2015 17th UKSIM-AMSS International Conference on Modelling and Simulation

Comment Volume Prediction using Neural Networks and Decision Trees

Kamaljot Singh*, Ranjeet Kaur
Department of Computer Science
DAV University, Jalandhar
Punjab, INDIA
Email: {kamaljotsingh2009, er.ranjeetsandhu}@gmail.com

Dinesh Kumar
Department of Information Technology
DAV Institute of Engineering and Technology
Jalandhar, Punjab, INDIA
Email: er.dineshk@gmail.com

***Abstract*—The leading treads towards social networking services had drawn massive public attention from last 'one and half' decade. The amount of data that is uploaded to these social networking services is increasing day by day. So, their is massive requirement to study the highly dynamic behavior of users towards these services. This is a preliminary work to model the user patterns and to study the effectiveness of machine learning predictive modeling approaches on leading social networking service Facebook. We modeled the user comment patters, over the posts on Facebook Pages and predicted that how many comments a post is expected to receive in next *H* hrs. To automate the process, we developed a software prototype consisting of the crawler, Information extractor, information processor and knowledge discovery module. We used Neural Networks and Decision Trees, predictive modeling techniques on different data-set variants and evaluated them under Hits@10(custom measure), Area Under Curve, Evaluation Time and Mean Absolute error evaluation metrics. We concluded that the Decision trees performed better than the Neural Networks under light of all evaluation metrics.**

Keywords–Neural Networks; RBF Network; Prediction; Facebook; Comments; Data Mining; REP Tree; M5P Trees.

I. Introduction

The leading trends towards social networking services had drawn massive public attention from 'one and half' decade. The merging up of computing with the physical things had enabled the conversion of everyday objects into information appliances[1]. These services are acting as a multi-tool with routine applications e.g.: news, advertisements, communication, commenting, banking, marketing etc. These services are revolutionizing day by day and much more are on the way. These all services have daily huge content generation in common, that is more likely to be stored on Hadoop clusters[2][3]. As in Facebook, 500+ terabytes of new data ingested into the databases every day, 100+ petabytes of disk space in one of FBs largest Hadoop (HDFS) clusters and their is 2.5 billion content items shared per day (status updates + wall posts + photos + videos + comments). The Twitter went from 5,000 tweets per day in 2007 to 500,000,000 tweets per day in 2013. Flickr features 5.5 billion images as that of Januray 31,2011 and around 3k-5k images are adding up per minute[4].

In this paper, we focused on the leading social networking service Facebook, in particularly 'Facebook Pages' (one of the product of Facebook), for automatic analysis of trends and patterns of users. So, for this work, we developed a software prototype that consist of crawler, Information extractor, information processor and knowledge discovery module. Our research is oriented towards the comment volume prediction(CVP) that a document is expected to receive in next *H* hours.

This is paper is organized as Section II discuss about the related works, Problem formulation is discussed in Section III. Section IV and Section V discusses about the experimental settings and Results. The paper is closed with Conclusion and Future work in Section VI, followed by Acknowledgment and References.

II. Related Works

The most closest related works to our research are: In the paper [5], the author had developed an industrial proof-of-concept demonstrating the fine-grained feedback prediction on Hungarian blogs using various prediction models and on variety of feature sets and evaluated the results using Hits@10 and AUC@10 measures. In the paper [6], the authors had modeled the relationship between content of political blog and the comment volume using Naive Bayes, Linear regression, Elastic regression and Topic-Poisson Models, and then evaluated them under the light of precision, recall and F1 measure.

In contrast to them, we haven't focused on Hungarian or on political blog, we focused on leading social networking service Facebook, For this research and targeted the state-of-the-art regression models like Multi-Layer Perceptron, RBF Network, M5P Tree and REP Tree.

