Intra News Category Classification using N-gram TF-IDF Features and Decision Tree Classifier

Dimpleveer Singh[1], Sumit Malhotra[2]
[1, 2] Department of Computer Science Engineering
[1, 2] Bhai Gurdas Institute of Engineering & Technology, Sangrur, Punjab

***Abstract**- During the last decade, the majority of important newspapers and magazines developed websites to present news and other materials. Reading important and interesting news is useful to users, but it is also time consuming since they would have to read all the news items. Therefore, a news classification method for receiving the relevant information quickly seems to be essential. Many researchers have allocated much work to the automation of news classification in order to develop such a text classification system. In a classification process, text analysis is an important preprocessing steps that leads to good features and then to finally building a good classifier. It is obvious that different text types need certain fine tuning to achieve better classification results. In this work, inter and intra news classification has been explored and multi-feature based news classification system has been proposed based on BBC news dataset. Sports category has been selected for intra class classification whereas technology, business, sports, politics and entertainment are used for inter class classification. At first, pre-processing has been carried out in which special characters, numbers etc. has been eradicated from each news sample from different news categories. Further feature extraction has been carried out in which TF-IDF of unigram, bigram and trigram word tokens has been used as concatenated feature vector. For classification different classifiers has been explored but decision tree gives more effective results than the others. Experiment results shows that proposed system gives about 96% accuracy in intra news classification. For inter class classification, about 98% accuracy has been achieved in true identification of tested news dataset.*

Keywords- News classification, Term Frequency-Inverse Document Frequency, N-gram, Decision Tree Classifier, etc.

I. INTRODUCTION

Text classification methods have drawn much attention in recent years and have been widely used in many programs. These techniques are essential because the textual data is swiftly rising with the passage of time. Text mining tools are required to perform indexing and retrieval of this rapidly growing text data. Text mining is the finding of some information which is previously unknown by extracting that information from large sets of unstructured text. Today, this un-structured data is growing as mostly the information is available in an electronic form such as emails, on World Wide Web, electronic publications and other documents. The term un-structured mean, the type of data in which the text is occurring in a natural free form or a sequence that may include word and sentence ambiguity.

This un-structured information cannot be used for further processing by computers. The computers typically handle text as simple sequences of character string and are unable to provide useful information from the given text, without any process performed on it. Therefore, specific processing and preprocessing methods are required in order to extract useful patterns and information from the unstructured text [1]. During the last decade, the majority of important newspapers and magazines developed websites to present news and other materials. Reading important and interesting news is useful to users, but it is also time consuming since they would have to read all the news items.

Therefore, a news classification method for receiving the relevant information quickly seems to be essential. Many researchers have allocated much work to the automation of news classification in order to develop such a text classification system. Titles of several news classification methods proposed in the literature include : Financial News Classification [2], Classification of Short Texts [3], Automatic News Headlines Classification [4], a Hybrid Text Classification Approach with Low Dependency on Parameter by Integrating K-Nearest Neighbor and Support Vector Machine [5], Classification of News Headlines for Providing User-Centered E-Newspaper [6], Emotions Extraction from News Headlines [7] ,and Short News Headlines Classification of Twitter [8].

Text Classification Process

The stages of text classification are discussing as following points:

- **Document Collection:**

In this step collect the different types of document like html, .pdf, .doc etc.

- **Pre-Processing:**

In this present the text document into clear word format. For example Perform Tokenization, stemming word, removing stop words, tokenization a document is treated as a string and then partition into list of tokens. In removing stop word remove the stop words for example “the”, “a”, “and”. In stemming words converts different words from into similar canonical form.

- **Indexing :**

In this provide the index to every document with this easily identify each document.

- **Feature selection:**

After preprocessing and indexing the important step of text classification is feature selection. The main idea of feature selection is to select subset of features from the original document. It is performed by keeping the words with highest score according to predetermined measure of the importance of the word. There are various types of features which can be used for text data i.e. term frequency, TFIDF, unigram, bigram or trigram features of the words

- **Classification:**

In this step, documents are classified into predefined categories. The documents can be classified by supervised, unsupervised methods.When the class label of each document is known that is called supervised classification when the class label of documents is not known that is called unsupervised classification. There are various types of classifiers available i.e. SVM, ANN, KNN, decision trees etc. Decision tree gives effective results and is used in the present work.

