Tax Complaints Classification on Twitter Using Text Mining

Prita Dellia[1] and Aris Tjahyanto[1]

***Abstract*—Twitter growth and utilization encourage the emergence of limitless textual information so that people can express their complaints easily This leads the Directorate General of Taxation uses twitter to deal with tax complaints faced by the community. However, the messages on twitter can contain any information, either the tax complaint or not. This will cause difficulties in handling complaints process. It is important to automatically identify so tax complaint handling can be done effectively and efficiently. Given these problems, it is necessary to do the twitter tax complaint classification with the support of text mining. There are several methods of classification such as Naïve Bayes classifiers, Support Vector Machine (SVM) and Decision Tree. This research aims to classify the tax complaint on twitter automatically by using text mining. The experimental results show the value of f-measure of SVM, Naïve Bayes and Decision Tree, respectively, are 89.3%, 85.6% and 76.9%.**

***Keywords*— Classification, Twitter, Tax Complaints, Text Mining.**

I. Introduction

Taxes as one of state revenues element, has a big role and increasingly dependable for the national development and spending interests. Taxes are an imposition of compulsory levies on individuals or entities by governments. Taxes are levied in almost every country of the world. A country needs economic development, which requires a relatively large funds. It is necessary for state revenue sources which potential. Other than that, national development which has been announced by the government aims to make the nation of Indonesia becomes an independent nation. Economic independence without the help of other countries is one parameter that is often seen in determining the position of a nation in the international relationship [1]. Indonesian government should be able to increase state revenue, one of which comes from taxes. One of the government's efforts to achieve that goal is by collecting taxes. Tax revenue contributed for 74.63% of all state revenues in Indonesia.

On the other hand, twitter growth and utilization encourage the emergence of limitless textual information. Twitter is a social networking and microblogging service that allows users to send and read text-based messages of up to 140 characters, known as “tweets”. Twitter users can send and receive the messages via a variety of mechanisms, including mobile phones, PCs, websites and desktop programs, and they are distributed in real time [2]. Tweets can express opinion on different topics, which can outbreaks of bullying [3], share consumer’s opinions concerning brands and products [4], acceptance or rejection of politicians [5], polarity prediction in political and sport discussion [6], event that generate insecurity [7], all in an electronic word-of-mouth way.

Indonesia became the third country in the use of twitter in the world after the US and Japan in 2013. Peoples can express their opinion on different topic via twitter easily, including opinions about the disappointment experienced.

The disappointment experienced by the public can be expressed in the form of complaints and lawsuits against the organization. It is because the public assumes that the quality they receive are not in accordance with expectations, giving rise to dissatisfaction or disappointment. Public complaint is an expression of public dissatisfaction caused by a product or a service [8]. One of the complaints expressed by the community via twitter is complaints about taxes. Management of complaint can be the fastest and easiest way to show products, services, systems or people that do not perform as expected. It also provides the ability test for organizations to fix mistakes [9]. This leads the Directorate General of Taxation uses twitter to deal with tax complaints faced by the community. However, the message on twitter can contain any information, either the tax complaint or not. This will cause difficulties in handling complaints process because twitter message must be manually sorted in which is a complaint among all twitter messages about taxes. It is important to automatically identify so tax complaint handling can be done effectively and efficiently.

Problems of tax complaint classification on twitter can be done with the support of text mining. Text mining can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools [10]. The purpose of text mining is to extract useful information from a collection of documents for a particular purpose. One part of the text mining is text classification. Many researchers who have done research on text classification, such as emotion classification [11][12], news classification [13], document classification [14], crime prediction [15] and sentiment analysis [16][17][18][19]. There are several

[1]Prita Dellia and Aris Tjahyanto are with Department of Information Systems, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia. E-mail: pritadellia@gmail.com; a[19]tjahyanto@gmail.com.

methods of text classification in text mining. This research focuses on Naïve Bayes, Support Vector Machine (SVM) and Decision Tree.

