

BIG DATA ANALYTICS IN FINANCIAL MARKET

Kavitha S¹, Raja Vadhana P², Nivi A N³

¹Department of Computer Science and Engineering, Dr. N.G.P. Institute of Technology, Coimbatore, Tamil Nadu, India

²Department of Computer Science and Engineering, Dr. N.G.P. Institute of Technology, Coimbatore, Tamil Nadu, India

³Department of Computer Science and Engineering, Dr. N.G.P. Institute of Technology, Coimbatore, Tamil Nadu, India

Abstract

Big data plays a serious role within the business for creating higher predictions over business information that is collected from the real world. Finance is that the new sector wherever the big data technologies like Hadoop, NoSQL are creating its mark in predictions from financial data by the analysts. It's a lot of fascinating within the stock exchange choices which might predict on a lot of profits of stock exchange. For this stock exchange analysis each regular information and historical information of specific stock exchange are needed for creating predictions. There are varied techniques used for analyzing the unstructured information like stock exchange reviews (day-to-day information) and historical statistic of economic information severally. This paper involves discussion regarding the strategies that square measure used for analyzing each varieties of information.

Keywords: Big data, prediction, finance, stock market, business intelligence

1. INTRODUCTION

Big data means a lot in financial services in the transformation of the organization services, profits, etc. It is more promising for financial analysts and also the investors for their services and investments. The large information gathered over stock message boards is being large assets for nearly 71 percent of organizations and it also uses the historical time series data for accurate predictions of the stock market. It also gains new insights of financial organizations as well as investors. Big data technologies create a value for these types of data in financial market. Once the data are collected from message board (unstructured) they have to be classified based on the users sentiment on data in order to predict the correct results by integrating the historical data. The historical data is analyzed by financial volatility models. This paper deals with the models that can be used with predictive analytics of big data in financial market for better predictions.

1.1 Sentiment Analysis

Sentiment analysis is the process of identifying user's view from their reviews or feedback from social media. These reviews are only unstructured and it is handled by the tools of big data which implements the machine learning algorithm for analyzing sentiments. The Sentiment Analysis is focused when it comes to big data analytics because of the reviews of customers who are majorly involved in the benefit of the organization. The sentiment of the reviews is based on the subjectivity of the content. For example the phrase like "not bad" means "good", the words directly give positive feedback are "good", "amazing", etc. the negative feedbacks include "poor", "not good", etc., These are the features to be extracted from each and every statement of

the review and classified as polarity represented in figure 1. It is mainly used in fields like marketing and advertisement to improve the quality and profit of the product by various organizations. In financial analysis it is the reviews on stock exchange dash board from the investors or analysts or by any organizations.

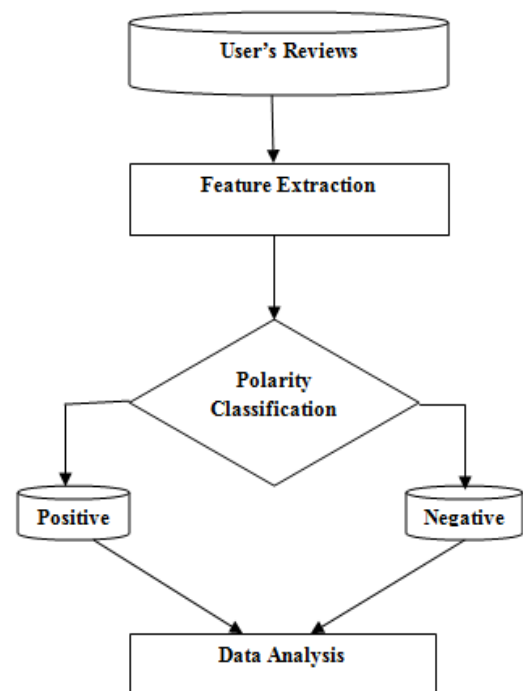


Fig -1 Sentiment Analysis

1.2 Financial Volatility

Even though online reviews are analyzed using sentiment analysis methods the analysis of financial volatility using traditional volatility methods is important to make accurate predictions on stock or other financial data. The veracity will be there in the mere analysis of online reviews and hence it requires historical data analysis for producing better result than complete uncertainty in data. In finance, volatility could be a measure for variation of worth of a financial mechanism over a period of time. Historic volatility springs from statistic of past market costs. An implied volatility springs from the value of a market listed spinoff (in explicit, an option). The symbol σ is employed for volatility, and corresponds to std. deviation, that shouldn't be confused with the equally named variance, that is instead the square, σ^2 . In stock market, the shares may go up and down day by day which is not consistent is said to be the volatility. The historic data can be structured but it holds a large amount of data that could be integrated with reviews.