- **Performance Evaluation:**

This is the last stage of text classification this is experimentally, rather than analytically. In this measure the performance. Many measures have been used like precision and recall.

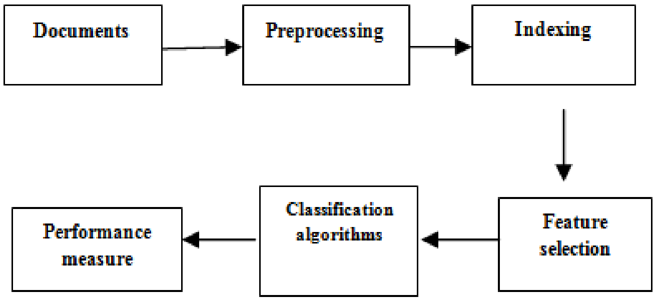


Figure 1. Document Classification Process

II. LITERATURE SURVEY

Ghosh et al. [9] compares these three probabilistic models for text clustering, both theoretically and empirically, using a general model-based clustering framework. For each model, they investigate three strategies for assigning documents to models: maximum likelihood (k-means) assignment, stochastic assignment, and soft assignment.

Vin-tan et al. [10] says that building up on the meta-classifier presented, based on 8 SVM components, they add to these a new Bayes type classifier which leads to a significant improvement of the upper limit that the meta- classifier can reach. Thus, the meta-classifier upper limit has increased from 94.21% when using 8 SVM classifiers to 98.63% when using the 8 SVM classifiers plus the Bayes classifier. Moreover, in the case of the 9-classifiers SBED meta-classifier they obtained even lower results, on average dropping from 92.04% to 90.38%. In the case of the 9-classifiers SBCOS, the classification accuracy of the meta-classifier has increased from 89.74% to 93.10%.

Srinivasagan et al. [11] developed the intelligent News Classifier and experimented with online news from web for the category Sports, Finance and Politics. The novel approach combining two powerful algorithms, Hidden Markov Model and Support Vector Machine, in the online news classification domain provides extremely good result compared to existing methodologies. By the introduction of several preprocessing techniques and the application of filters they reduced the noise to a great extent, which in turn improved the classification accuracy. Preprocessing in the training data set significantly reduced the training computational time.

Thakare et al. [12] represent overview of data mining system and some of its application. Information play important role in every sphere of human life. It is very important to gather data from different data sources store and maintain the data. Generate the information, generate knowledge and disseminate data, information and knowledge to every stakeholder. Due to vast use of computers and electronics devices and tremendous growth in computing power and storage capacity, there is explosive growth in data collection. The storing of the data in data warehouse enables entire enterprise to access a reliable current database.

Zaveri et al. [13] developed text classification system using machine learning tools. Text classification can be automated successfully using machine learning techniques, however pre-processing and feature selection steps play a crucial role in the size and quality of training input given to the classifier, which in turn affects the classifier accuracy. Sophisticated text classifiers are not yet available for several regional languages, which if developed would be useful for several governmental and commercial projects.

Prasad et al. [14] introduced the main components of such forecasting systems. The main objective is to predict the Classify the Financial News based on the contents of relevant news articles which can be accomplished by building a prediction model which is able to classify the news as either rise or drop. Making this prediction model is a binary classification problem which uses two types of data: past intraday price and past news articles. The prediction model applying all the types of news related to auto industry in general and the ones related to competitors and compare the results with the current prediction model.

Mahender et al. [15] focused review on the existing literature and explored the documents representation and an analysis of feature selection methods and classification algorithms were presented. It was verified from the study that information Gain and Chi square statistics are the most commonly used and well performed methods for feature selection, however many other feature selection methods are proposed. This gives a brief introduction to the various text representation schemes. The existing classification methods are compared and contrasted based on various parameters namely criteria used for classification, algorithms adopted and classification time complexities.

