International Journal of Research in Advent Technology, Vol.7, No.5S, May 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

An Overview and Survey on Techniques Used for Text Mining

[1]K.Meena, [2]R.Lawrance
[1]Research Scholar,
Research and Development Centre,
Bharathiar University, Coimbatore, India
[2]Director, Department of Computer Applications,
Ayya Nadar Janaki Ammal College, Sivakasi, India
msdmeena@gmail.com,lawrancer@yahoo.in

Abstract - Data mining is used to extract important information from the data sources. Text mining is used to retrieve useful information from the textual databases. Text mining techniques used to classify the new documents into predefined classes. Researchers are developing a new algorithm for classifying the text document or modifying the existing algorithm with their ideas to produce the better results than the others algorithm. This survey discuss about the various researchers work related to text mining, text preprocessing, text clustering and text classification. Accordingly, the literature review is separated into three parts. They are pre processing, text clustering and text classification. This survey provides an idea about how to retrieve the significant word from the text document and also guide the researchers to improve the accuracy of text document clustering and classification algorithms.

Keywords – Text mining; Text clustering; Text classification

1. INTRODUCTION

Text mining is the procedure of obtaining valuable information from unstructured text. It processes the unstructured input then converts it into structured one, from the structured text obtain patterns then performs evaluation and interpretation of the output. Text analysis involves pattern recognition, information extraction, tagging, visualization, lexical analysis, predictive analysis and information retrieval. Text mining also termed as text analytics. Text analytics deals with the problems in the business because most of the business information available in the form of unstructured format. In Text Mining, patterns are extracted from natural language text. It may be freely characterized as the process of analyzing text to extract information that is useful for particular purposes. When compared with the kind of data stored in databases, text is unstructured, amorphous, and difficult to deal with algorithmically. Text mining encompasses web mining, information retrieval, computational linguistics and natural language processing. Text mining or knowledge discovery from text deals with the machine supported analysis of text. It uses techniques from information retrieval, information extraction as well as natural language processing (NLP) and connects them with the algorithms and methods of knowledge discovery in databases (KDD), data mining, machine learning and statistics.

In recent years, applications of text mining are widely used in the fields of multimedia, biomedical, patent analysis, anti-spam filtering of emails, linguistic profiling and opinion mining etc. To fulfill these requirements, it is necessary to analyze the content of text in depth. In order to extract useful patterns from unstructured text, various tasks such as text preprocessing, text transformation, attribute selection, pattern discovery and evaluation are performed on it. To solve the problems related to text, several researchers proposed various algorithms.

Text preprocessing means split a document into a stream of words by removing all punctuation marks and by replacing tabs and other non-text characters by single white spaces. This tokenized representation is then used for further processing. Text can be transformed by using filtering and stemming methods. A standard filtering method is stop word filtering. The idea of stop word filtering is to remove words that allow little or no content information, like articles, conjunctions, prepositions, etc. Furthermore, remove words that occur extremely often and also words that occur very rarely are likely to be of no particular statistical relevance. Stemming methods try to build the basic forms of words, i.e. strip the plural ’s’ from nouns, the ’ing’ from verbs, or other affixes. A stem is a natural group of words with equal or very similar meaning. After the stemming process, every word is represented by its stem. To further decrease the number of words that should be used, indexing or keyword selection algorithms can be used. In this case, only the selected keywords are used to describe the documents. A simple method for keyword selection is to extract keywords based on their entropy. The entropy gives a measure how well a word is suited to separate documents by keyword search. For instance, words that occur in many documents will have low entropy. The entropy can be seen as a measure of the importance of a word in the given domain context. The basic method of representing a document is, by considering it an element in a vector space. Each component of the vector is the frequency of occurrence of a word in the document. The size of the vector can be reduced by selecting a subset of most important words according to some criterion. Several vector space models such as term frequency, TF_IDF, probabilistic topic model, latent semantic analysis(LSA), hyperspace analog to language(HAL), correlated occurrence analog to lexical semantics(COALS) and high dimensional explorer(HIDEx) etc are useful for construct document vector. Then the text mining process merges with the traditional data mining process. Classic data mining techniques such as association rule mining, classification and clustering etc are used on the document vector that resulted from the previous stages. Results obtained from the mining techniques are well-suited for preferred application. Text mining processing steps are shown in Figure 1.