III. Problem Formulation

We focused on fine grained predictive modeling techniques. For fine grained prediction, we address this problem as a *regression problem*. Given some posts that appeared in past, whose target values (comments received) are already known, we simulated the scenario. The task is to predict that

978-1-4799-8713-9/15 $31.00 © 2015 IEEE
DOI 10.1109/UKSim.2015.20

how many comments that a post is expected to receive in next *H hrs*. For this, we crawled the Facebook pages for raw data, pre-processed it, and made a temporal split of the data to prepare the training and testing set. Then, this training set is used to train the regressor and performance of regressor is then estimated using testing data(whose target value is hidden) using some evaluation metrics. This whole process is demonstrated in Figure 1 and detailed in this section.

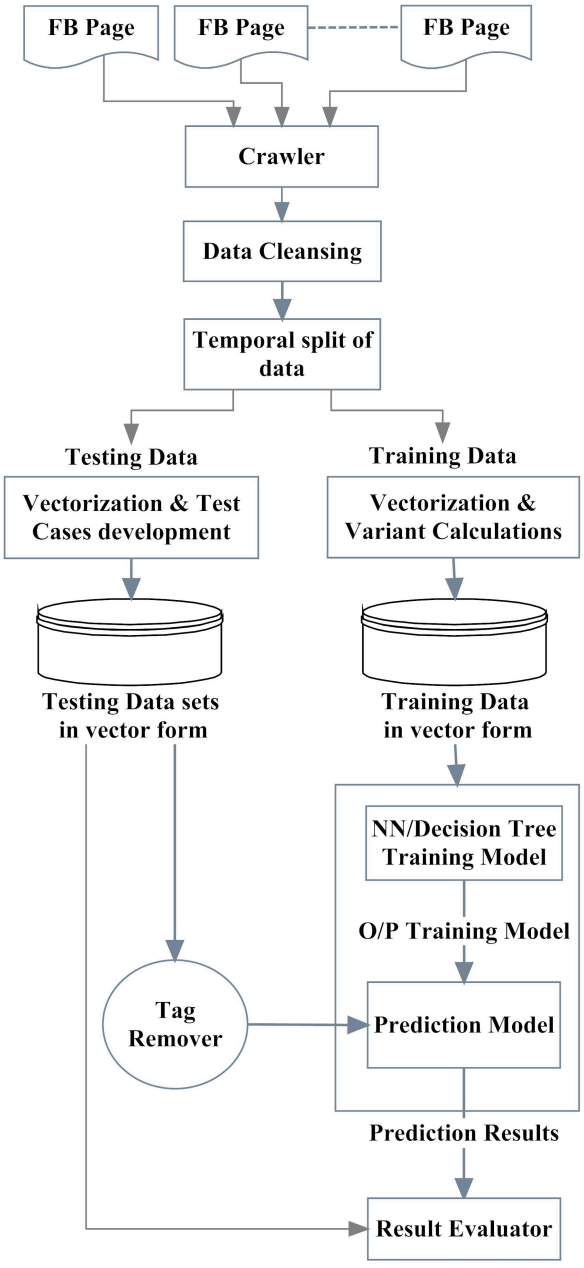

Figure 1. CVP Process Flow.

A. *Feature set used for this work*

We had identified 53 features, and 1 as target value for each post and categorized these features as:

1) *Page Features:* We identified 4 features of this category that includes features that define the popularity/Likes, category, checkin's and talking about of source of document. *Page likes :* It is a feature that defines users support for specific comments, pictures, wall posts, statuses, or pages. *Page Category :* This defined the category of source of document eg: Local business or place, brand or product, company or institution, artist, band, entertainment, community etc. *Page Checkin's :* It is an act of showing presence at particular place and under the category of place, institution pages only. *Page Talking About :* This is the actual count of users who are 'engaged' and interacting with that Facebook Page. The users who actually come back to the page, after liking the page. This include activities such as comments, likes to a post, shares by visitors to the page.
2) *Essential Features:* This includes the pattern of comment on the post in various time intervals w.r.t to the randomly selected base date/time demonstrated in Figure 2, named as C1 to C5.