Liu et al. [16] reports a controlled study with statistical significance tests on five text categorization methods: the Support Vector Machines (SVM), a k-Nearest Neighbor (kNN) classifier, a neural network (NNet) approach, the Linear Least- squares Fit (LLSF) mapping and a Naive Bayes (NB) classifier. They focus on the robustness of these methods in dealing with a skewed category distribution, and their performance as function of the training-set category frequency.

Araghi et al. [17] aims to classify news into variety of groups so that users can identify the most popular news group in the gievn country at a time. Based on Term Frequency-Inverse Document Frequency (TF-IDF) and Support Vector Machine (SVM), a news classification method was proposed. The proposed approach is comprised of three different steps: 1) text preprocessing, 2) feature extraction based on TF-IDF, and 3) classification based on SVM.

Wei-Ta Chu et al. [18] presented an advertisement detection, segmentation, and classification framework to facilitate advertisement studies in newspaper images and website snapshots. We classify advertisement based on visual analysis that attracted little attention before.

Dyah Rahmawati et al. [19] included unlabeled data in solving multilabel classification problem for Indonesian news article through Word2vec to construct vector representation of the words. This representation can catch the semantic similarity between words and we used these vectors to extract the classification features.

Ghulam Mujtaba et al. [20] presents a holistic analysis of the entire email classification domain by assembling almost all major research efforts in this regard to assist researchers in this field to gain a better understanding of the existing solutions in the major areas of email classification. Articles on email classification published in 2006–2016 were comprehensively reviewed. The selected articles were examined from five rationale aspects: email classification application areas, datasets used in each application area, features sets used in each application area, classification techniques, and performance metrics.

Bekir Parlak et al. [21] show that the most successful setting is the combination of Bayesian Network classifier, distinguishing feature selector, and TF term weighting method. As a future work, a new dataset containing Turkish versions of the documents in the self-constructed dataset may be compiled and classification performances of these two datasets having same documents in different languages can be extensively analyzed.

III. PROPOSED WORK

This approach is comprised of three steps: text preprocessing, feature selection based on TF-IDF, and Inter news classification using Decision tree classifier. Each step is individually described in the following section.

- **Text Preprocessing**

Text preprocessing is the first step in the process of news classification.A text is effectively preprocessed when unstructured data, mostly combinations of useful and useless data, are first received. The data are collected from different sources, and they should be cleaned. First, the text data are cleaned from distortion and useless information such as punctuations, exclamations, semicolons, irrelevant sentences, quotations, dates, etc.

- **Transforming Cases**

The transform case operator transforms all the (upper case and lower case) characters existing in the text into lower case characters. This operator is used to eliminate homologous words which are different only in terms of their case.

- **Tokenizing**

The majority of studies conducted on text mining include words or sentences which should be separated word by word for increased processing. Thus, all the words are separated in sentences, and all the punctuations are disposed of since they cannot represent any group. This simplifies computations in the next steps [22].

- **Filtering Stopwords**

Filtering stopwords are one of the methods used since the first studies were conducted on information retrieval. This algorithm is mainly employed to delete unnecessary things such as words occurring too frequently or too infrequently in sentences or text documents. It is also used to delete unimportant words or the words with no specific meanings such as a, an, or the. Like the previous steps, this step is applied in order to reduce processing time or computational complexity [22].

- **Feature Extraction based on TF-IDF of unigram, bigram and trigram**

Upon creating the glossary, the weight of each word is calculated with respect to TF-IDF which is one of the most famous algorithms used in text mining research. The word frequency means the number of times a term is repeated in a text, and IDF stands for Inverse Document Frequency, an algorithm used to calculate the inverse probability of finding a word in a text [22]. Equation 1 is a classic TF-IDF equation used to calculate weight:

$$W_{ij} = tf_{ij} * \log \frac{N}{df_i} \qquad (1)$$

In this equation, wij is the weight of weight of the word i in the document j, N is the number of documents in the set of total documents, tfij is the frequency of the word i in the document j, and dfi is the number of documents containing the word i [23].