2. SENTIMENT ANALYSIS

Sentiment analysis is the classification of sentiment features from the real life data such as comments posted on review boards. It plays a major role in big data analytics to provide predictive results with the machine learning algorithms. The sentiment is categorized into positive, negative and neutral [1]. There are two main tasks (1) product features are identified from the comments of reviewers and (2) the comments are classified as positive or negative. These are very challenging tasks. The classification can be done with document level, sentence level, phrase level, etc., using algorithms such as machine learning technologies such as supervised and unsupervised algorithms [2].

2.1 Unsupervised Machine Learning

Unsupervised learning is that of tries to search out the hidden structure in untagged knowledge. Because the examples provided to the learner are untagged, there's no error or reward signal to gauge a possible resolution. The unsupervised learning is distinguished from supervised learning and reinforcement learning when there is no error. Unsupervised learning traditionally uses the lexicon based approach for sentiment classification [3]. These methods uses sentiment lexicon to identify entire document's sentiment polarity [4] - [6].

2.1.1 Lexicon Based Approach:

It is the sentiment based on each word and phrase. It is identified by Turney [7] as semantic orientation of the reviews. Later the lexicon based approach is used for sentiment extraction. A lexicon based approach is a practical and easy approach for Sentiment Analysis of data without a requirement for training. A Lexicon based approach is good on how the lexicon is used. A lexicon-based approach is mainly projected to perform task using opinion bearing words (or simply opinion words).Opinion words are the commonly used words to express positive or negative opinions (or sentiments), For example, "good", "bad",

"poor" and "excellent". The count of positive and negative opinion words is used to determine the feature of the product in each sentence of the review. When the positive opinion words are more than the negative opinion words, then final conclusion on the feature will be positive or otherwise negative [8].

2.2 Supervised Machine Learning:

Supervised machine learning uses a trained label set to classify the sentiment on data. The training corpus is used for learning new classification of data [9]. A set of training examples will be there for training data. Each example consists of an input object (vector) and a desired output value (supervisory signal).A literature study shows often SVM yields higher accuracy results than other techniques [2].

2.2.1 Naïve Bayes Classifier:

Naïve Bayes classification method is simple method and comparatively produces good results with sensible accuracy. It uses a bag of words for classifying the subjectivity of the content. It is based on Naïve Bayes rule assuming conditional independence which is a main drawback of this classification [10]. For classifying a document d and class c using Bayes theorem is given by Eq. 1 and Eq. 2,

$$p\left(\frac{c}{d}\right) = \frac{p\left(\frac{d}{c}\right)p(c)}{p(d)} \quad (1)$$

Naïve Bayes Classifier,

$$c^* = \operatorname{argmax}_c P\left(\frac{c}{d}\right) \quad (2)$$

2.2.2 Support Vector Machine:

SVM is a classification and regression model used for data analysis. It constructs a set of hyper planes used for linear classification and regression shown in figure 2. It uses kernel mapping for non linear classification [11] which is used for erroneous classification. It produces more accurate results on both classification and regression than other machine learning techniques [2].