SVM has outperformed other machine learning algorithms for various text classification tasks [20][21][22]. SVM are capable of effectively processing feature vectors of some 10000 dimensions, given that these are sparse [23]. In other words, high generalization ability of the method makes it particularly suited for high dimensional data such as text. Besides these two classifier, Decision Tree are considered to be one of the most popular approaches for representing classifiers [24]. Decision Tree classifier is used for comparison with Naïve Bayes classifier and SVM classifier.

Text mining can assist management of complaints to organize and separate the contents of message on twitter as shown in Figure 1. This research aims to identify twitter messages that containing either the tax complaint or not automatically by using text mining. The rest of this paper is structured as follows: Section I gives an introduction and reason for this research. Steps and technique used in this research is described in Section II. Results and discussion described in Section III, includes experiments of three classifier and evaluation to show the effectiveness of classification methods. Finally, Section IV show the conclusions of this research.

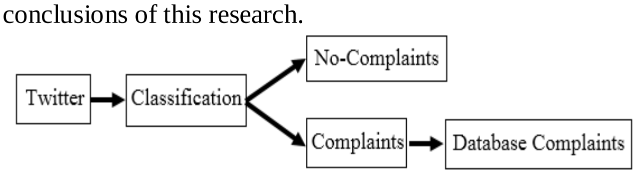

Figure 1. Management of complaints

II. METHOD

The research method goes through a number of stages. Figure 2 shows a general overview for the framework of typical classification system. As shown in Figure 2, this research is done mainly in three stage, preprocessing, classification, validation and evaluation, which are described in the next section.

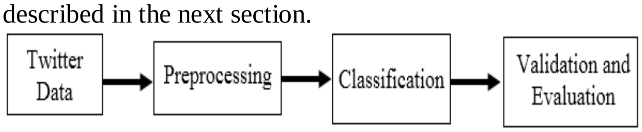

Figure 2. Research Methods

A. Data Collection

Twitter data is collected from the social networking twitter via Twitter API (Application Programming Interface) using the package "twitter" in R. The attribute in this study using twitter message content which known as tweets. The keywords used are '@DitjenPajakRI' and '@kring_pajak'. Therefore, the tweets containing those keywords will be drawn. Tweets collected are still mixed, containing both a complaints and not a complaint. In addition, that data collected contains keywords, including the hashtag (#) or mention (@) in the keyword.

B. Preprocessing

Twitter data have noisy text. It should be prepared first so that it can be used at next stage. The process of preparing the raw data is also called preprocessing. Preprocessing aims to transform unstructured text data into structured data.

Preprocessing stage is performed as follow:

1) Cleaning is remove hashtag, username, url and email on Twitter data collected.
2) Case folding is uniforms the letters to lowercase and remove the characters other than letters 'a' to 'z', including the removal of numbers and punctuation marks.
3) Tokenizing is break tweet into words using whitespace characters as breaker.
4) Filtering took the important words from the result of tokenizing step. This step will remove word that have no particular meaning using stopword dictionary.

C. Classification

Classification stage grouping twitter data into class complaints or not complaints. This stage uses three different classification method such as Naïve Bayes, Support Vector Machine (SVM) and Decision Tree. Brief description of classification methods in this research are as follows:

1) Naïve Bayes

The Naïve Bayes has been developed by C.T Yu and G. Salton [25] and also S. Roberson and K.Spark [26] in the 1970 respectively. The Naïve Bayes classifier is a probabilistic classifier that assumes the statistical independence of each feature (or word) and is a conditional model based on Bayes' formula [27][28]. This classifier estimates the probabilities that an object from each class falls in each possible discrete value of vector variable x [29]. Then, Bayes theorem is used to generate classification.

The number of probabilities that must be estimates are in order of $O(k^P)$ for *pk-valued* variables; when time *p* grows, the estimation becomes impractical. Appropriate independence assumption allows to approximate the full conditional distribution requiring $O(k^P)$ probabilities with a product of univariate distributions, requiring $O(k^P)$ probabilities per class. For m classes with $1 \leq k \leq m$, Naïve Bayes will be defined as [30]:

$$p(x|c_k) = p(x_1, \dots, x_p|c_k) = \prod_{j=1}^{p} p(x_j|c_k) \tag{1}$$

2) Support Vector Machine (SVM)

The SVM is a non-probabilistic classifier that works by constructing a decision surface on a high-dimensional space [31][32]. The principle of this algorithm is to find a decision surface, called hyperplane that optimally divides the training set. The training set is mapped into a high dimensional space. Algorithm find hyperplane in this space by looking at the largest margin, and then separating data into different groups.