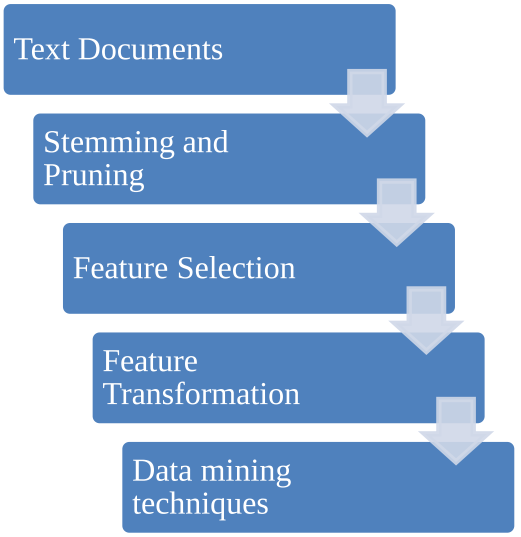

Fig. 1. Text mining processing steps

T.Yukiko and A.Yumi[1] illustrates how text mining is important and effective technique for analyze the users accent in the social media. Social media allow users to share their information. Text mining is used to find the frequent word and also used to derive the opinion of the shared information then the opinions were compared using various approaches. Hence the author concludes that the text mining is a valuable method used for take advantage of users' tone of voice.

Text emotion investigation has become a prosperous extremity in the text mining. Li et al. [2] proposed an algorithm to analyze the online forums and detect the poignant divergence of a text then use text mining approaches. This proposed algorithm is pooled with Support Vector Machine (SVM) and K-means clustering and the results shows that the SVM provides more consistent result than K-means clustering.

Mining useful patterns from the text document is very tedious process. Several techniques have been proposed for extracting correct patterns. Some of the techniques deal with term based some of the techniques deal with pattern based. Zhong, Ning, Yuefeng Li, and Sheng-Tang Wu [3] proposed novel and useful technique based on pattern. This technique is helpful to find relevant and remarkable information. The proposed technique overcomes the low-frequency and misunderstanding problems for text mining.

D.S.McNamara[4] discuss various computational methods such as Latent Semantic Analysis (LSA), Hyperspace Analog to Language(HAL), Correlated Occurrence Analog to Lexical Semantics (COALS), Contextual Similarity Log-Odds (CS-LO) and High Dimensional Explorer(HIDEx) to take out meaning from the text and advance theories of human cognition and also shows how statistical models used in various situations and its contribution in the field of cognitive science.

Scandariato, Riccardo, et al. [5] proposed text mining based approach to find which software application components has safety vulnerabilities. The components were grouped based on the terms and their connected frequencies in the source code using text mining techniques. Based on the features then detect whether the component has vulnerabilities. This process is used to spot out the components that need individual inspection.

Bao, Shenghua, et al. [6] analyze the communal emotions mining problem. The proposed emotion-topic representation used to determine the connections between user-generated social emotions and online documents. This work permits connecting the terms and sensations based on the topic. The proposed work performs well in mining the meaningful topics and advances the social emotion calculation performance.

Harpaz, Rave, et al. [7] presents current advances in pharmacovigilance and also explain about several data sources related to text mining for pharmacovigilance. This work proves that the text mining is useful to analyze several text data sources related to pharmacovigilance. It also shows that the pharmacovigilance is a developing regulation and text mining used to play an important position in its alteration.

Nassirtoussi, Arman Khadjeh, et al. [8] illustrates and conduct a review about how text mining has been used for market prediction analysis. The review based on three aspects such as preprocessing, machine learning and assessment mechanism. This proposal is used for financial institutions and bank for making better decisions to improve the profit.