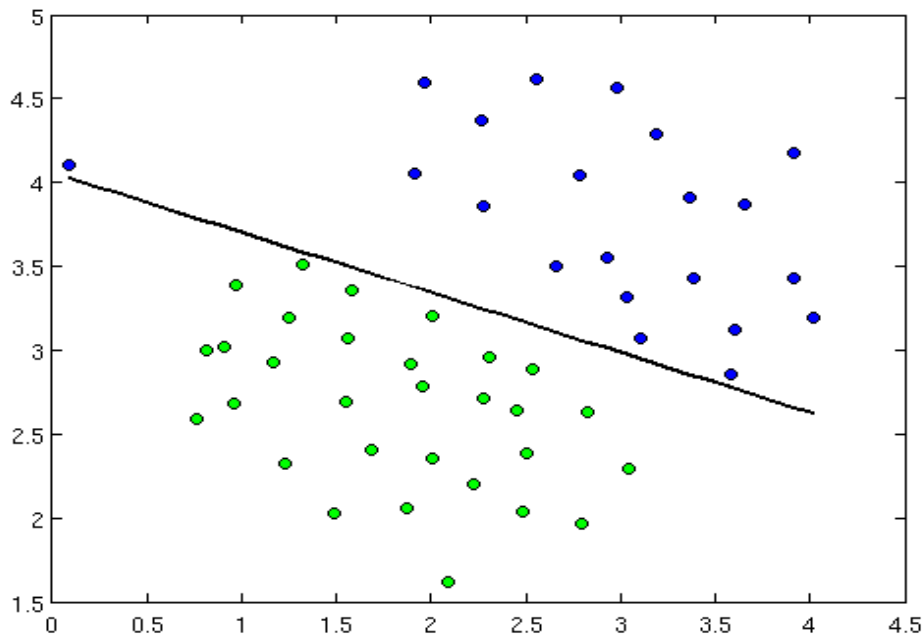


Fig -2 SVM Linear Classifier

3. FINANCIAL VOLATILITY

Financial volatility is predictable and deals with the historical data which is a time series model. These large data is handled with some classic methods of financial volatility to determine the future. It is very important for financial investments of investors [12]. Volatility is the change in prices of assets during particular period of time. It is measured with the standard deviation continuously with the specific period of time [13]. Gaussian process is the widely used process for this measure but the efficient method ARCH was first introduced by Engle [14]. Later it is extended to GARCH which is more persistent even other models are derived from it for the stock market volatility.

3.1 ARCH

Autoregressive Conditional Heteroscedasticity (ARCH) is introduced by R.F Engle[14] for the calculation of financial volatility. ARCH is a regression model to find the maximum likelihood estimation by the calculation of mean and variance V using Eq. 3.

$$V(\epsilon) = \sigma^2 \quad (3)$$

ARCH model uses conditional variance and the variance function can be generally expressed as in Eq. 4.

$$h_t = h(y_{t-1}, y_{t-2}, \dots, y_{t-p}, \alpha) \quad (4)$$

Where y is the variance, p is the order of ARCH process and α is a vector of unknown parameters

3.2 GARCH

Generalized Autoregressive Conditional Heteroscedasticity (GARCH) is the derived process of ARCH to allow the lagged conditional variance which is not allowed in ARCH process. This is also a learning mechanism which can be used for measuring stock volatility for efficient results than ARCH process [15]. In GARCH (p,q) process p, q is the order of process and GARCH (1,1) is more efficient and useful. It is given by Eq. 5.

$$h_t = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \beta_1 h_{t-1}, \quad \alpha_0 > 0, \alpha_1 \geq 0, \beta_1 \geq 0 \quad (5)$$

Where α and β are vector of unknown parameters

4. SVR BASED ON GARCH MODEL

SVR based on GARCH model is a review over using SVM and GARCH model for predictive analysis which is a part of big data predictive analytics which can predict the future with the day to day information and historical data of finance. The following review gives the results on these models showing the better predictions can be made for future in the application of finance.

4.1 Desheng Dash Wu, Lijuan Zheng, and David L. Olson [16]:

A popular financial website Sina Finance is used to collect the financial reviews. This data also identified the communication between investors are during opening and closing time of stock market. From this data the sentiments are classified using SVM where the chi square test is used to

compare the sentiment analysis techniques. The significance shows the accuracy that SVM gives better result than Naïve Bayes classifier. For the same period of time Stock index values (historical) data are downloaded to calculate the financial volatility using GARCH model. The volatility is calculated at both individual and industry level. GARCH-SVM predicts better volatility trend at individual stocks which keeps the parameters in order of GARCH terms. It is good for individual level than industry level because of the effect on stock features over sentiment of stock forum. This model works better for small organizations than larger organizations.