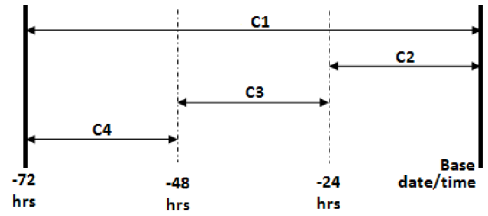

Figure 2. Demonstrating the essential feature details.

C1: Total comment count before selected base date/time. *C2:* Comment count in last 24 hrs w.r.t to selected base date/time. *C3:* Comment count is last 48 hrs to last 24 hrs w.r.t to base date/time. *C4:* Comment count in first 24 hrs after publishing the document, but before the selected base date/time. *C5:* The difference between C2 and C3. Furthermore, we aggregated these features by source and developed some derived features by calculating min, max, average, median and Standard deviation of 5 above mentioned features. So, adding up the 5 essential features and 25 derived essential features, we got 30 features of this category.

3) *Weekday Features:* Binary indicators$(0, 1)$ are used to represent the day on which the post was published and the day on selected base date/time. 14 features of this type are identified.
4) *Other basic Features:* This include some document related features like length of document, time gap between selected base date/time and document published

date/time ranges from $(0, 71)$, document promotion status values $(0, 1)$ and post share count. 5 features of this category are identified.

B. Crawling

The data originates from Facebook pages. The raw data is crawled using crawler, that is designed for this research work. This crawler is designed using JAVA and Facebook Query Language(FQL). The raw data is crawled by crawler and cleaned on basis of following criteria:

- We considered, only those comments that was published in last three days w.r.t to [1]base date/time as it is expected that the older posts usually don't receive any more attention.
- We omitted posts whose comments or any other necessary details are missing.

This way we produced the cleaned data for analysis.

C. Pre-processing

The crawled data cannot be used directly for analysis. So, it is carried out through many processes like split and vectorization. We made *temporal split* on this corpus to obtain training and testing data-set as we can use the past data(Training data) to train the model to make predictions for the future data(Testing data)[7][8]. This is done by selecting a threshold time and divide the whole corpus in two parts. Then this data is subjected to *vectorization*. To use the data for computations it is required to transform that data into vector form. For this transformation, we had identified some features as already discussed in this section, on which comment volume depends and transformed the available data to vector form for computations. The process of vectorization is different in training and testing set:

1) Training set vectorization: Under the training set, the vectorization process goes in parallel with the variant generation process. *Variant* is defined as, how many instances of final training set is derived from single instance/post of training set. This is done by selecting different base date/time for same post at random and process them individually as described in Figure 2. Variant - X, defines that, X instances are derived form single training instance as described in example of facebook official page id: 103274306376166 with post id: 716514971718760, posted on Mon Aug 11 06:19:18 IST 2014, post crawled on Fri Aug 15 11:51:35 IST 2014. It received total of 515 comments at time of crawling as shown in Figure 3.

[1]Base date/time, It is selected to simulated the scenario, as we already know what will happen after this. There is one more kind of time we used in this formulation: is the post published time, which comes before the selected base date/time.

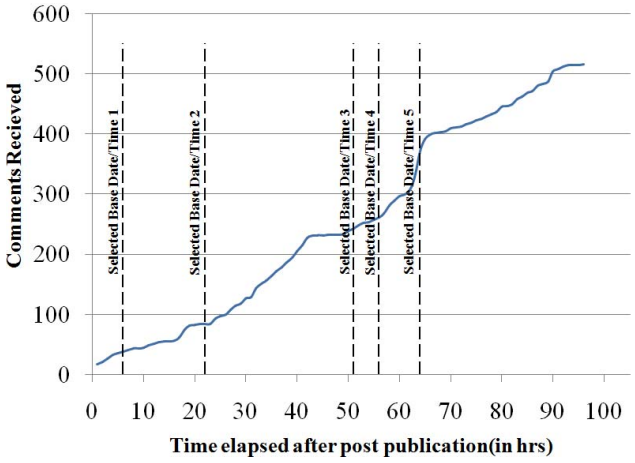

Figure 3. Cumulative Comments and different selected base date/time.