Suarez-Tangil, Guillermo, et al. [9] proposed a new system Dendroid based on information retrieval techniques and text mining. This system first analyzes the malware families of android operating system. Then identify the resemblance between the malware families using vector space model. Based on the likeness, the malware families are grouped automatically. The results of this study show that the performance of this approach is accurate and also good for large malware databases.

The above discussions show that the text mining has been analyzed in various ways with different types of datasets and used in a range of fields to take fruitful decision.

2. PRE PROCESSING

Pre processing means data can be transformed into a format that can be without difficulty and successfully processed by the user. If the data is used without any pre processing, training phase is more complicated, further steps in knowledge discovery also take more time and also it can lead to erroneous results. So it is the most important one in the research. It consists of data cleaning, feature transformation and feature selection. Data cleaning in text mining means pruning and stemming can be applied to distill the text document into structured format. Feature transformation means formatted text can be converted into vectors. It refers to the process of creating new features from already existing features. These features may not be same as original features but it is more valuable than the original features. Text transformation in text mining analysis is the process of converting text documents into vector space model. Feature selection means select the related features which is used for effectively construct the model. Feature selection is used for decrease the size of the data sets. Data sets with tens or hundreds or thousands of features are reducing during the feature selection methods. The following discussion shows some of the work related to preprocessing used by various researchers.

Uysal et al. [10] illustrate how much preprocessing is important for text classification. This analysis is takes place with different aspects such as text domain, dimension reduction, text language and classification accuracy. E-mail and news of two different languages such as English and Turkish are considered for analysis. Experimental results shows that exact mixture of preprocessing task provide vital upgrading of classification accuracy.

Bullinaria et al. [11] proposed best computational method to take out more meaningful information from the large document. This method uses pruning, stemming and Single Value Decomposition (SVD) method for dimensionality reduction. The proposed one shows good results for the large text corpora.

Jivani, Anjali Ganesh. [12] discuss various stemming algorithms with their advantages and disadvantages. They discussed truncating stemmer methods, statistical stemmer methods, inflectional and derivational stemmer methods. They concludes that the lots of correspondence between the stemming algorithms and if one algorithm produce good results for one region and the other algorithm produce good results in another region.

Willett, Peter.[13] presented an algorithm to produce stem words for the English language words. It reviews the important of the algorithm and tells about how it is useful in the information retrieval research. This algorithm has been accepted and extended for large range of languages.

Moh'd A Mesleh, Abdelwadood [14] proposed a procedure for Arabic text classification based on SVM. This procedure use chi square method for feature selection. Chi square method gives better results compared with other method for text classification.

Zareapoor et al. [15] compare dimension lessening techniques for text classification. This paper takes email text for classification. Feature selection methods such as Information Gain Ratio (IGR) and chi square and feature extraction techniques such as Latent Semantic Analysis (LSA) and Principle Component Analysis(PCA) are considered for analysis. The result shows that the feature extraction methods give constant classification results.

Zhang et al. [16] proposed a method for multi label classification using naïve bayse classifier. This method uses PCA for feature extraction then use genetic algorithm is for feature selection. This method produce better results when compared with other multi label text classification methods.

Aghdam et al. [17] proposed a novel feature selection algorithm based on Ant Colony Optimization (ACO) algorithm. This paper also compares the ACO method with other feature selection methods for text categorization. The comparison results show that the ACO based method select valuable features than the other methods.

Jing et al. [18] explain about the Term Frequency – Inverse Document Frequency (TF-IDF). Apply TF-

IDF as feature selection method in various text documents classification and analyze the results. Based on the results, they proposed a new modified TF-IDF feature selection method to increase the accuracy.

Dash et al. [19] proposed a filter method for feature selection mainly used for clustering. Traditional feature selection methods used for clustering provide results based on the parameters of the clusters. The filter method produces good results irrespective of the parameters of clustering methods.