- **Decision Trees for Classification**

A decision tree [24] is represented in the form of a tree and it is a classifier used for text categorization where each node can act as a leaf or decision node. In medical science [25], decisions are to be taken rapidly and minor delay may lead to serious issues. Performing conceptual analysis [24] and decision making proves appropriate for such situations and system may work better than before. For manipulating such situations, decision trees have helped researchers a lot where medical experts can make better examinations on basis of results they obtain and rules can be deducted on their basis. Decision trees are easy to understand and rules can be easily generated through them. They can solve complex problems very easily. But training through decision trees is very expensive and is very costly. Moreover, a news headline class can only be connected to one branch only. A single mistake in higher upper level can cause whole sub tree invalid. There are issues of continuous variables as well as over fitting in decision trees [26].The main purpose of the decision tree algorithms is to split the feature space into unique regions corresponding to the classes [27]. An unknown feature vector is assigned to a class via a sequence of Yes/No decisions along a path of nodes of a decision tree. C4.5 is an algorithm used to generate a decision tree and it is known as one of the successful decision tree classification algorithms. System module for the proposed work has been given in figure below.

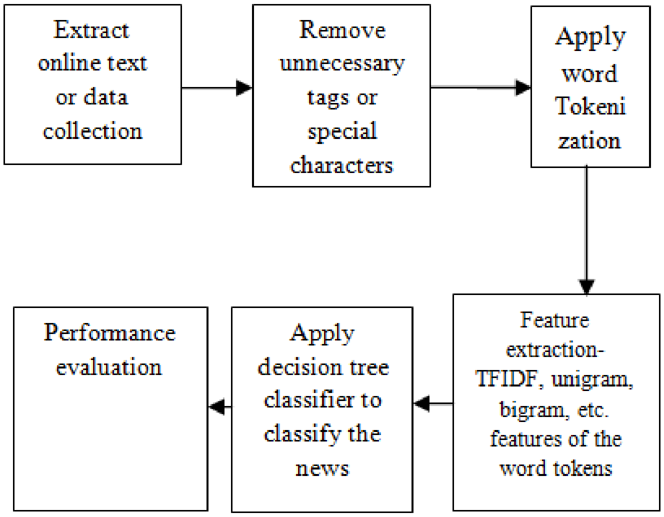

Extract online text or data collection
Remove unnecessary tags or special characters
Apply word Tokeni zation
Feature extraction- TFIDF, unigram, bigram, etc. features of the word tokens
Apply decision tree classifier to classify the news
Performance evaluation

Figure 2. News Classification Process.

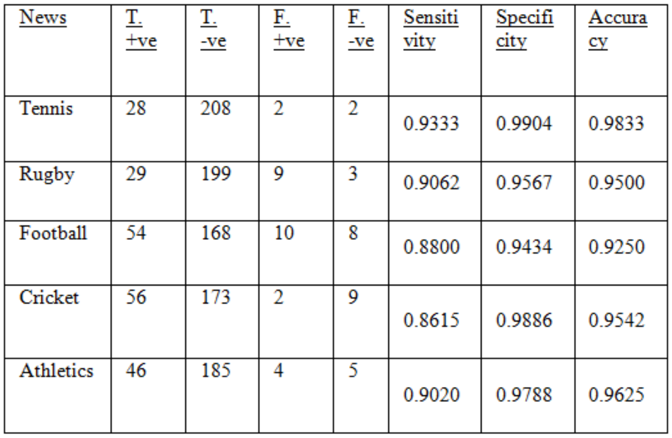
IV. RESULTS

Dataset used for this research work is BBC dataset which has various categories i.e. Technology, entertainment, business, politics and sports . In sports category, further five different sports have been provided named as tennis, rugby, football, cricket and athletics. Intra news classification has been carried out in which sports category has been considered for performance evaluation of the proposed news classifier system.

