre-processing is completed and text is transformed into a number of features with a numeric representation, machine learning algorithms can be used. Different methods are used in Foreign Exchange Rates prediction. These methods are distinguishable from each other by what they hold to be constant into the future. Basically the systems are using the input data to learn to classify an output usually in terms of the movement of the market in classes such as Up, Down and Steady.

a) SVM: is a non-probabilistic binary linear classifier used for supervised learning. The main idea of SVMs is finding a hyper plane that separates two classes with a maximum margin. The training problem in SVMs can be represented as a quadratic programming optimization problem. They take the directions of the news things in an imagined archive map as elements. A report map is a low-dimensional space in which every news thing is situated on the weighted normal of the directions of the ideas that happen in the news thing. SVMs can be reached out to nonlinear classifiers by applying bit mapping (portion trap).

In SVM light [12] Training and testing of algorithm is done. Sliding window, semantics and syntax are not mentioned. Use of SVM principally yet they too evaluate the capacity to anticipate the discrete estimation of the stock return utilizing SVR. They foresee returns and figure the R2 (squared connection coefficient) between anticipated and really watched return. The streamlining behind the SVR is fundamentally the same to the SVM, yet rather than a double measure (i.e. positive or negative), it is prepared on really watched returns. While a double measure must be ”genuine” or ’false’, this measure gives more weight to more noteworthy deviations in the middle of real and anticipated returns than to littler ones.

b) Regression Algorithm: They studied training, testing and volume sampling is which not mentioned. Sliding window has negative feedback while semantic and syntax has beendone, while here the software used and data is not mentioned. In regression based, it takes different forms in the research done. One approach is Support Vector Regression (SVR) which is a regression based variation of SVM as discussed earlier.

c) Naive Bayes: In machine learning, naive bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes theorem with strong independence assumptions between the features. It is probably the oldest classification algorithm. But it is still very popular and is used among many of the works. It is based on the Bayes Theorem and it is called naive because it is based on the nave assumption of complete independence between text features. It differentiates itself from approaches such as k-Nearest Neighbours (k-NN), Artificial Neural Networks (ANN), or Support Vector Machine (SVM) in that it builds upon probabilities. The Naive Bayes algorithm is applied on sentiment analysis to examine the effect of multiple sources of social media along with the effect of conventional media and to investigate their relative importance and their interrelatedness. The paper [11] make use of naive bayesfor text classification.

d) Combinatory Algorithms: The author used stacked classifier it is referring to a class of algorithms which are composed of a number of machine learning algorithms stacked or grouped together. Then a stacked classifier is used which is a trainable classifier that combines the predictions of multiple classifiers via a generalized voting procedure. The voting step is a separate classification problem. They use a decision tree based on information gain for handling numerical attributes in conjunction with an SVM with sigmoid kernel to design the stacked classifier. The paper [15] uses self-organizing fuzzy neural network as combinatory algorithm.

5) Sentiment Analysis: Sentiment analysis is used for detecting the general sentiment that is hidden in online resources and social media to understand and analyse how people feel about a topic. Sentiment analysis of news is a good source for market prediction as it expresses the point of view and sentiment of opinion leaders and forms public opinion. The paper [2] is based on the impact of News headlines on FOREX. Once the seed words have been extended, both the seed words and extended words are utilized to arrange the conclusion of the news articles. Their exploratory results demonstrate that the utilization of the extended feeling words enhanced arrangement execution, which was further made strides by fusing their comparing intensities which brought about precision results to go from 52% to 91.5% by fluctuating the distinction of power levels from positive and negative classes from (0.5 to 0.5) to >9.5 individually.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

6) Text Mining: In foreign market exchange prediction, directional classification of news impact plays a major role. All aspects of Text mining need to be closely studied in the field of market prediction. In paper [1], the immediate impact of news stories on the stocks based on the Efficient Markets Hypothesis is studied. The tertiary movements on the stock price movements are identified by a novel piecewise linear approximation approach is the main advantage. Here more challenging and interesting issues, such as intraday stock price predictions, could not be achieved. A systematic framework for predicting the tertiary movements of stockprices can be achieved by analysing the news stories on the stocks. The paper [3] discusses on how money market news headlines can be used to forecast intraday currency exchange rate movements.

