Text Mining Based on Tax Comments as Big Data Analysis Using SVM and Feature Selection

Mihuandayani
Magister of Informatics Engineering
Universitas AMIKOM Yogyakarta
Yogyakarta, Indonesia
mihuandayani20@gmail.com

Ema Utami
Magister of Informatics Engineering
Universitas AMIKOM Yogyakarta
Yogyakarta, Indonesia
emma@nrar.net

Emha Taufiq Luthfi
Magister of Informatics Engineering
Universitas AMIKOM Yogyakarta
Yogyakarta, Indonesia
emhataufiqluthfi@amikom.ac.id

***Abstract*— The tax gives an important role for the contributions of the economy and development of a country. The improvements to the taxation service system continuously done in order to increase the State Budget. One of consideration to know the performance of taxation particularly in Indonesia is to know the public opinion as for the object service. Text mining can be used to know public opinion about the tax system. The rapid growth of data in social media initiates this research to use the data source as big data analysis. The dataset used is derived from Facebook and Twitter as a source of data in processing tax comments. The results of opinions in the form of public sentiment in part of service, website system, and news can be used as consideration to improve the quality of tax services. In this research, text mining is done through the phases of text processing, feature selection and classification with Support Vector Machine (SVM). To reduce the problem of the number of attributes on the dataset in classifying text, Feature Selection used the Information Gain to select the relevant terms to the tax topic. Testing is used to measure the performance level of SVM with Feature Selection from two data sources. Performance measured using the parameters of precision, recall, and F-measure.**

Keywords—Text Mining; Tax Comments; Support Vector Machine; Feature Selection

I. INTRODUCTION

The economy of a country is strongly supported by the number of the country's State Budget. Sources of State Budget include Taxes, Non-Tax State Revenues (PNBP) and Grant receipts both from within and outside the country. Taxes became the largest source of contributions to the state treasury, accounting for 85.6% of all State Budgets [1]. The types of taxes that are often charged are the income tax (PPh), Value-Added Tax (VAT), Sales Tax on luxury goods (PPnBM), stamp duty, Land and Building Tax (PBB) and other taxes. In order to realize the target of the State Budget in the coming year, it is necessary to have various efforts and cooperation to improve the tax system in Indonesia both from the government, society or practitioner side. The policies of the Directorate General of Tax under the auspices of the Ministry of Finance continue to make efforts to increase public awareness about the importance of tax and improvement of the tax service system. Efforts made by the Directorate General of Taxation are the online complaint services and the development of tax applications including the State Acceptance Module (MPN-G2) such as e-billing, e-invoice, and e-filing SPT facilities. The growth of the internet and social networking today has made it easy for people to express their opinions. Big data also play its role because of the high velocity and the variety in every transactions or activity on the social networking. The Social media such as Twitter, Facebook, Instagram, and Path are used by the community as a tool to channel the opinions and conditions around. Since 2013, the use of social media such as Twitter and Facebook in Indonesia is in high categorization, it made the Directorate General of Taxes participate in utilizing the use of media through the official online account as an alternative media in conveying information to the community as well as a means of communication between the public with the Directorate General of Taxes. Complaints submitted by the public through Facebook and Twitter can be extracted into consideration in the evaluation of the quality of tax services. In addition, for further research, information in the form of public opinion results can be used as one of decision support for planning tax policies. The use of social media for the user has encouraged the increase of unlimited textual information so that there is a need to utilize textual data to be presented without reducing the value of the information. This can be done with text mining. Text mining is a text analysis where data sources are usually obtained from documents with the aim of searching for words that can represent the contents of a document so that interrelationships and inter-document classes [2] can be analyzed. It is used to know the pattern of issues and problems that occur in the community in real time so that it can be taken into consideration in preparing a more appropriate policy. Text mining can be done through classification (classifier) or just by looking at the frequency (word cloud) and followed by doing sentiment analysis [3].