For a training set with labeled pair $(x_i, y_i), i = 1,2, \dots$ where $x_i \in R^n$and $y \in \{1, -1\}^l$. the SVM method nedd to solve the following optimization problem, which can be presented as

$$min_{w,b,\xi} \frac{1}{2} W^T W + C \sum_{i=1}^{l} \xi_i$$

$$subject\ to\ y_i\ (w^T\phi(X_i) +\ b)\ \geq 1 -\ \xi_i, \quad (2)$$
$$\xi_i\ \geq 0$$

where 'W' is the weight parameter assigned to variables, ξ is the slack or error correction added and 'C' is the regularization factor [33]. Since the objectives of the problem is to minimize "$\frac{1}{2}W^TW + C\ \sum_{i=1}^{l}\xi_i$ ". Where value of "$y_i\ (w^T\phi(X_i) +\ b)$ " needs to be greater than "$1 - \xi_i$" and the value of 'ξ' is considered to be very small i.e., nearly equal to 0. Here training vector 'x_i' is mapped to higher dimensional space by ' '.

Since SVM requires input in the form of a vector of numbers, the reviews of text file for classification need to be converted to numeric value. After the text file is converted to numeric vector, it may go through a scaling process, which helps to manage the vectors and keep them in the range of [1, 0].

3) Decision Tree

Decision Tree is a classifier in the form of a tree structure. The Decision Tree has decision nodes and leaf nodes. The decision nodes check features of examples, while the leaf nodes will match the label for examples according to its features [34].

Decision Tree algorithm choose informative words based on the criteria of information gain, and predict the categories of each document according to the occurrence of word combinations in the document. Decision Tree classify examples by starting to initial decision node known as root node of Decision Tree. Root node contains a condition that is used to check one of examples's features. Then, this node select a branch according to that feature's example. The branch arrives at a new decision node with a new condition. This process goes on until it arrives at a leaf node which will provide a label for example.

D. Validation and Evaluation

Validation is to split the data into training data and test data. The validation process use the k-fold cross-validation, which uses k = 10 folds. It means twitter dataset is randomly divided into 10-fold. Each turn, one data fold is exploited for testing and the remaining folds are exploited for training. Teen-fold cross validation has becomes the standard method for validation process [35].

Performance of classification algorithm can be evaluated using parameters based on the confusion matrix [36] as shown in Table 1. True Positive (TN) means the number of tweets those are labeled as positive (complaints) and correctly classified as positive by classifier. False Positive (FP) indicates positive tweets, but classifier does not classify it as positive. True Negative (TN) represent tweets which are labeled as negative (no-complaints) and also classified as negative by classifier. False Negative (FN) are negative tweets but classifier does not classify it as negative. Those four parameters can be used to evaluate the performance of classification using the values of accuracy, precision, recall, and f-measure.

TABLE 1.
CONFUSION MATRIX

	Correct labels	
	Positive	**Negative**
Predict Positive	TP (True Positive)	FP (False Positive)
Predict Negative	FN (False Negative)	TN (True Negative)

Accuracy is common measure for classification performance. It is defined as the ratio of correctly classified example to the total number of examples. Accuracy can be calculated as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Precision is defined as the ratio between the numbers of examples correctly labeled as positive divided to the total number that are classified as positive. Precision is defined as follows:

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

Recall is defined as the ratio between the numbers of examples correctly labeled as positive divided on the total number of examples that truly are positive. Recall is defined as follows:

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

Many researcher use measure such as f-measure which combines precision and recall [37]. F-measure is defined as follows:

$$F - measure = \frac{2*Recall*Precision}{(Recall+Precision)} \quad (6)$$

Receiver Operating Characteristic Curve (ROC Curve) is depicted in a two-dimensional graph illustrating the performance of the classifier. Area under ROC curve is often used as a measure of quality of a probabilistic classifier. A random classifier has an area under curve 0.5, while a perfect classifier has 1 [38]. Classifier used in practice, so it should be between these values, preferably close to 1

III. RESULT AND ANALYSIS

A. Dataset

The data is a real-time data. Twitter data has been collected and categorized into complaint and not a complaint classes. Total tweet used is 1001 tweets with 758 tweets labeled as no-complaints and 243 tweets labeled as complaints.