In this paper [1] the author studied, display the new multilingual variant of the Columbia News blaster news outline framework. The framework addresses the issue of client access to searching news from numerous dialects from various locales on the web. The framework naturally gathers, composes, also, compresses news in various source dialects, permitting the client to search news points with English outlines and think about points of view from diverse nations on the points.

In the paper [18] the author observed work that display SENTIWORDNET 3.0, a lexical asset expressly formulated for supporting supposition characterization and feeling mining applications. SENTIWORDNET 3.0 is an enhanced rendition of SENTIWORDNET 1.0, a lexical asset freely accessible for examination purposes, now right now authorized to more than 300 exploration bunches and utilized as a part of an assortment of exploration tasks around the world. Both SENTIWORD-NET 1.0 and 3.0 are the after effect of consequently expounding all WORDNET syn sets as indicated by their degrees of inspiration, antagonism and impartiality. SENTIWORDNET 1.0 and 3.0 contrast (a) in the renditions of WORDNET which they comment (WORDNET 2.0 and 3.0, individually),(b) in the calculation utilized for consequently expounding WORDNET, which now incorporates (furthermore to the past semi-managed learning step) an irregular walk venture for refining the scores. SENTIWORDNET 3.0 is particularly focussing on the enhancements concerning viewpoint that it epitomizes as for form 1.0.

In the paper [14] the author studied two novel Natural Language Processing (NLP) order methods are connected to the examination of corporate yearly reports in the errand of monetary gauging. The speculation is that printed content of yearly reports contain crucial data for surveying the execution of the stock throughout the following year. The primary technique depends on character ngram profiles, which are created for every yearly report, and afterward named taking into account the CNG order. The second strategy draws on a more conventional methodology, where clarity scores are joined with execution inputs and afterward supplied to a bolster vector machine (SVM) for grouping. Both routines reliably outflanked a benchmark portfolio, and their blend ended up being much more compelling and effective as the consolidated models yielded the most noteworthy returns with the least exchanges. SENTIWORDNET 3.0 against a piece of WORDNET 3.0 physically expounded for energy, pessimism, and lack of bias; these outcomes show exactness changes of around 20% as for SENTIWORDNET 1.0.

In the paper [6] the author demonstrated that the returns for substantial and mid-top stocks recorded on the New York Ex-change are most certainly not serially indigent. Interestingly, arrange lopsided characteristics on the same stocks are exceptionally tenacious from every day. These two exact certainties can be accommodated if complex financial specialists respond to request irregular characteristics inside of the engaging so as to exchange day in countervailing exchanges adequate toevacuate serial reliance over the everyday skyline. The example of intraday serial reliance, over interims extending from five minutes to 60 minutes, uncovers hints of effectiveness making activities. For the stocks in our example, it takes longer than five minutes for shrewd financial specialists to start such exercises. By thirty minutes, they are well along on their every day journey.

Principle commitment of paper [16] is the use of more expressive components to speak to message and the job of business sector criticism as a major aspect of our pledge choice procedure. In our study, we demonstrate that a vigorous Feature Selection permits lifting characterization exactness fundamentally above past approaches when joined with complex element sorts. That is on account of our methodology permits selecting semantically significant elements and

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

in this way, lessens the issue of over-fitting while applying a machine learning methodology. The system can be exchanged to some other application territory giving printed data and relating impact information.

Paper [5] study actualizes a disorder based model to anticipate the outside trade rates. In the first stage, the deferral direction inserting is utilized to remake the in secret stage space (or state space) of the conversion scale elements. The stage space shows the intrinsic crucial normal for the trade rate and is suitable for monetary displaying and gauging. In the second stage, part indicators such as bolster vector machines (SVMs) are built for estimating. Contrasted and customary neural systems, unadulterated SVMs or disarray based neural system models, the proposed model performs best. The root mean squared anticipating blunders are essentially lessened.