In doing text processing required the use of classification methods such as Naïve Bayes (NB), Artificial Neural Network (ANN), and Support Vector Machine (SVM). NB is used to handle document classification problems through simple models so that the calculation of Naïve Bayes is easy and works well on large datasets. In addition, the hierarchical model of Naïve Bayes is considered to improve the efficiency of multi-grade text classification models [4]. ANN applied to the classification based on the extracted rule usually has a low error rate [5]. SVM has a better degree of accuracy in classification [6] and AdaBoost's combination with SVM can

978-1-5386-0954-5/18/$31.00 ©2018 IEEE

provide better generalization performance on an unbalanced class dataset [7]. SVM is a method that overcomes over machine learning, but one of the problems with text classification is the number of attributes used on a dataset [8]. Many attributes make accuracy low, even though the dynamic data needed a better technique to handle the dataset. To get better accuracy, the existing attributes must be selected with the right algorithm [9]. Feature selection is an important part of text processing, especially in the process of optimizing the performance of the classifier. Feature selection is based on a subset that works by minimizing features that are not relevant to the classification [10]. In this research, feature selection used Information Gain to calculate entropy value in the dataset. Information Gain is one of the most widely used feature selection criteria for classification applications [11].

This research proposed a text mining processing through SVM method with classification optimization with Feature Selection. Feature Selection is used to select the relevant feature of the dataset in order to get a better performance of SVM as a classifier. Text mining aims to generate a classification on the sentiment about the problem of taxation based on data sources the public comments on Facebook and Twitter. In this study, the results of positive and negative sentiments are based on time period and the type of tax data namely service, website system, and tax news. For further research, information generated from this text mining can be used as considerable of taxation and support services for future policies.

II. Related Work

Various research related to text processing is discussed using certain methods to optimize performance levels including improving accuracy and reducing the error rate in classification. One of them is in analyzing detection pattern with different dataset through SVM approach which is widely used as machine learning algorithm on sentiment analysis. The study used three different ratios between training data and test data that is 70:30, 50:50 and 30:70. The measured performance is from the level of precision, recall, and f-measure for each dataset. The study measures SVM performance that results in values that depend on the dataset as well as the ratio of training data and data testing [12]. In research on the use of data mining to find useful data from the World Wide Web repository is divided into Content Mining, Usage Mining and Structure Mining in the case of using text, images, audio and video on extract unstructured data information. The study used the Machine Learning approach to compare the accuracy of Naïve Bayes, Random Forest, and SVM algorithms. Of the 500 datasets used indicate the best accuracy with SVM algorithm is about 97.40% compared with Naïve Bayes and Random Forest algorithms [6]. Document categorization as the important issues of mining the text refers to the automatic classification of documents in a data class based on category or topic. In a study focused on the Machine Learning approach for automatic text categorization to be used on typical web structures. The study extracted data by the SVM method in partitioning documents into training datasets and data testing ratios of 60-40 and 80-20 accuracy improved in 80-20 dataset categories [13]. Related research on classification and feature selection in the Fast Moving Consumer Goods (FMCG) industry case study to identify relevant factors affecting the performance of customer loyalty classification. The study used a feature selection method that is chi-square as the best test in the selection of features with an accuracy of 83, 2% [10].

Sentiment analysis is often interpreted as a view or opinion or emotion in the form of text, sound, and images from social media such as Facebook, Twitter and various sources of websites. In doing sentiment analysis involves the process of classification of data in various classes be it positive, negative, and neutral. Therefore, classification plays an important role in Natural Language Processing. For example in international journals, proposed a variety of techniques used to analyze sentiment through machine learning including SVM, Rule-Based and Lexicon-based. Various methods performed through comparison of methods for sentiment analysis concluded that machine learning such as SVM and naïve Bayes have the best accuracy and can be a recommendation as a learning-base strategy [14]. Other research related to text mining in sentiment analysis that is discussing the utilization of Twitter data in a movie review. In the study used a method to extract features from the data source. Feature extraction is done in two stages by classifying the technique of the tweet engine correctly using the training data. Machine learning became the efficient technique which does not need the words in the database such as in knowledge base. By studying the highest accuracy model in the sentiments analysis of Bollywood or Hollywood movie, the highest accuracy in sentiment analysis using machine learning SVM and Naïve Bayes was obtained by 75% higher SVM accuracy and naïve Bayes of 65% [15].