Table 1. Samples of each sport category in collected dataset

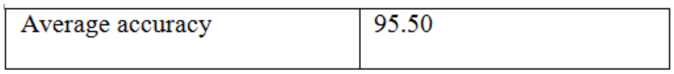
Category of Sports	Total News Taken	News used for Training	News used for Testing
Tennis	100	70	30
Rugby	147	115	32
Football	265	203	62
Cricket	124	58	66
Athletics	101	51	50

Table 2. Performance evaluation using different parameters



News	T. +ve	T. -ve	F. +ve	F. -ve	Sensitivity	Specificity	Accuracy
Tennis	28	208	2	2	0.9333	0.9904	0.9833
Rugby	29	199	9	3	0.9062	0.9567	0.9500
Football	54	168	10	8	0.8800	0.9434	0.9250
Cricket	56	173	2	9	0.8615	0.9886	0.9542
Athletics	46	185	4	5	0.9020	0.9788	0.9625

Table 3. Average accuracy



Average accuracy	95.50

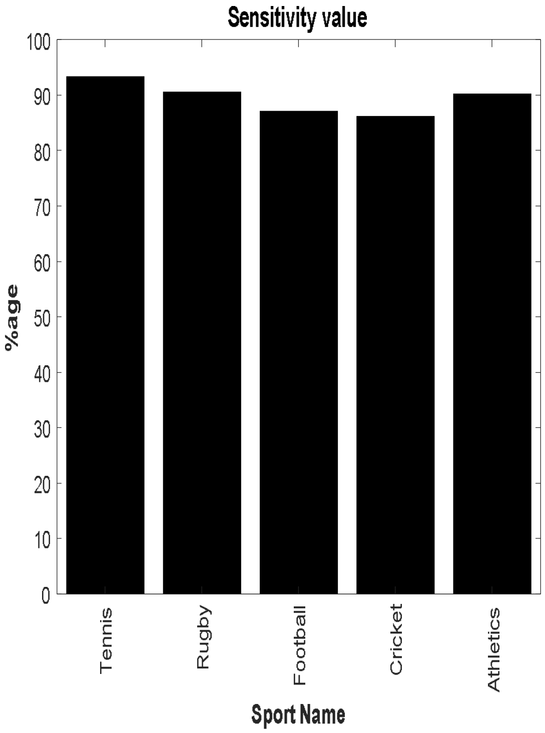


Figure 3. Bar Graph of Sensitivity Value

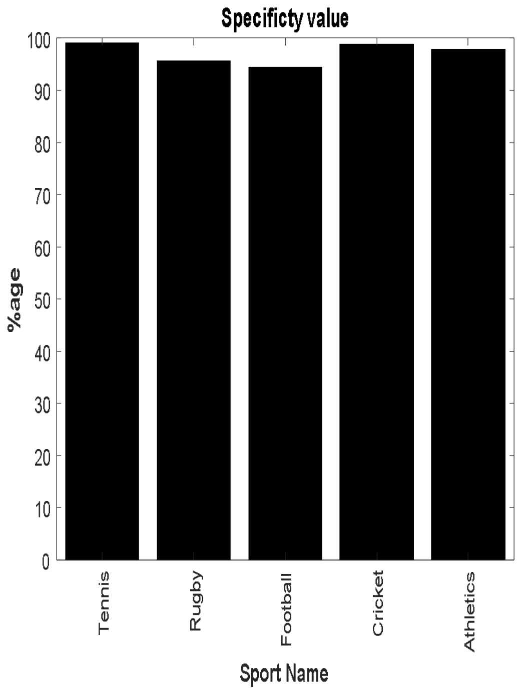


Figure 4. Bar Graph of Specificity Value

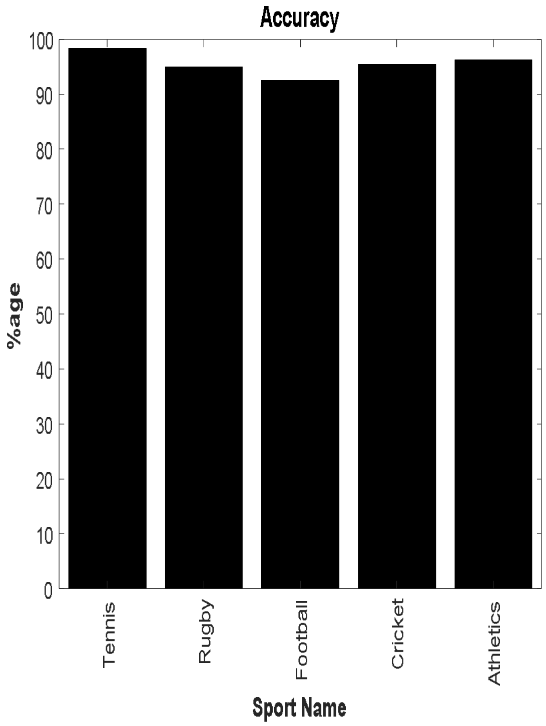


Figure 5. Bar Graph of Accuracy Value

There are five categories in sports category in which Tennis has 100 news samples out of which seventy has been used for training and thirty has been used for testing. Similarly for other categories about 70% percent news documents have been used for training and thirty percent are used for testing. Performance evaluation has been carried out only for tested data-sets as they are not included in training the classifier. Sensitivity, specificity and accuracy has been given in tabular as well as graphical form which shows about 96% accuracy in true classification of news documents

V. CONCLUSION

Category classification, for news, is a multi-label text classification problem. The goal is to assign one or more categories to a news article. A standard technique in multi-label text classification is to use a set of binary classifiers. For each category, a classifier is used to give a “yes” or “no” answer on if the category should be assigned to a text. The previous methods, typically, require both positive and negative examples for training data. The initial set of training data requires that each document is assigned all positive labels. Support Vector Machines offer state-of-the-art performance, however they are slow to train and updating the training data is not really a viable option. Naive Bayesian Classifiers can give good