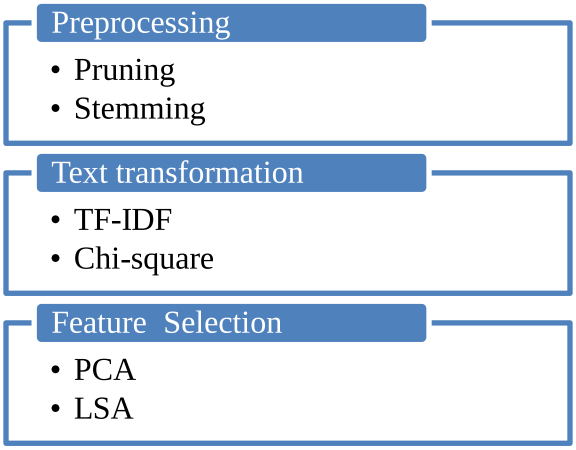

Fig. 2. Preprocessing methods discussed in various papers

Preprocessing methods discussed in various paper presented in this section is represented in this Figure 2. This survey tells more about what are the preprocessing methods available, which method is mostly used in text transformations, how many methods are available for selecting important features from the text document and also the performance of the methods in text mining. From the above survey, pruning and stemming are used to reduce the dimensionality of the text documents. TF-IDF and Chi-square methods are used to select the important words from the document. These important words are transformed into vector using PCA and LSA.

3. TEXT CLUSTERING

Text clustering is used to group the similar set of text documents. It uses natural language processing and machine learning techniques to process and classify the unstructured text document. Google search engine is one of the most excellent examples for text clustering. Many researchers proposed many algorithms and many new novel methods for clustering. Some of them are discussed here.

Pang et al. [20] proposed classifier based on cluster centroid. This classifier is a combination of K-Nearest-Neighbor (KNN) classifier and Rocchio classifier. The proposed method gives better classification results for both Chinese and English corpora and also the comparison proposed with SVM, KNN and Rocchio shows that the proposed one performs well in text classification.

Agnihotri et al. [21] deliberate to mine the important information from text. They use Guttenbergs William Shekespeare stories data set for investigation. Frequent items are collected from the text document using frequent pattern mining and using the threshold value measured the association between two words. Then find the distance between two words using cosine similarity method. This paper used hierarchical agglomerative clustering algorithm k-means clustering algorithm.

Berkhin, Pavel [22] provides survey mainly focused on clustering techniques available in data mining. This paper explained detail about partitioning relocation clustering algorithms, hierarchical clustering, grid based clustering algorithms, density based clustering algorithms and constraint based clustering algorithms and also discuss about algorithmic issues.

Brugger et al. [23] proposed an extended version of clustering clusot algorithm. This algorithm works in two steps. In the first step it calculate the clusot surface using the information obtained from the trained SOM then find the clusters using this surface automatically. This method is useful for any n-dimensional grid topology.

Kohonen, Teuvo [24] discuss in detail about how SOM is a significant clustering algorithm using various textual data sets and bioinformatics data sets. In this paper, author introduced a novel verdict. It cluster the input item based on combination of few best- matching models.

Kaufman et al. [25] proposed a work that identified the patterns of syntax and structure of the songs of Humpback whales using Self-Organizing map and Hyperspace Analog to Language. In this paper, first songs were categorized by using Self-Organizing Map (SOM) and then identify the international co-occurrence similarity in the human language using Hyperspace Analog to Language (HAL) model. HAL used to identify the particular patterns exclusive of the songs related with the specific geographical areas. These patterns are very useful to show that the humpback whales songs are specific to the particular region.

Y.Liu et al. [26] proposed fast SOM clustering technology for text information representation. To sustain high quality and efficiency in text clustering, clustering method done in two phases that is on line and offline. From the text document, feature extraction and semantic quantization are done in offline. In online, the documents are applied onto the output layers of SOM. This proposed one gives high clustering efficiency and clustering quality when compared with other conventional methods.