Know, by selecting different base date/time at random for single post, different variants are obtained for above example shown in Table I.

TABLE I
VARIANTS OBTAINED.

Variant	Selected Base Date/Time	Comments received in last 72 Hrs w.r.t Base Date/ Time	Comments target value
1	6	38	88
2	22	83	149
3	51	242	180
4	56	261	184
5	64	371	112

2) Testing set vectorization: Out of the testing set, 10 test cases are developed at random with 100 instances each for evaluation and then they are transformed to vectors.

D. Predictive Modeling

For the fine-grained evaluation, we have used the Decision Trees(REP Tree[9] and M5P Tree[10]) and Neural Networks(Multi-Layer Preceptron[11], RBF Network[12]) predictive modeling techniques.

E. Evaluation Metrics

The models and training set variants are evaluated under the light of Hits@10, AUC@10, M.A.E and Evaluation Time as evaluation metrics:

1) Hits@10: For each test case, we considered top 10 posts that were predicted to have largest number of comments, we counted that how many of these posts are among the top ten posts that had received largest number of comments in actual. We call this evaluation measure *Hits@10* and we averaged Hits@10 for all cases of testing data [5].

2) AUC@10: For the AUC [13], i.e., area under the receiver-operator curve, we considered as positive the 10 blog pages receiving the highest number of feedbacks in the reality. Then, we ranked the pages according to their predicted number of feedbacks and calculated AUC. We call this evaluation measure AUC@10. It is represented as:

$$AUC = \frac{T_p}{T_p + F_p} \quad (1)$$

where, T_p is True positive's and F_p is False positive's.
3) M.A.E: This measures defines that how close to actual comment volume, are the eventual outcomes. The mean absolute error is given by Equation 2.

$$M.A.E = \frac{1}{n}\sum_{i=1}^{n} \mid f_i - y_i \mid = \frac{1}{n}\sum_{i=1}^{n} \mid e_i \mid \quad (2)$$

where, f_i is the prediction and y_i the true value. All the test cases and all examples(n) are considered for this evaluation.
4) Evaluation time: It is the duration of the work performed describing the efficiency of the model.

IV. Experiment Settings

For our experiment, we crawled facebook pages collect the data for training and testing of our proposed model. In total 2,770 pages are crawled for 57,000 posts and 4,120,532 comments using JAVA and Facebook Query Language(FQL). The crawled data adds upto certain Giga bytes and this process of crawling had taken certain weeks. After crawling, the crawled data is cleaned(After cleansing 5,892 posts are omitted and we left with 51,108 posts).

We divided the cleaned corpus into two subsets using temporal split, (1) Training data(80%, 40988) and (2) Testing data(20%, 10120) and then these datasets are sent to preprocessor modules for preprocessing where:

1) *Training Dataset :* The training dataset goes through a parallel process of Variant calculations and Vectorization and as a result of training set pre-processing, we are obtained with these five training sets as:

TABLE II
Training set Variants Obtained.

Training Set Variant	Instances count
Variant - 1	40,949
Variant - 2	81,312
Variant - 3	121,098
Variant - 4	160,424
Variant - 5	199,030

2) *Testing Dataset :* Out of 10,120 testing data items, 1000 test posts are selected at random and 10 test cases are developed are described earlier.

The models that are used for experiments are Multi-Layer preceptron(MLP), RBF Networks, Decision Trees(REP Tree and M5P Tree). We used [2]*WEKA*((The Waikato Environment for Knowledge Analysis)) implementations of these regressors.

Neural Network - Multi Layer Perceptron Learning is used in 2 forms: (1)Single Hidden layer with 4 neurons. and (2) two hidden Layers, 20 neurons in 1^{st} hidden layer and 4 in 2^{nd} hidden layer. For both of the cases, the training iterations are fixed to 100, while the learning rate to 0.1 and momentum to 0.01. For Radial Basial function (RBF) Network, the cluster count is set to 90 clusters and default parameters are used for REP and M5P Tree.