4.2 Fernando Perez-Cruz, Julio A Afonso Rodriguez and Javier Giner [17]:

In general, standard ML procedure is used to evaluate the parameters of the GARCH model. Here SVM is used to evaluate the parameters of GARCH model. In ML estimation the fit should be Gaussian distribution to produce best fit or else it provides more error terms. But SVM does not rely on previous data and also it tries to produce the best fit. SVM also does not produce best fit on Gaussian distribution because of pdf model. Here SVM is used as linear distribution not as nonlinear distribution which can be used in future. With linear distribution SVM evaluates the GARCH parameters to produce best fit.

4.3 Peter R. Hansen and Asger Lunde [18]:

In order to find the ability of volatility models in forecasting conditional variance, a large number of models are compared using the DM-\$ exchange rates and IBM stock returns. The volatility models are compared with its parameters like realized variance using first R observations by considering null hypothesis as the benchmark model. Nearly 330 ARCH models are compared but everything is outperformed by GARCH (1,1) model in terms of exchange rates. SPA test is used as the significance test to estimate the GARCH (1,1) model to make conclusion on selecting the volatility model. It is proven that in terms of IBM returns GARCH (1,1) is inferior but in terms of exchange rates, no other model can beat GARCH(1,1) model.

4.4 Robert P.Schumaker and Hsinchun Chen [19]:

A predictive machine learning technique is used to examine the financial news articles using textual representations. The two data sources, company generated and independently generated data sources are used. SVM is used to predict the stock specific variables and also for discrete numeric prediction. Here a new model called AZFinText system is designed for identifying the performance of future stock profit by closeness measure, accuracy and trading. Textual representation is good while using proper nouns and SVM is used to learn the change in share prices and adjust it due to the change of price severity. SVM produce better results than other machine learning techniques.

4.5 Jun Hua Zhao, Zhao Yang Dong, Zhao Xu, Kit Po Wang [20]:

The forecasting of electricity price is estimated using SVM. The estimation is done for electricity price series and its interval. SVM is the best regression technique used for forecasting price value and here non linear heteroscedastic model is required where GARCH model is linear and hence new Nonlinear Conditional Heteroscedastic Forecasting (NHCF) is introduced to evaluate the time changing variance. The electricity time series is estimated with Australian National Electricity Market data. Non linear SVM function is used to estimate the parameters of NHCF model and to find the maximum likelihood. The experiments are conducted to compare the GARCH and NHCH models where NHCF is good for non linear distribution like deregulating electricity market.

4.6 Reinaldo C. Garcia, Javier Contreras, Marco van Akkeren, and João Batista C. Garcia [21]:

The forecasting of electricity price is estimated with GARCH model. In historical time series analysis of data, the present and future values of the data is estimated using maximum likelihood. For non linear networks, selection of model is done by comparing ARIMA and GARCH model with time varying variance. Statistical significance and hypothesis tests are carried out to verify GARCH parameters such as auto correlation and residuals. The GARCH methodology is used to estimate the electricity prices for Spain and California electricity markets. The result proves that GARCH model outperforms the generalized time series model ARIMA. The extreme ability of the volatility model is also tested to prove that GARCH model can be used for non linear distribution like deregulating electricity market.

4.7 Altaf Hossain, Mohammed Nasser [22]:

The ARMA-GARCH, SVM, Relevance Vector Machine (RVM), Recurrent SVM (RSVM) are applied for estimating the financial volatility. The comparison of GARCH and ARMA-GARCH is made with RRVM and RSVM. For this evaluation India's Bombay Stock Exchange (BSE) SENSEX Index data and Japan's stock market NIKKEI225 is used. The prices of stock index are collected from Yahoo Finance. The collected data should be transformed log returns for the analysis. The models SVM and RVM are applied to GARCH for comparing GARCH and ARMA-GARCH model to forecast stock market volatility. DD, MSE, MAE and R^2 are used for evaluation. The conclusion states that ARMA-GARCH is better than pure GARCH model, relevance vector produced by RRVM is smaller than RSVM but both RRVM and RSVM performs equally. In terms of DS and MSE RRVM is better than RSVM whereas in terms of MAE, RSVM is better than RRVM. The result also proven RRVM is better than RSVM for forecasting stock market volatility.