performance as well, but depending on the features used they can require for previous training data to be kept. These problems have been solved using decision tree classifier which gives higher accuracy in accurate classification of news documents. Results has also been improved from existed work as only single word TF-IDF features were considered in most of existed algorithms of news classification. In this work, we have introduced two word and three-word cluster based TF-IDF features which are combined to single vector before providing it to the classifier. In future, work can be furtherextended for variety of purposes. News headlines text can beclassified for languages other than English also e.g. Urdu, Gurmuki, Hindi etc. and news setiment analysis can be carried out

REFERENCES

[1] M. I. Rana, S. Khalid, and M. U. Akbar, "News classification based on their headlines: A review", in IEEE 17th International Multi-Topic Conference (INMIC) , pp. 211-216, 2014.

[2] B. Drury, L. Torgo, and J. J. Almeida, "Classifying news stories to estimate the direction of a stock market index", in 6th Conference on Information Systems and Technologies Iberian (CISTI), pp. 1- 4, 2011.

[3] A. Heß, P. Dopichaj, and C. Maaß, "Multi-value classification of very short texts", in proceedings of the 31st annual German conference on Advances in Artificial Intelligence, Springer- Verlag Berlin, pp.70-77, 2008.

[4] M. W. Pope, "Automatic classification of online news headlines", University of North Carolina at Chapel Hill, November 2007.

[5] C. H. Wan, L. H. Lee, R. Rajkumar, and D. Isa, "A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine", Expert Systems with Applications, pp. 11880-11888, 2012.

[6] R. Deshmukh and M. D. Kirange, "Classifying news headlines for providing user centered e-newspaper using SVM", in International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), vol 2, Issue 3, 2013.

[7] D. Kirange, "Emotion classification of news headlines using svm", Asian Journal of Computer Science and Information Technology, pp. 104-106, 2012.

[8] I. Dilrukshi, K. De Zoysa, and A. Caldera, "Twitter news classification using SVM", in 8th International Conference on Computer Science & Education (ICCSE), pp. 287-291, 2013.