Text processing is effectively utilized in the analysis of sentiments about a particular topic or product to know the reviews or recommendations of various public opinion. Research on text processing methods for sentiment analysis is continuously conducted to optimize the performance of text mining in obtaining the appropriate opinions as in studies that measure the effect of training data size using SVM and Naïve Bayes by forming two ensembles. The study states that the change in training set size does not significantly affect the level of classification accuracy using SVM or Naïve Bayes but by combining SVM and Naïve Bayes using AND-type fusion suggests increased accuracy and F-Score from SVM [16]. In addition, the development of research related techniques in managing data such as transaction data and customer interaction on social media is done to analyze various methods and analytical tools that can be applied in big data applications [17] and support decision makers to gain insights based on data extraction from the dataset.

III. Research Methodology

In general, the phases performed in this study consisted of identifying issues about tax topics, collecting data derived from social media, text processing, feature selection with Information Gain, classification using SVM and testing phase to measure the performance of algorithms and represent the results in text mining. Identification of problems is the first step to find out the main problems, tax functions, and tax business processes in order to determine what part or object of tax to know the sentence of public sentiment. Objects in question are

the public opinion on tax services including services in the tax office, tax news, and tax website system including e-billing system. In this stage, it is necessary to perform functional requirement analysis for text mining to work optimally to adjust the target of opinion. The requirements analysis can be explained in Table I.

TABLE I. REQUIREMENTS ANALYSIS

Requirements	Explanation
Crawling	Take comments using the Facebook and Twitter API
Case Folding	Change the word on tweets into lowercase and by cleaning the hashtag or symbol
Convert Emoticon	Converts emoticons into words to be interpreted
Tokenizing	Break a comment into a snippet
Filtering	Delete any irrelevant words
Stemming	Reduce every word to get the word base
Feature Selection	Select the relevant data subset
Classifier	Classify comments in positive and negative classes based on type
Diagram Visualization	Represents the results of classification in the form of diagrams

The data collection phase is performed to manage the dataset that is used as training data and data testing. The data used sourced from Facebook comments through the official account of the Directorate General of Taxes is DitjenPajakRI. In addition to data from Facebook, used Twitter comment data on the official account @DitjenPajakRI and @kring_pajak. The technique used for data retrieval is crawling using the Application Programming Interface (API) provided by the provider. After performing data collection, the next step is to do text processing on the dataset. Text processing is done through five stages: case folding, convert emoticon, tokenizing, filtering and stemming. Text processing is done to regulate the structure of words obtained from social media comments in order to facilitate the classification based on sentiment and tax object. The parts of text processing are explained on Table I. The feature selection phase used the Information Gain which is one of the most widely used feature selection criteria for text classification applications. The advantage of information is based on the theory of information theory by measuring how much class label information is obtained when observing the value of some features [11]. In information theory, the value of Entropy is information contained in several distributions such as the distribution of class P (c). Therefore, the Information Gain of some features f measures the value of Entropy P (c) that changes after observed f as the following equation:

$$IG(w) = -\sum_{c \in C} P(c) \log P(c) + \sum_{w \in \{0,1\}} P(w) \sum_{c \in C} P(c|w) \log P(c|w) \quad (1)$$

The relevant features in the information gain can be shown when the feature has a high value of information gain. In this study, Fast Correlation Based Filter (FCBF) used to identify which one the relevant features and redundancy. FCBF was effective to handle the feature redundancy in selection the feature [18]. The FCBF worked to select the set of feature and identify the high correlation to the class with symmetrical uncertainty $SU \geq \rho$ as follows:

$$SU(f_i, C) = 2\frac{IG(f_i, C)}{H(f_i) + H(C)} \quad (2)$$

To describe the implementation phase of this research in text mining is generally shown in Fig. 1.