HP Pavilion dv4-1241tx is used for this evaluation, whose configuration includes Windows 7 Operating System, Intel Core 2 Duo CPU with 2.00GHz 2.00GHz clock rate processors, with 3.00 GB of RAM and 320 GB of Hard Drive. Evaluation time, may vary on varying system configuration.

V. Result and Discussion

We had evaluated the models with several configurations and we are presenting results of some of them in this section. Table III, demonstrates the Hits@10, AUC@10, Evaluation Time & M.A.E on different implemented models. Moreover, the effect of different training set variants is also shown.

A. Hits@10

Hits@10 is one of the important accuracy parameter for the proposed work. It tells about the prediction accuracy of the model.

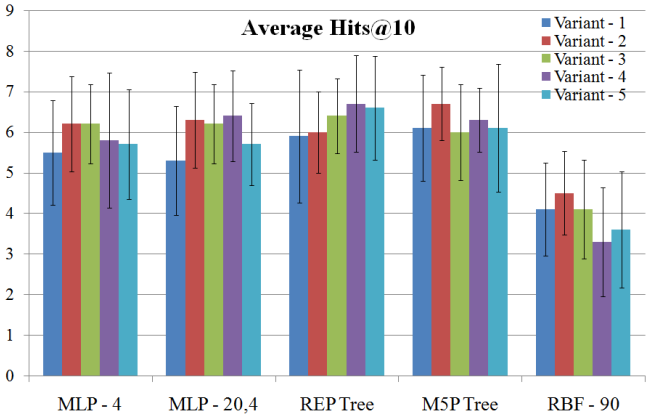

Figure 4. Hits@10 with Standard Deviation.

From the graph shown in Figure 4, it is clear that the prediction Hits@10 accuracy in case of Decision Trees(M5P Tree for Variant-2 and REP Tree for Variant - 4) is higher as compared to other modeling techniques and RBF Model shown minimal Hits@10 accuracy.

[2]WEKA is a tool for data analysis and includes implementations of data pre-processing, classification, regression, clustering, association rules, and visualization by different algorithms.

TABLE III
EXPERIMENTAL RESULTS.

Model		Variant - 1	Variant - 2	Variant - 3	Variant - 4	Variant - 5
MLP (4)	Hits@10	5.500 ± 1.285	6.200 ± 1.166	6.200 ± 0.980	5.800 ± 1.661	5.700 ± 1.345
	AUC@10	0.656 ± 0.164	0.807 ± 0.189	0.852 ± 0.180	0.795 ± 0.232	0.670 ± 0.205
	Time Taken	40.882 Sec	190.809 Sec	132.469 Sec	162.377 Sec	193.465 Sec
	M.A.E	47.699%	11.897%	7.921%	6.4622%	31.961%
MLP (20,4)	Hits@10	5.300 ± 1.345	6.300 ± 1.187	6.200 ± 0.980	6.400 ± 1.114	5.700 ± 1.005
	AUC@10	0.674 ± 0.157	0.831 ± 0.193	0.809 ± 0.206	0.832 ± 0.190	0.734 ± 0.205
	Time Taken	166.804 Sec	335.025 Sec	474.729 Sec	629.820 Sec	777.803 Sec
	M.A.E	42.23%	2.650%	0.962%	6.440%	18.56%
REP Tree	Hits@10	5.900 ± 1.640	6.000 ± 1.000	6.400 ± 0.917	6.700 ± 1.187	6.600 ± 1.281
	AUC@10	0.784 ± 0.127	0.827 ± 0.121	0.768 ± 0.109	0.807 ± 0.098	0.756 ± 0.137
	Time Taken	10.844 Sec	9.885 Sec	28.618 Sec	41.483 Sec	46.871 Sec
	M.A.E	21.650%	0.7334%	24.024%	16.170%	23.735%
M5P Tree	Hits@10	6.100 ± 1.300	6.700 ± 0.900	6.000 ± 1.183	6.300 ± 0.781	6.100 ± 1.578
	AUC@10	0.761 ± 0.143	0.708 ± 0.165	0.711 ± 0.165	0.693 ± 0.199	0.730 ± 0.185
	Time Taken	34.440 Sec	71.520 Sec	117.599 Sec	177.850 Sec	518.638 Sec
	M.A.E	15.430%	17.870%	26.650%	36.659%	14.737%
RBF Network (90 Clusters)	Hits@10	4.100 ± 1.136	4.500 ± 1.025	4.100 ± 1.221	3.300 ± 1.345	3.600 ± 1.428
	AUC@10	0.899 ± 0.110	0.912 ± 0.087	0.945 ± 0.083	0.937 ± 0.077	0.912 ± 0.086
	Time Taken	298.384 Sec	491.002 Sec	614.138 Sec	1831.946 Sec	1602.836 Sec
	M.A.E	16.56%	21.39%	29.39%	15.50%	17.63%