B. Experimental Result

The stages in this research such as preprocessing, classification, validation and evaluation, assisted by WEKA software application. The preprocessing stage is done by using the filter function "*StringToWordVector*" in Weka. This filter convert a string attribute to a vector that represents word occurrence frequencies from the text contained in the string [38]. The result of this stage is then used in the classification stage. Experiments were conducted to test classification methods.

Experiments were performed using three different classification methods which are Naïve Bayes, Support Vector Machine (SVM) and Decision Tree. This experiment is used to examine the performance of classification methods. We use algorithm included in the WEKA data mining packages [38]. For Naïve Bayes, we use "NaiveBayes" algorithm which is standard probabilistic Naïve Bayes classifier. For SVM classifier, we use a polynomial kernel with the sequential minimal optimization "SMO" algorithm according to [39]. For Decision Tree, we use "J4.8" which reimplements C4.5 algorithm.

The validation process used same k-fold cross validation for tree classification method which used 10-fold. The learning procedure is executed a total of 10 times on diferrent training sets. Whereas the performance evaluation of classifier is done by using the measurement methods such as accuracy, precision, recall, f-measure and Receiver Operating Characteristics (ROC).

The performance results of all three classification methods are presented in Table 2. Table 2 shows that the SVM method has a high degree of accuracy, precision, recall and f-measure. It is the highest among the three methods of classification. SVM has an accuracy of 89,6%, followed by Naïve Bayes and Decision Tree with respectively 85,9% and 80,4%. Precision value for SVM 89,3%, while the Naïve Bayes 85,5% and Decision Tree 79,7%. Recall value for each classification method is SVM 89,6%, Naïve Bayes 85.9% and Decision Tree 80,4%. SVM still has the highest value in f-measure which amounted to 89,3%, then Naïve Bayes 85.6% and Decision Tree 76,9%. It is relevance with research in [21][22][23] that the SVM outperformed other various machine learning algorithms for text classification tasks, including the tax classification complaints on twitter. The experimental results contradictive with study in [14] because the performance of Naive Bayes can only seen through the value of ROC and not from value of overall performance evaluation. The highest ROC value owned by Naïve Bayes amounted to 0.888, followed by SVM 0.832 and Decision Tree 0.756. In contrast to studies [25], Decision Tree is considered incapable of representing the classifiers for this case when compared to the other methods. This is because the Decision Tree has the lowest value on the overall performance evaluation.

TABLE 2.
PERFORMANCE CLASSIFICATION

Method	Accuracy	Precision	Recall	F-Measure	ROC
Naïve Bayes	85,9%	85,5%	85,9%	85,6%	**0,888**
SVM	**89,6%**	**89,3%**	**89,6%**	**89,3%**	0,832
Decision Tree	80,4%	79,7%	80,4%	76,9%	0,756

IV. CONCLUSION

This research aims to automatically identify groups of twitter data which are classified as either complaints or not. It is important to do it so handling tax complaint can be done effectively and efficiently. The experimental results show that the automatic classification of a tax complaint on twitter can be done with the support of text mining. SVM classification method has the best value of accuracy, precision, recall and f-measure is respectively 89,6%, 89,3%, 89,6%, 89,3%. Whereas Naïve Bayes has the best value of ROC that is equal to 0,888.

The classification results of tax complaint data in this study was a collection of some types of complaints. Further research can classify the complaints data based on the type of complaint so that complaint can be solved according to the type of problem. Additionally, the performance of tax complaints classification can be improved by using feature selection method, i.e Markov Random Field (MRF), to determine which features are relevant.