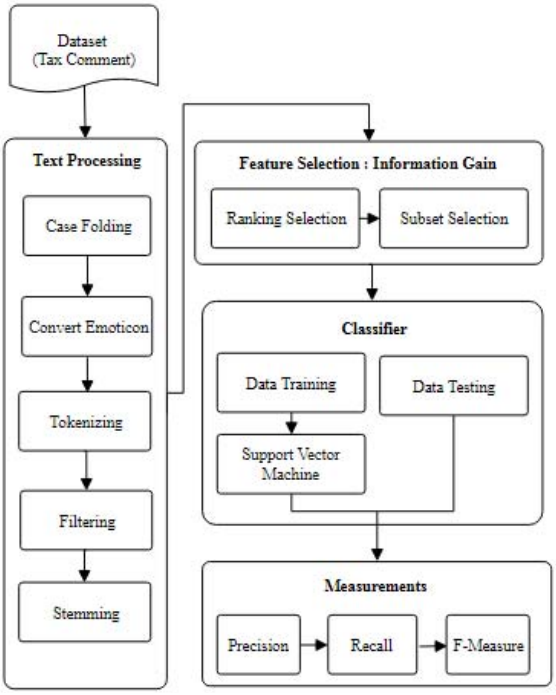

Fig. 1. Text Mining Flowchart

Based on Fig. 1 the next phase of selecting the relevant features is the classification phase using SVM algorithm which purposed to find the function of the separator (hyperplane) with the largest margin, so as to separate the two data sets optimally [19]. The SVM algorithm is very effective used to handle the text classification problem [20]. SVM is capable of working on high-dimensional datasets using the trick kernel. SVM uses only some of the selected data points that contribute (Support Vector) to form the model used in the classification process. In this study, SVM was used to find hyperplane that separated the data on positive comments and negative comments. SVM used the kernel to transform the input into the feature space or implement the model to a higher dimension so that the nonlinear case separable on the input becomes linear separable on a feature space. The kernel function used in this research is Radial Basis Function (RBF) kernel. SVM is defined through model stages Data points are known in SVM such as $x_i = \{x_1, x_2 \ldots x_n\} \varepsilon R_n$ with the data class $y_i \epsilon \{-1, +1\}$. Then paired the data and classes by condition $0 \leq \alpha_i \leq C$ to find the optimum SVM function as follows :

$$Opt = \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (3)$$

To calculate the value of w and b using the following formula:

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i \qquad b = -\frac{1}{2}(w.x^{+} + w.x^{-}) \tag{4}$$

After the value of w and b is known, the next step is to know the decision classification function sign (f(x)) as follows:

$$f(x) = \sum_{i=1}^{m} \alpha_i y_i K(x, x_i) + b \tag{5}$$

After obtaining training results with SVM in the form of positive and negative sentiments on tax cases based on services, website systems, and tax news. The results obtained are visualized in graphical form and analyzed to find out the results of performance level measurement on the application of text mining. The measurements used the precision, recall and F-measure parameter to know the performance result of the algorithm to the tax sentiment.

IV. Result And Discussion

In this section, presented the results of tax sentiments from two data sources based on the tax object. In addition, the classification displayed the test results with Precision, Recall and F-Measure on the use of SVM algorithm with Feature Selection. The datasets that have been collected through crawling data consist of Twitter data and Facebook data. The data is obtained from a period of 6 months during July to December 2017 or second semester of 2017. The comments of social media users about taxation using the Indonesian language. Various tax comments have been done by text processing and feature selection, categorized based on the type of tax object that has been defined previously ie service, website, and news. The kind of features which is used of the public comments related to the service, website and news for the unique term or word such as “Pajak”, “SSE”, “Billing”, “NPWP”, “DJP”, “Pph”, “Error”, “Lapor”, “SPT”, “Daftar”, and others about 258 selected features. Service included the comments related to the performance of services in the Tax Office (KPP) including the establishment of Taxpayer Identification Number (NPWP) and response to complaints. The website in question is related to the use of e-billing system, e-filling, and payment using the online system. While News is a comment or news related to taxation such as tax amnesty, tax abuse cases and so on. Sentiment analysis of positive and negative comments of the topic object is classified using SVM. Text mining data obtained in Facebook and Twitter can be displayed in Table II.