Jing, Liping, et al [27] proposed an algorithm used to solve text clustering sparsity problem. This method mechanically computes the feature weights during the clustering process. The comparison result shows that the modified method produce good results than the k-means and bisection k-means method.

Li, Yanjun et al. [28] proposed a new text clustering algorithm with chi square statistic method. This algorithm finds the related features recurrently using the chi square method then cluster the text document. This algorithm is tested with various data sets. It produces improved clustering accuracy in terms of the purity and F-measure.

Several clustering algorithms such as k-means and SOM are discussed in various papers for text documents. But many of the papers use SOM or the combination of SOM for text document clustering. This discussion helps to be familiar with the performance of the various clustering techniques and also assists to know about which technique is suitable for text clustering.

4. TEXT CLASSIFICATION

Text classification is used to assign class label to a text document based on their content. Types of classifications are binary classification, multi label classification, multiple classification and complex taxonomy classification. Binary classification classifies and assigns only two labels. For example positive or negative has been assigned as label for sentiment analysis. Multi label classification may assign all labels to a single document. For example let us consider news in a news paper as a set of several documents. In this, one document might be of type education, finance, religion or sports at the same time or none of these. Multiple classifications select and assign one label from many choices. For example classification of group of fruits which may be mango, pine apple, pomegranate or orange. Text classification is an important task for classifying the lots of text documents available in the web. Many researchers proposed many text classification algorithms. Some of them are discussed here.

Ko, Youngjoong [29] proposed a new method of weighting schemes in the text classification algorithm. This method introduces new modified inverse document frequency for text weighting. Then classify the document using KNN and SVM. Experimental results shows that the proposed method performs well than the existing other weighting methods.

Aggarwal et al. [30] illustrate about various feature selection, feature transformation and classification methods for text document classification. They discussed mutual information, chi-square statistics and gini index methods for feature selection. Then Latent Semantic Index(LSI) , Singular Value Decomposition(SVD) and Linear Discriminant Analysis(LDA) methods for feature transformation. They discussed decision tree classifiers, linear classifiers, rule based classifiers and proximity based classifiers with several applications such as news group filtering, target marketing and medical diagnosis.

Moraes et al. [31] discussed document classification. This paper discussed sentiment classification in two steps. First step is feature selection and extraction process and then classify the document using Artificial Neural Network and Support Vector Machine (SVM).

Alsaleem, Saleh [32] proposed a method to classify the Arabic text document. For this Arabic text classification, they use Naïve Bayes(NB) classification method and SVM method. From the experimental results, SVM performance is better than NB.

Pawar et al. [33] present various methods for text classification. This paper compares various classifiers such as NB, KNN and SVM with the standard datasets and also provides the results of various classifiers. This result shows that SVM is the best method for text classification.

Desmet et al. [34] discussed text classification to analyze the information related to suicidal feelings on the social media. This paper collect information from post then used genetic algorithm to select informative features from that. Selected features transformed using LSA model then it is classified using SVM. Discussion show that the text classification methods useful for finding suicide related problems and provide some suggestion to the deprived people.

Jindal et al. [35] present a literature survey with various feature selection methods and techniques used in the field of text mining. This paper illustrates and deliver clear idea about what are related journals with text mining, which methods are used for feature selection and feature extraction, which type of dataset is used for experiment and which techniques are used for classify the text document. This paper concludes that the mainly used dataset by researchers for text classification is Reuter-21578, web KB and 20-news group. In the field of feature selection, researchers mainly used chi-square statistics and Information Gain(IG). And in the machine learning algorithms, SVM and KNN are mostly used for text classification.

Briscoe et al. [36] discussed in detail about how to use the text mining techniques for assessment of answers provided by learners for questions. This paper discussed various methods for feature transformation but finally confirmed that the LSA technique is better one and discussed various techniques for ranking purpose.

Zhong et al. [37] proposed a novel technique for web page classification automatically. This paper discussed NB, TFIDF and SVM ensemble classifier for classify the web pages. The experimental results state that the ensemble classifier performance is better than others.