ACKNOWLEDGEMENTS

The authors are thankful to LPDP (the scholarship thesis program).

REFERENCES

[1] I. T. Review, "1108: SPT Masa PPN Hard Copy," *Indonesian Tax Review*, vol. 7, no. 20, SMARTaxes Publishing, pp. 13–19, 2008.

[2] T. O'Reilly and S. Milstein, *The Twitter Book*, Second Edi. Sebastopol: O'Reilly Media, Inc., 2009.

[3] J. Xu, K. Jun, X. Zhu, and A. Bellmore, "Learning from Bullying Traces in Social Media," *Proc. 2012 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol.*, pp. 656–666, 2012.

[4] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 11, pp. 2169–2188, 2009.

[5] P. H. C. Guerra, A. Veloso, W. M. Jr, and V. Almeida, "From Bias to Opinion : a Transfer-Learning Approach to Real-Time Sentiment Analysis," *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, pp. 150–158, 2011.

[6] N. A. Diakopoulos and D. A. Shamma, "Characterizing Debate Performance via Aggregated Twitter Sentiment," *Chi2010 Proc. 28Th Annu. Chi Conf. Hum. Factors Comput. Syst. Vols 1-4*, pp. 1195–1198, 2010.

[7] M. Cheong and V. C. S. Lee, "A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter," *Inf. Syst. Front.*, vol. 13, no. 1, pp. 45–59, 2011.

[8] W. Reiboldt, "Factors That Influence a Consumer Complainer's Rating of Service Received from a Third Party Complaint-Handling Agency - the Los Angeles Department of Consumer Affairs," *J. Consum. Satisf. Dissatisfaction Complain. Behav.*, vol. 16, pp. 166–177, 2003.

[9] Suryadi, "Penanganan Keluhan Publik pada Birokrasi Dinas Perijinan," *Masyarakat, Kebud. dan Polit.*, vol. 23, no. 4, pp. 293–303, 2010.

[10] R. Feldman and J. Sanger, *The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press, 2007.

[11] N. Chirawichitchai, "Emotion classification of Thai text based using term weighting and machine learning techniques," in *2014 11th Int. Joint Conf. on Computer Science and Software Engineering: "Human Factors in Computer Science and Software Engineering" - e-Science and High Performance Computing: eHPC, JCSSE 2014*, 2014, pp. 91–96.

[12] Y. A. Sari, E. K. Ratnasari, S. Mutrofin, and A. Z. Arifin, "User Emotion Identification in Twitter Using Specific Features: Hashtag, Emoji, Emoticon, and Adjective Term," *J. Ilmu Komput. dan Inf. (Journal Comput. Sci. Information)*, vol. 7, no. 1, pp. 18–23, 2014.

[13] A. N. Chy, M. H. Seddiqui, and S. Das, "Bangla news classification using naive Bayes classifier," in *16th Int'l Conf. Computer and Information Technology, ICCIT 2013*, 2014, no. March, pp. 366–371.

[14] Y. Jiang, Q. Shen, J. Fan, and X. Zhang, "The classification for e-government document based on SVM," in *Proceedings - 2010 International Conference on Web Information Systems and Mining, WISM 2010*, 2010, vol. 2, pp. 257–260.

[15] X. Chen, Y. Cho, and S. Y. Jang, "Crime Prediction Using Twitter Sentiment and Weather," in *Systems and Information Engineering Design Symposium (SIEDS)*, 2015, vol. 0, no. c, pp. 63–68.

[16] H. A. Aldahawi and S. M. Allen, "Twitter mining in the oil business: A sentiment analysis approach," in *Proceedings - 2013 IEEE 3rd International Conference on Cloud and Green Computing, CGC 2013 and 2013 IEEE 3rd International Conference on Social Computing and Its Applications, SCA 2013*, 2013, pp. 581–586.

[17] D. Arora, K. F. Li, and S. W. Neville, "Consumers' sentiment analysis of popular phone brands and operating system preference using twitter data: A feasibility study," in *Proceedings - International Conference on Advanced Information Networking and Applications, AINA*, 2015, vol. 2015–April, pp. 680–686.