TABLE II. Dataset Used

Source Data	Service		Website		News	
	Positive	*Negative*	*Positive*	*Negative*	*Positive*	*Negative*
Facebook	714	652	491	478	540	708
Twitter	865	871	560	779	825	992

Based on the data in Table II, this study obtained data sources Facebook is 3583 data and 4892 of Twitter data for training. The dataset sources from Facebook and Twitter had been divided into training data and testing data. The ratio used for training and testing data is 80:20. In this study, the positive and negative sentiments of each data are also divided according to the time period in this case that is 6 months so that in each period the government can conduct an evaluation and see the development of public comments about the efforts to improve the service tax. By being able to monitor various sources of social media, service performance improvements can utilize this big data analysis technique. Furthermore, the tax sentiment chart with the target variables that are summarized and sourced from the comments of Facebook users can be illustrated in Fig. 2.

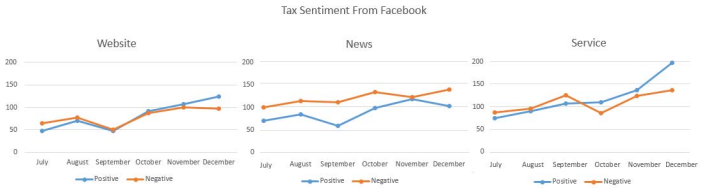

Fig. 2. Tax Sentiment of Facebook During 2nd Semester of 2017

Based on the graph in Fig. 2, the positive sentiments of Facebook users towards taxation, especially in the tax service section shows a significant value on the rising line. This means that during the months of taxation comments on the services of social media Facebook shows better performance or tax service. Next graph of taxation according to Twitter can be seen in following Fig. 3.

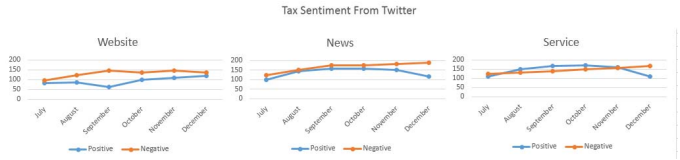

Fig. 3. Tax Sentiment of Twitter During 2nd Semester of 2017

In the graph in Fig. 3, the three graphs illustrated the negative sentiment has increased during the months in the website, news and service comments. From both sources of data monitored can provide an overview of the increasing or decrease in tax conditions perceived by the communities’ comments as consideration of the government. After sentiments are displayed, tests are performed to measure the performance level of each sentiment result, especially the use of the SVM with feature selection in the case of taxation. The parameters to evaluate the performance of algorithm are precision, recall and, F-measure. Precision means the percentage of predicted data as positive is correct or how much item relevant, recall means the percentage of positive data predicted as positive or how much the selected relevant item and F-measure value is obtained by weighting F = 2 to find out the result of evaluation calculation in information retrieval that combines recall and precision. The table of precision, recall and, F-measure from Facebook data can be shown in following Table III.

TABLE III. FACEBOOK DATASET MEASUREMENTS

Type of Comment	Target Variable	Precision	Recall	F-Measure
Website	Positive	0.68	0.57	0.62
	Negative	0.81	0.77	0.79
News	Positive	0.68	0.62	0.65
	Negative	0.79	0.72	0.75
Service	Positive	0.82	0.79	0.80
	Negative	0.75	0.69	0.72
Average		0.76	0.69	0.72

Based on the data in Table III, the highest precision and recall values were 82% and 79% in service and positive sentiment respectively. Besides, the F-measure showed the highest is 80 %. The average of all precision, recall, and F-measure from Facebook data were 76 %, 69%, and 72% respectively. For the precision, recall, and F-measure calculation results from Facebook data are shown in Fig. 4.