Bin et al. [38] proposed a multi-classifier to automatically asses the essay. This paper use Vector Space Model (VSM) to represent the text as a vector. Then use Document Frequency (DF), IG and Chi-square statistics to select features from the essay. To classify the essay, combination of NB, KNN and SVM techniques are used. This multi classifier technique gives better results when evaluating essay.

The survey about text classification discussed what classification methods available for text document classification, which method is suitable for text classification and the performance of the classification algorithm with different types of text datasets. The papers discussed here used various classification algorithms such as KNN, Naïve Bayes and SVM for text document classification. Hence, this survey is most helpful to select the most suitable algorithm for text document classification.

5. CONCLUSION

This survey shows and discussed about the different types of methodological aspects of text mining, significance of pre processing methods, various text clustering approaches and classification algorithms for text mining. These discussions provide a platform for understanding the text clustering and classification problem and also give an idea for developing new text clustering and classification algorithm or encourage the researchers to modify the existing algorithm to attain the high accuracy and efficiency. This survey may determine or provides the direction of research work related to text document.

REFERENCES

[1] T.Yukiko and A.Yumi, “User needs search using text mining”, Lecture Notes in Computer Science, 2013, Volume 8018, pp. 607-615.

[2] Li, Nan, and Desheng Dash Wu. "Using text mining and sentiment analysis for online forums hotspot detection and forecast." Decision support systems 48.2 (2010): 354-368.

[3] Zhong, Ning, Yuefeng Li, and Sheng-Tang Wu. "Effective pattern discovery for text mining." IEEE transactions on knowledge and data engineering 24.1 (2012): 30-44.

[4] D.S.McNamara, “Computational methods to extract meaning from text and advance theories of human cognition”, Topics in Cognitive Science, 2011, Volume 3, Issue 1, pp. 3-17.

[5] Scandariato, Riccardo, et al. "Predicting vulnerable software components via text mining." IEEE Transactions on Software Engineering 40.10 (2014): 993-1006.

[6] Bao, Shenghua, et al. "Mining social emotions from affective text." IEEE transactions on knowledge and data engineering 24.9 (2012): 1658-1670.

[7] Harpaz, Rave, et al. "Text mining for adverse drug events: the promise, challenges, and state of the art." Drug safety 37.10 (2014): 777-790.

[8] Nassirtoussi, Arman Khadjeh, et al. "Text mining for market prediction: A systematic review." Expert Systems with Applications 41.16 (2014): 7653-7670.

[9] Suarez-Tangil, Guillermo, et al. "Dendroid: A text mining approach to analyzing and classifying code structures in android malware families." Expert Systems with Applications 41.4 (2014): 1104-1117.

[10] Uysal, Alper Kursat, and Serkan Gunal. "The impact of preprocessing on text classification." Information Processing & Management 50.1 (2014): 104-112.

[11] Bullinaria, John A., and Joseph P. Levy. "Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD." Behavior research methods 44.3 (2012): 890-907.

[12] Jivani, Anjali Ganesh. "A comparative study of stemming algorithms." Int. J. Comp. Tech. Appl 2.6 (2011): 1930-1938.

[13] Willett, Peter. "The Porter stemming algorithm: then and now." Program 40.3 (2006): 219-223.

[14] Moh'd A Mesleh, Abdelwadood. "Chi square feature extraction based SVMs Arabic language text categorization system." Journal of Computer Science 3.6 (2007): 430-435.

[15] Zareapoor, Masoumeh, and K. R. Seeja. "Feature extraction or feature selection for text classification: A case study on phishing email detection." International Journal of Information Engineering and Electronic Business 7.2 (2015): 60.

[16] Zhang, Min-Ling, José M. Peña, and Victor Robles. "Feature selection for multi-label naive

Bayes classification." Information Sciences 179.19 (2009): 3218-3229.