[18] L. Bing, K. C. C. Chan, and C. Ou, "Public sentiment analysis in twitter data for prediction of a company's stock price movements," in *Proceedings - 11th IEEE International Conference on E-Business Engineering, ICEBE 2014 - Including 10th Workshop on Service-Oriented Applications, Integration and*

Collaboration, SOAIC 2014 and 1st Workshop on E-Commerce Engineering, ECE 2014, 2014, pp. 232–239.

[19] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums," *ACM Trans. Inf. Syst. …*, vol. 26, no. 3, p. 12:1-12:34, 2008.

[20] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2002, pp. 79–86.

[21] A. Abbasi and H. Chen, "Analysis to Extremist- Messages," no. October, pp. 67–75, 2005.

[22] R. Zheng and J. Li, "A framework for authorship analysis of online messages: Writing-Style features and techniques," *J. Am. Soc. Inf. Sci. Technol.*, vol. 57, no. 3, pp. 378–393, 2006.

[23] E. Leopold and J. Kindermann, "Text Categorization with Support Vector Machines. How to Represent Texts in Input Space?," *Mach. Learn.*, vol. 46, no. 1, pp. 423–444, 2002.

[24] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*. Springer US, 2010.

[25] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, 2009.

[26] C. Horn, "Analysis and Classification of Twitter Messages," Graz University of Technology, 2010.

[27] J. Ortigosa-Hernández, J. D. Rodríguez, L. Alzate, M. Lucania, I. Inza, and J. A. Lozano, "Approaching Sentiment Analysis by using semi-supervised learning of multi-dimensional classifiers," *Neurocomputing*, vol. 92, pp. 98–115, 2012.

[28] X. Glorot, A. Bordes, and Y. Bengio, "Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach," in *Proceedings of the 28th International Conference on Machine Learning*, 2011, no. 1, pp. 513–520.

[29] A. Holts, C. Riquelme, and R. Alfaro, "Automated Text Binary Classification Using Machine Learning Approach," in *2010 XXIX International Conference of the Chilean Computer Science Society*, 2010, pp. 212–217.

[30] D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*. MIT Press eBooks, 2001.

[31] B. Gokaraju, S. S. Durbha, R. L. King, S. Member, and N. H. Younan, "A Machine Learning Based Spatio-Temporal Data Mining Approach for Detection of Harmful Algal Blooms in the Gulf of Mexico," *Ieee J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 4, no. 3, pp. 710–720, 2011.

[32] T. A. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing Contextual Polarity: an exploration of features for phrase-level sentiment analysis," *Comput. Linguist.*, vol. 35, no. 3, pp. 399–433, 2009.

[33] and C.-J. L. Chih-Wei Hsu, Chih-Chung Chang, "A Practical Guide to Support Vector Classification," *BJU Int.*, vol. 101, no. 1, pp. 1396–400, 2008.

[34] Y. Wang, W. Fu, A. Sui, and Y. Ding, "Comparison of Four Text Classifiers on Movie Reviews," in *2015 3rd International Conference on Applied Computing and Information Technology/2nd International Conference on Computational Science and Intelligence*, 2015, pp. 495–498.

[35] I. H. (Ian H. . Witten, E. Frank, and M. A. (Mark A. Hall, *Data mining : practical machine learning tools and techniques*. Morgan Kaufmann, 2011.

[36] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Syst. Appl.*, vol. 57, pp. 117–126, 2016.

[37] M. Farhoodi and A. Yari, "Applying machine learning algorithms for automatic Persian text classification Applying Machine Learning Algorithms for Automatic Persian Text Classification," in *International Conference on Advanced Information Management and Service (IMS)*, 2010, pp. 318–323.

[38] M. Vuk, "ROC Curve , Lift Chart and Calibration Plot," *Metod. Zv.*, vol. 3, no. 1, pp. 89–108, 2006.

[39] J.~Platt, *Fast Training of Support Vector Machines using Sequential Minimal Optimization*, no. April. 1999.