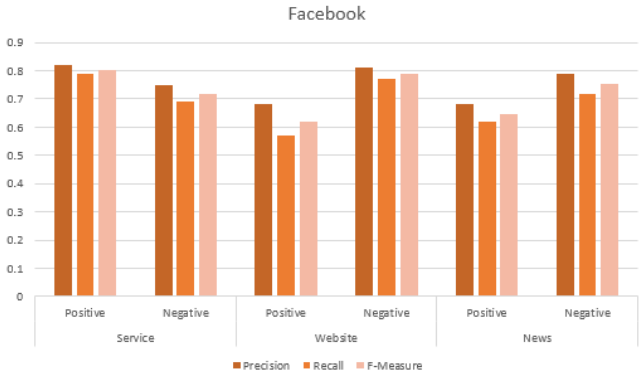

Fig. 4. Facebook Precision, Recall and F-measure

In addition to data from Facebook, testing is also done on Twitter data which is displayed on the table of precision, recall, and F-measure in following Table IV.

TABLE IV. TWITTER DATASET MEASUREMENTS

Type of Comment	Target Variable	Precision	Recall	F-Measure
Website	Positive	0.68	0.64	0.66
	Negative	0.69	0.68	0.68
News	Positive	0.82	0.79	0.80
	Negative	0.84	0.85	0.84
Service	Positive	0.77	0.72	0.74
	Negative	0.66	0.56	0.61
Average		0.74	0.71	0.72

The results of precision, recall, and F-measure based on Twitter data showed the highest value for precision and recall respectively 84% and 85% in negative news sentiment comments. Besides, the F-measure is about 84 %. The average of all precision, recall, and F-measure from Twitter data were 74 %, 71%, and 72% respectively. In addition, the measurement graph of Twitter is illustrated in Fig. 5.

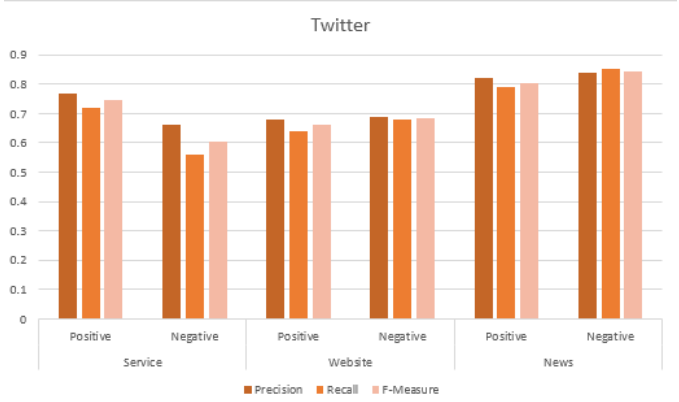

Fig. 5. Twitter Measurements

After measuring the algorithm of the tax comments sourced from Facebook and Twitter, need to conclude on the performance of systems of the classification with SVM and feature selection using Information Gain. The performance algorithm which is measured by the average of precision, recall and F-measure from Facebook and Twitter data can be calculated to know the overall performance of proposed method. The precision, recall and F-measure average of all dataset were 75 %, 70% and 72% achieved.

V. CONCLUSION

Text mining in this research is obtained through the stages of text processing, feature selection using Information Gain and classification using SVM. The datasets obtained are derived from Facebook and Twitter comments on taxes. Classification results based on positive and negative sentiments with three categories namely service, website, and news. From the measurement of precision, recall and F-measure obtained on the dataset from Facebook and Twitter showed the results of the performance algorithm that the average of the proposed method about 75%, 70% and 72% achieved. This research can be used as a basis for big data analysis in tax case an evaluation of tax service based on public opinion.

ACKNOWLEDGMENT

Universitas AMIKOM Yogyakarta for the support by giving all the facility which is needed for the study. This research also can be done by a good cooperation among all sectors and big effort to provide a paper which hopes can be useful for others.