[17] Aghdam, Mehdi Hosseinzadeh, Nasser Ghasem-Aghaee, and Mohammad Ehsan Basiri. "Text feature selection using ant colony optimization." Expert systems with applications 36.3 (2009): 6843-6853.

[18] Jing, Li-Ping, Hou-Kuan Huang, and Hong-Bo Shi. "Improved feature selection approach TFIDF in text mining." Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on. Vol. 2. IEEE, 2002.

[19] Dash, Manoranjan, and Poon Wei Koot. "Feature selection for clustering." Encyclopedia of database systems. Springer, Boston, MA, 2009. 1119-1125.

[20] Pang, Guansong, and Shengyi Jiang. "A generalized cluster centroid based classifier for text categorization." Information Processing & Management 49.2 (2013): 576-586.

[21] Agnihotri, Deepak, Kesari Verma, and Priyanka Tripathi. "Pattern and cluster mining on text data." Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on. IEEE, 2014.

[22] Berkhin, Pavel. "A survey of clustering data mining techniques." Grouping multidimensional data. Springer, Berlin, Heidelberg, 2006. 25-71.

[23] Brugger, Dominik, Martin Bogdan, and Wolfgang Rosenstiel. "Automatic cluster detection in Kohonen's SOM." IEEE Transactions on Neural Networks 19.3 (2008): 442-459.

[24] Kohonen, Teuvo. "Essentials of the self-organizing map." Neural networks 37 (2013): 52-65.

[25] Kaufman, A. B., Green, S. R., Seitz, A. R., & Burgess, C. (2012). Using a self-organizing map (SOM) and the hyperspace analog to language (HAL) model to identify patterns of syntax and structure in the songs of humpback whales. International Journal of Comparative Psychology, 25(3).

[26] Liu, Yuan-Chao, Chong Wu, and Ming Liu. "Research of fast SOM clustering for text information." Expert Systems with Applications 38.8 (2011): 9325-9333.

[27] Jing, Liping, et al. "Subspace clustering of text documents with feature weighting k-means algorithm." Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Berlin, Heidelberg, 2005.

[28] Li, Yanjun, Congnan Luo, and Soon M. Chung. "Text clustering with feature selection by using statistical data." IEEE Transactions on knowledge and Data Engineering 20.5 (2008): 641-652.

[29] Ko, Youngjoong. "A study of term weighting schemes using class information for text classification." Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012.

[30] Aggarwal, Charu C., and ChengXiang Zhai. "A survey of text classification algorithms." Mining text data. Springer, Boston, MA, 2012. 163-222.

[31] Moraes, Rodrigo, JoãO Francisco Valiati, and Wilson P. GaviãO Neto. "Document-level sentiment classification: An empirical comparison between SVM and ANN." Expert Systems with Applications 40.2 (2013): 621-633.

[32] Alsaleem, Saleh. "Automated Arabic Text Categorization Using SVM and NB." Int. Arab J. e-Technol. 2.2 (2011): 124-128.

[33] Pawar, Pratiksha Y., and S. H. Gawande. "A comparative study on different types of approaches to text categorization." International Journal of Machine Learning and Computing 2.4 (2012): 423.

[34] Desmet, Bart, and Véronique Hoste. "Online suicide prevention through optimised text classification." Information Sciences 439 (2018): 61-78.

[35] Jindal, Rajni, Ruchika Malhotra, and Abha Jain. "Techniques for text classification: Literature review and current trends." webology 12.2 (2015).

[36] Briscoe, Ted, Ben Medlock, and Øistein Andersen. Automated assessment of ESOL free text examinations. No. UCAM-CL-TR-790. University of Cambridge, Computer Laboratory, 2010.

[37] Zhong, Shaobo, and Dongsheng Zou. "Web Page Classification using an ensemble of support vector machine classifiers." journal of Networks 6.11 (2011): 1625.

[38] Bin, Li, and Yao Jian-Min. "Automated essay scoring using multi-classifier fusion." International Conference on Information and Management Engineering. Springer, Berlin, Heidelberg, 2011.