B. AUC@10

AUC@10 metrics tells about the prediction precision of the models.

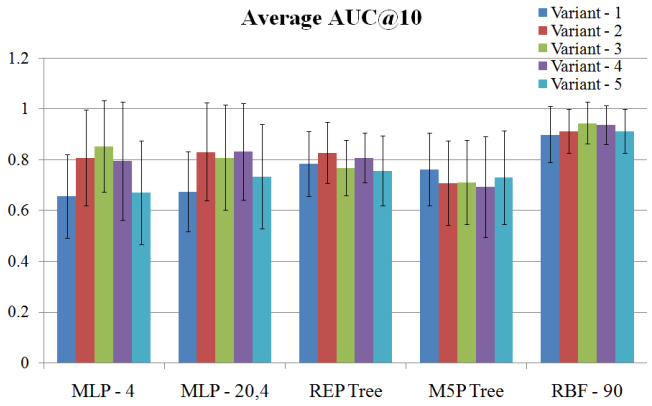

Figure 5. AUC@10 with Standard Deviation.

From the graph shown in Figure 5, it is depicted that the RBF Network had performed very well with 0.945 AUC@10 value of variant - 3 and it is higher among the other models. MLP-4 performed minimal with 0.656 AUC@10 variant - 1 value.

C. M.A.E

This measure shows the mean absolute prediction error produced by the models. The graph in Figure 6 depicts that the REP Tree is producing minimal error of 0.7334% under variant - 2 followed by MLP-20,4 with 0.962% under variant - 3. This is also depicted that the NN-MLP are producing very high error under variant - 1.

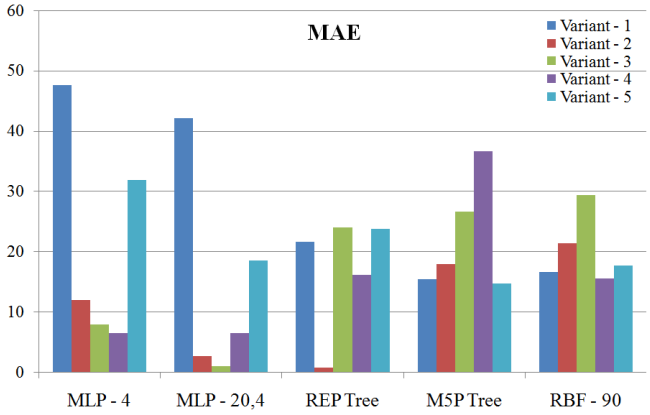

Figure 6. M.A.E.

D. Evaluation Time

This measure includes the time to train the regressor and to evaluate the test cases. From the graph in Figure 7, it is depicted that the REP Tree under Variant - 2 had predicted in minimal time of $9.885Sec$ followed by M5P Tree under Variant - 1 with $34.44Sec$ and the RBF Network had taken very high prediction time of $1831.946Sec$ under variant - 4.