4.8 Robert Engle [23]:

The usefulness of ARCH and GARCH models are described in financial forecasting. The standard deviation is calculated to find the usefulness of both the models. The weights for forecasting the variance is determined by the ARCH model. GARCH model also estimate the weights but it never decline it to complete zero. It is easy and simple form to estimate by conditional variances. Both the models are used in various applications but successful especially in financial forecasting due to the risk measurement. These models and their derivatives provide the storage to analyze and test the pricing and its change over time.

4.9 Pichhang Ou, Hengshan Wang [24]:

For classification and regression, RVM is used in forecasting financial time series. RVM uses Bayesian approach which is identical to SVM. Here GARCH(1,1) is used with RVM to forecast Shanghai composite Index. This GARCH-RVM is compared with other machine learning techniques such as recurrent Least Square SVM and RSVM and also basic GARCH(1,1) model. The predictive capability of RVM is better than LSSVM and SVM because of its larger memory to rely on previous series of information.

4.10 Wei Sun, Jian-Chang Lu, Ming Meng [25]:

SVM is used to forecast market clearing price (MCP) of electricity market. For this analysis nonlinear SVM function called kernel SVM is used. It is compared with RBF kernel of BP neural network forecasting to estimate its efficiency to predict MCP. In this analysis past price is not included except some influencing factors of electricity price. On comparison it is proved that SVM has lesser errors and better performance on generalization than BP neural network model.

5. CONCLUSION

In this paper we discussed about how the big data analytics especially predictive analytics is influenced in the financial market mainly in the stock exchange with its emerging technologies. The data analytics is also the main focused area on big data analytics from the volume, variety of data collected over the internet with its velocity but the value of data results in applications where the data could give some meaning. For its value it is analyzed to predict the future with predictive analytics of big data and veracity or uncertainty is something that always remains in any type of analysis. So the historical data is used here with some traditional techniques to give optimized predictions. For all these analysis some techniques and methods are unconditionally required to complete the required process. We mainly focused on such methods that are used to predict the future of the stock market using methods and algorithms for both day to day information and historical data. Among the entire methods the study shows that SVM algorithm is majorly used for sentimental classification because of its accurate results compared to other machine learning algorithms. Similarly in the financial volatility GARCH is

the widely used method for the historical or time series data analysis. The integration of SVM and GARCH model largely helps for the analysts of stock market and also for the investors to predict the stocks for more profit.

REFERENCES

- [1] Rudy Prabowo and Mike Thelwall, "Sentiment Analysis: A combined approach," *Journal of Informatics* 3 (2009) 143-157.
- [2] B. Pang, L. Lee, and S. Vaithyanathan. "Thumbs up? Sentiment classification using machine learning techniques". In *Proceedings of ACL*, pages 79–86, 2002.
- [3] xia.hu, jiliang.tang, huiji.gao, huan.liu, "Unsupervised Sentiment Analysis with Emotional Signals", WWW'13, May 13–17, 2013, Rio de Janeiro, Brazil. ACM 978-1-4503-2035-1/13/05.
- [4] B. O Connor, R. Balasubramanian, B. Routledge, and N. Smith. "From tweets to polls: Linking text sentiment to public opinion time series". In *Proceedings of ICWSM*, 2010.
- [5] J. Wiebe, T. Wilson, and C. Cardie. "Annotating expressions of opinions and emotions in language". *Language Resources and Evaluation*, 39:165–210, 2005.
- [6] T. Wilson, J. Wiebe, and P. Hoffmann. "Recognizing contextual polarity in phrase-level sentiment analysis". In *Proceedings of HLT and EMNLP*, 2005.
- [7] P. Turney. 2002. "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews". In *ACL-2002*.
- [8] M. Hu and B. Liu. "Mining and summarizing customer reviews". *KDD'04*, 2004.
- [9] Erik Boiy, Pieter Hens, Koen Deschacht, Marie-Francine Moens, "Automatic Sentiment Analysis in On-line Text", *Proceedings ELPUB2007 Conference on Electronic Publishing – Vienna, Austria – June 2007*.
- [10] Pablo Gamallo, Marcos Garcia, "Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets", *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 171–175, Dublin, Ireland, August 23-24, 2014.
- [11] Christopher J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition", 1–43, Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.
- [12] S.-H. Poon and C. W. J. Granger, "Forecasting volatility in financial markets: A review," *J. Econ. Lit.*, vol. 41, no. 2, pp. 478–539, 2003.
- [13] P. Ou and H. Wang, "Financial volatility forecasting by least square support vector machine based on GARCH, EGARCH and GJR models: Evidence from ASEAN stock markets," *Int. J. Econ. Fin.*, vol. 2, no. 1, pp. 51–64, 2010.
- [14] R.F. Engle, "Autoregressive conditional heteroscedasticity with estimates of variance of United Kingdom inflation," *J. Econometrica*, vol. 50, no. 4, pp. 987–1008, 1982.