[9] Shrizhong and Joydeep Ghosh ,” A Comparative study of generative models for document clustering”, the university o texas at Austin ,TX 78712-1084,Vol 4, No1,2008.

[10] D. Morariu, R. Cretulescu and L. Vintan,"Improving a SVM Meta-classifier for Text Documents by using Naïve-Baye", Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844, vol.No.3, pp.351-361,2010.

[11] Krishnalal G, S Babu Rengarajan, K G Srinivasagan ,“A new text mining approach based on HMM -SVM for web news classification”, International Journal of Computer Applications (0975 - 8887), Volume 1 – No. 19, pp.98-104, 2010.

[12] Mr.S.PDeshpande and Dr. V.M Thakre,” Data mining system and applications:a review”,'international journal of distributed and parallel system(IJDPS), Vol.1.No.1, pp.32-44, September 2010.

[13] Mita K. Dalal, Mukesh A.Zaveri,“Automatic text classification:A technical review”, International Journal of Computer Applications (0975 – 8887) , Volume 28– No.2, pp.37-40, August 2011.

[14] Rama Bharath Kumar, Bangari Shravan Kumar, Chandragiri Shiva Sai Prasad,”financial news classification using SVM”, International Journal of Scientific and Research Publications, Volume 2, Issue 3, pp.1-6, March 2012 .

[15] VandanaKorde, C Namrata Mahender,”Text classification and classifier a survey”, International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012

[16] Yiming yang and xin liu,”A re-examination of text categorization methods”, caenegie mellon university pittsburg , PA 15213-3702 , USA ,Vol 18, No.2, March 2012.

[17] S. M. H. Dadgar, M. S. Araghi and M. M. Farahani, "A novel text mining approach based on TF-IDF and Support Vector Machine for news classification", 2016 IEEE International Conference on Engineering and Technology (ICETECH), Coimbatore, pp. 112-116, 2016.

[18] W. T. Chu and H. Y. Chang, "Advertisement Detection, Segmentation, and Classification for Newspaper Images and Website Snapshots", 2016 International Computer Symposium (ICS), Chiayi, pp. 396-401, 2016.

[19] D. Rahmawati and M. L. Khodra, "Word2vec semantic representation in multilabel classification for Indonesian news article", 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), George Town, pp. 1-6, 2016.

[20] G. Mujtaba, L. Shuib, R. G. Raj, N. Majeed and M. A. Al-Garadi, "Email Classification Research Trends: Review and Open Issues", in IEEE Access, vol. 5, pp. 9044-9064, 2017.

[21] Bekir Parlak, Alper Kursat Uysal, "On Feature Weighting and Selection for Medical Document Classification", Published in: Developments and Advances in Intelligent Systems and Applications pp 269-282, 2017.

[22] A. A. Hakim, A. Erwin, K. Eng, M. Galinium, and W. Muliady, "Automated document classification for news article in bahasa Indonesia based on term frequency inverse document frequency (TFIDF) approach", in 6th International Conference on Information Technology and Electrical Engineering (ICITEE), pp. 1-4, 2014.

[23] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of (TF-IDF), LSI and multi-words for text classification", Expert Systems with Applications, vol. 38, pp. 2758-2765,2011.

[24] A. M. Mahmood, N. Satuluri, and M. R. Kuppa, "An Overview of Recent and Traditional Decision Tree Classifiers in Machine Learning", International Journal of Research and Reviews in Ad Hoc Networks, Vol. 1, No.1, 2011.

[25] PodgorelecV, Kokol P, Stiglic B, Rozman I ”Decision trees: An overview and their use in medicine”, Journal of Medical Systems, 26:445–463, 2002.

[26] Amey K.Shet Tilve, Surabhi N.Jain "A Survey on Machine Learing Techniques for Text Classification", International Journal of Engineering Sciences & Research Technology, pp.513-520, Feb2017.

[27] A.K.Uysal, S.Gunal ”A novel probabilistic feature selection method for text classification”, Knowledge-Based Systems, Vol.36, 226–235, 2012.