VI. CONCLUSION AND FUTURE SCOPE

This paper examines the Neural Networks and Decision Tree's, came to the conclusion that Decision Trees outperforms very well than Neural Networks in our proposed comment volume prediction model. Moreover, with this examination we also shown that this model can be used for forecasting the comment volume perhaps choosing up of right variant is must. Our model is producing very good results, but their is further a room for improvement using more

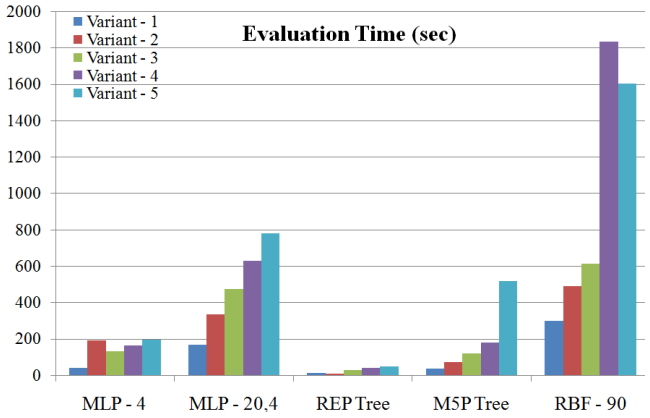

Figure 7. Evaluation Time.

features and with other regression techniques. The outcome of this work is a software prototype for comment volume prediction which can be further enhanced using category based predictor and by including multi-media features etc.

ACKNOWLEDGMENT

The authors would like to thanks Facebook for providing the necessary API's for data crawling, without which the proposed work was not feasible.

REFERENCES

[1] A. Kamilaris, A. Pitsillides, Social networking of the smart home, in: Personal Indoor and Mobile Radio Communications (PIMRC), 2010 IEEE 21st International Symposium on, 2010, pp. 2632–2637. doi:10.1109/PIMRC.2010.5671783.

[2] K. Shvachko, H. Kuang, S. Radia, R. Chansler, The hadoop distributed file system, in: Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on, 2010, pp. 1–10. doi:10.1109/MSST.2010.5496972.

[3] I. Polato, R. Ré, A. Goldman, F. Kon, A comprehensive view of hadoop researcha systematic literature review, Journal of Network and Computer Applications 46 (2014) 1–25.

[4] T. Reuter, P. Cimiano, L. Drumond, K. Buza, L. Schmidt-Thieme, Scalable event-based clustering of social media via record linkage techniques., in: ICWSM, 2011.

[5] K. Buza, Feedback prediction for blogs, in: M. Spiliopoulou, L. Schmidt-Thieme, R. Janning (Eds.), Data Analysis, Machine Learning and Knowledge Discovery, Studies in Classification, Data Analysis, and Knowledge Organization, Springer International Publishing, 2014, pp. 145–152. doi:10.1007/978-3-319-01595-8_16.

[6] T. Yano, N. A. Smith, What's worthy of comment? content and comment volume in political blogs., in: ICWSM, 2010.

[7] T. M. Pelusi, D., Optimal trading rules at hourly frequency in the foreign exchange markets, Mathematical and Statistical Methods for Actuarial Sciences and Finance, Springer (2012) 341–348ISBN: 978-88-470-2341-3.

[8] D. Pelusi, M. Tivegna, P. Ippoliti, Improving the profitability of technical analysis through intelligent algorithms, Journal of Interdisciplinary Mathematics 16 (2-3) (2013) 203–215.

[9] Y. Zhao, Y. Zhang, Comparison of decision tree methods for finding active objects, Advances in Space Research 41 (12) (2008) 1955–1959.

[10] E. Onyari, F. Ilunga, Application of mlp neural network and m5p model tree in predicting streamflow: A case study of luvuvhu catchment, south africa, in: International Conference on Information and Multimedia Technology (ICMT 2010), Hong Kong, China, 2010, pp. V3–156.

[11] D. Pelusi, Designing neural networks to improve timing performances of intelligent controllers, Journal of Discrete Mathematical Sciences and Cryptography 16 (2-3) (2013) 187–193.

[12] A. G. Bors, Introduction of the radial basis function (rbf) networks, in: Online symposium for electronics engineers, Vol. 1, 2001, pp. 1–7.

[13] S. M. Tan, P.N, V. Kumar, Introduction to data mining, Pearson Addison Wesley Boston, 2006.