- [15] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *J Econometr.*, vol. 31, no. 3, pp. 307–327, 1986.
- [16] Desheng Dash Wu, Lijuan Zheng, and David L. Olson, "A Decision Support Approach for Online Stock Forum Sentiment Analysis", *IEEE Transactions on Systems, Man, And Cybernetics: Systems*, Vol. 44, No. 8, August 2014.
- [17] Fernando Perez-Cruz, Julio A Afonso Rodriguez and Javier Giner, "Estimating GARCH models using Support Vector Machines", *Quantitative Finance Volume 3 (2003) 1-10*.
- [18] P. R. Hansen and A. Lunde, "A forecast comparison of volatility models: Does anything beat GARCH (1,1)?" *J. Appl. Econometr.*, vol. 20, no. 7, pp. 873–889, 2005.
- [19] R. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The AZFinText system," *ACM Trans. Inf. Syst.*, vol. 27, no. 2, pp. 1–19, 2009.
- [20] J.-H. Zhao, Y. D. Zhao, X. Zhao, and K. P. Wang, "A statistical approach for interval forecasting for the electricity price," *IEEE Trans. Power Syst.*, vol. 23, no. 2, pp. 267–276, May 2008.
- [21] Reinaldo C. Garcia, Javier Contreras, Marco van Akkeren, and João Batista C. Garcia, "A GARCH Forecasting Model to Predict Day-Ahead Electricity Prices", *IEEE Transactions On Power Systems*, Vol. 20, No. 2, May 2005.
- [22] Altaf Hossain, Mohammed Nasser, "Recurrent Support and Relevance Vector Machines Based Model with Application to Forecasting Volatility of Financial Returns", *Journal of Intelligent Learning Systems and Applications*, 2011, 3, 230-241.
- [23] Robert Engle, "GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics", *Journal of Economic Perspectives - Vol 15, Number 4- Fall 2001-Pages 157-168*.
- [24] Pichhang Ou, Hengshan Wang, "Predict GARCH Based Volatility of Shanghai Composite Index by Recurrent Relevant Vector Machines and Recurrent Least Square Support Vector Machines", *Journal of Mathematics Research Vol.2, No. 2, May 2010*.
- [25] Wei Sun, Jian-Chang Lu, Ming Meng, "Application of Time Series Based SVM Model on Next Day Electricity Price Forecasting under Deregulated Power Market", *Proceedings of the Fifth International Conference on Information Technology and Quantitative Management*.

BIOGRAPHIES



Kavitha S is currently pursuing her Masters in Computer Science and Engineering in Dr. N. G. P. Institute of Technology, Anna University, Coimbatore, Tamil Nadu, India. She was a Software Trainer in NIIT Pvt Ltd. in the year 2011. She worked as a Computer Instructor in Kendriya Vidyalaya, Coimbatore for elementary board for the year 2013. She is specialized in

.Net Framework and other areas of interest are Data Mining and Cloud Computing.
kvth.sgm@gmail.com



Raja Vadhana P is currently pursuing her Masters under the department of Computer Science and Engineering in Dr. N. G. P. Institute of Technology, Anna University, Coimbatore, Tamil Nadu, India. She was a Software Engineer and Associate Consultant in HCL Technologies Ltd. from 2008 to 2011. She worked as an Associate Consultant with Larsen and Toubro Infotech Ltd. under Enterprise Application Integration for the year 2011. She is specialized in BFSI domain and has extensive experience in Service Oriented Architecture technologies like TIBCO, ORACLE BPM and IBM BPM.



Nivi A N is currently pursuing her Masters in Computer Science and Engineering in Dr. N. G. P. Institute of Technology, Anna University, Coimbatore, Tamil Nadu, India. Her areas of interest are Data Mining and Cloud Computing.