In the paper [17], a money related news feature operators is proposed to helping the financial specialists in choosing to purchase and to offer stocks in Taiwan market in the wake of getting the key ongoing news feature scattered by the operators. Weighted affiliation principles and content mining methods are utilized to infer the essentialness level of each recently arrived news feature on the vacillation of Taiwan Stock Exchange Financial Price Index on the following exchanging day. The test results uncover that the proposed work to be sure accomplishes critical execution and show its possibility in the utilizations of constant data dispersal, for example, money related news features through Internet.

II.CONCLUSION

Getting to the major information covered up in unstructured news content is a challenging issue. This paper reviewed various techniques used for feature selection, feature reduction,representation, sentiment analysis and machine learning algorithms useful for the context. News Headlines are always a good source to retrieve information which can be used for predicting Foreign Exchange Market. This paper discuss the problem occurred in the previous work and also discussed the work done by the different researchers in the existing system.

REFERENCES

[1] Aghdam, M. H., Ghasem-Aghaee, N., & Basiri, M. E. Text feature selection using ant colony optimization. Expert Systems with Applications,2009
[2] Anastasakis, L., & Mort, N., Exchange rate forecasting using a combined parametric and nonparametric self-organising modelling approach,2009
[3] 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.
[4] S., Pang, G., Wu, M., & Kuang, L, Categorization. Expert Systems with Applications.,2012.
[5] D.,& Wong, R. K. (2002) Currency exchange rate forecasting from news headlines
[6] Ani Nenkova1 and Kathleen McKeown, Automatic Summarization, University of Pennsylvania, USA, 2011.
[7] Jakub Piskorski and Roman Yangarber, Information Extraction: Past, Present and Future, Institute for Computer Science, Polish Academy of Sciences, Warsaw, Poland, 2013.
[8] C. Aggarwal, ChengXiang Zhai, IBM T. J. Watson Research Center Yorktown Heights, NY, 2010
[9] Nassirtoussi, Arman Khadjeh, et al. "Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment." Expert Systems with Applications 42.1 (2015): 306-324.
[10] Nassirtoussi, Arman Khadjeh, et al. "Text mining for market prediction: A systematic review." Expert Systems with Applications 41.16 (2014): 7653-7670.
[11] Yu, Yang, Wenjing Duan, and Qing Cao. "The impact of social and con-ventional media on firm equity value: A sentiment analysis approach." Decision Support Systems 55.4 (2013): 919-926.
[12] Fung, Gabriel Pui Cheong, Jeffrey Xu Yu, and Wai Lam. "Stock prediction: Integrating text mining approach using real-time news." Computational Intelligence for Financial Engineering, 2003. Proceed-ings. 2003 IEEE International Conference on. IEEE, 2003.
[13] Li, Feng. "The information content of forwardlooking statements in corporate filingsA nave Bayesian machine learning approach." Journal of Accounting Research 48.5 (2010): 1049-1102.
[14] Butler, Matthew, and Vlado Keelj. "Financial forecasting using character n-gram analysis and readability scores of annual reports." Advances in artificial intelligence. Springer Berlin Heidelberg, 2009. 39-51.
[15] Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market." Journal of Computational Science 2.1 (2011): 1-8.
[16] Huang, C.-J., Liao, J.-J., Yang, D.-X., Chang, T.-Y., & Luo, Y.-C. (2010). Realization of a news dissemination agent based on weighted association rules and text mining techniques. Expert Systems with Applications, 37, 64096413
[17] Premanode, B., & Toumazou, C. (2013). Improving prediction of ex-change rates using differential EMD. Expert Systems with Applications, 40, 377384
[18] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. "Senti-WordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." LREC. Vol. 10. 2010.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

BIOGRAPHY

Ms. Remya Vijayan received the BTech degree in Information Technology. She is pursuing post-graduation in Computer Engineering from Savitribai Phule PuneUniversity of Pune. At D. Y. Patil College of Engineering, Akurdi, Pune.

Ms. Soudamini S Pawar. received the ME degree in Computer Engineering from Savitribai Phule Pune University of Pune. She is currently working as an Assistant Professor in D Y Patil College of Engineering, Akurdi, Pune.