REFERENCES

[1] Admin. (2017, October 15), State Revenue Realization - Ministry of Finance of the Republic of Indonesia. Available online:www.bps.go.id

[2] Leismester, C. 2015. *Mastering Machine Learning with R*. Published by Packt Publishing Ltd. Livery Place 35 Livery Street. Methods. USA: A Wiley-Interscience Publication.

[3] Pathak, M, A. 2014. *Beginning Data Science with R*. Springer International Publishing Switzerland 2014.

[4] Jadon, E., Sharma R. *Data Mining: Document Classification using Naive Bayes Classifier*. International Journal of Computer Applications Volume 167 - No. 6, June 2017.

[5] *Lu, H., Setiono, R. Liu, H. NeuroRule: A Connectionist Approach to Data Mining*. 2017. Available online : arXiv:1701.01358v1 [cs.LG].

[6] Sheshasaayee, A. & Thailambal G. *Comparison of Classification Algorithms in Text Mining*. International Journal of Pure and Applied Mathematics 2017, Vol. 116 No.22 pp 425-433.

[7] Dedhia, C and Ramteke, J. *Ensemble model for Twitter Sentiment Analysis*. International Conference on Inventive Systems and Control 2017.

[8] Wang, S., Li, D., Song, X., Wei, Y., & Li, H. (2011). *A feature selection method based on improved fisher's discriminant ratio for text sentiment classification*. Expert Systems with Applications, 38(7), 8696–8702. doi:10.1016/j.eswa.2011.01.077.

[9] Xu, T., Peng, Q., & Cheng, Y. (2012). *Identifying the semantic orientation of terms using S-HAL for sentiment analysis.* Knowledge-Based Systems, 35, 279–289. doi:10.1016/j.knosys.2012.04.011.

[10] Sulistiani, H., & Tjahyanto, A. *Comparative Analysis of Feature Selection Method to Predict Customer Loyalty*. Journal of Engineering, Vol. 3, No. 1, 2017 (eISSN:2337-8557).

[11] Croft, W.B, Metzler D, and Strohman T. 2015. *Search Engines: Information Retrieval in Practice.* Pearson Education.

[12] Ahmad, Munir, & Shabib Aftab. *Analyzing the Performance of SVM for Polarity Detection with Different Datasets*.International Journal Modern Education and Computer Science. DOI: 10.5815/ijmecs.2017.10.04.

[13] Fatima, S., & Srinivasu, B. *Text Document Categorization using Support Vector Machine*. International Research Journal of Engineering and Technology (IRJET). 2017, Vol. 4 Issue 2.

[14] Kathuria, A. & Upadhyay S. *A Novel Review of Various Sentimental Analysis Techniques*. International Journal of Computer Science and Mobile Computing (IJCSMC). 2017, Vol. 6 Issue 4 pp. 17-22.

[15] Amolik, A., Jivane N., Bhandari M., Venkatesan M. *Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques*.International Journal of Engineering and Technology (IJET), 2016, Vol. 7 No. 6.

[16] Abdelwahab, O., Bahgat, M., Christopher J. Lowrance1, C.J, & Elmaghraby, A. *Effect of Training Set Size on SVM and Naïve Bayes for Twitter Sentiment Analysis*. 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT).

[17] Elgendy N., Elragal A. 2014. *Big Data Analytics: A Literature Review Paper*. In: Perner P. (eds) Advances in Data Mining. Applications and Theoretical Aspects. ICDM 2014. Lecture Notes in Computer Science, vol 8557. Springer, Cham.

[18] Tang, Jiliang and Alelyani, Salem and Liu, Huan. 2014. Feature selection for classification: A review. In: Data Classification: Algorithms and Applications. CRC Press, p. 37.

[19] Jiawei, H., Kamber, M., & Pei, J. 2012. *Data Mining: Concepts and Techniques Third Edition*. Waltham, MA: Morgan Kaufmann.

[20] Ronen, F, & James, S. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge University Press.2006.