International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835 Volume-4, Issue-3, Mar.-2017
http://iraj.in

IMPACT OF TOPIC MODELLING METHODS AND TEXT CLASSIFICATION TECHNIQUES IN TEXT MINING: A SURVEY

[1]MINO GEORGE, [2]P. BEAULAH SOUNDARABAI, [3]KARTHIK KRISHNAMURTHI

[1]Mphil Scholar, [2]Assoc. Professor, [3]Assist. Professor, Dept. of Computer Science, Christ University, Bangalore, IndiaChrist University, Bangalore, India Christ University, Bangalore, India
E-mail: [1]mino.george@res.christuniversity.in, [2]beaulah.s@christuniversity.in, [3]karthik.k@christuniversity.in

Abstract - The continuous growth of Information technology increases the amount of data explosively. Organize and analyse large document collection has become a big challenge. Text classifiers and topic models are used to sort out this problem. This paper mainly focuses on these two categories. First category discusses the three methods of topic modeling. They are Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA). The second category is text classification models. This includes Naïve Bayes Classifier, K-Nearest Neighbor and Support Vector Machines (SVM).A literature survey has done and explored the two categories. Finally, mentioned the combination of text classifiers and topic models can improve the classification accuracy. A combined approach of LDA and SVM show better performance than the others.

Keywords - K- Nearest Neighbor, Latent Dirichlet Allocation, Latent Semantic Analysis, Naïve Bayes, Probabilistic Latent Semantic Analysis, Support Vector Machine, Topic Modeling, Text Classification.

I. INTRODUCTION

To have a good way of organizing and managing huge volumes of corpus, we need new computational tools or techniques. And that should deal with automatic organization, searching and exploring large corpus. In recent years, the research on text classification is a hot topic among researchers. Manual text classification is not only consumes a lot of time but the financial resources also. When there are huge volumes of data, mining useful information is a critical problem. This can be solved by text classification with the help of topic modeling. Topic modeling is the process of discovering hidden patterns that reflects the underlying topics in documents. The main idea of topic models is that it considers entire corpus as a collection of documents, each document as a mixture of topics and each topic as a collection or distribution of terms. It is based on probabilistic distribution of words and topics in documents [1]. Here we have done a survey on the most popular topic modeling methods LDA, LSA and PLSA. All of this topic models have the capability to successfully improve classification accuracy in finding topics.

Text classification is another area where we can classify and label large disordered texts. There are lot of text categorization algorithms which works very effectively on large textual data. The traditional text categorization techniques are not at all useful in this era. In this paper, we have analysed categorization models which works well with large documents. We have discussed the basic ideas of SVM (Support Vector Machine), Naïve Bayes Classifier and K-Nearest Neighbor. The main objective of a classifier is to build a model from the training data and predict the values of the test data [2].It has been also discussed the combination of topic model method and text classifier. We can perform a text classification based on topic models.

This paper is organized as follows. Section II provides the Literature Review of methods of topic modeling and Text classification techniques. Section III and IV overviews the detailed study of topic modeling methods and text classifiers. Section V presents a comparative study on two categories. Last, conclusion is in section VII.

II. LITERATURE REVIEW

A. Methods of Topic Modelling

This section deals with the literature survey of three topic model methods. They are LSA, PLSA and LDA.

Latent Semantic Analysis (LSA)

Deerwester et.al, 1990 [3], proposed a new method for automatic indexing and information retrieval. They were tried to solve the problems in the existing retrieval techniques to match words of queries with the words of documents. In the past, the model was known as Latent Semantic Indexing (LSI) but changed to Latent Semantic Analysis (LSA) for information retrieval tasking. The model tries to overcome the drawbacks of term- matching retrieval by using latent semantic indexing by using a singular value decomposition of the matrix to identify the linear space. The large term by document matrix is decomposed into a set of orthogonal factors. This way of approach can achieve major compression in large corpus. They argue that some features of LSI can capture some aspects of synonymy and polysemy. In 1997, T. Dumais et.al, [4], described about a method for fully automated Cross-Language (CL)

document retrieval. In this method no query translation is required i.e., queries in one language can retrieve documents in other language. This is accomplished by constructing a multilingual semantic space using Latent Semantic Indexing (LSI). They state that by using all available measures CL-LSI shows better performance.

Probabilistic Latent Semantic Analysis (PLSA)

A major step forwarded in this area was made by Hofmann, 1999 [5], who proposed a statistical method called probabilistic LSI (PLSI), which is closely related to LSA. The model uses a generative latent class model to perform probabilistic mixture decomposition. PLSA model has many applications. They presented perplexity results for different types of texts and they mentioned its applications in document indexing. The experimental results show that the PLSA outperforms the standard LSA method. Paik et.al, 2010 [6] proposed a method, which extracts a word, which is most semantically relative to two words by the document classification. In this paper PLSA using web search engine as a corpus. The main aim of this paper was to get relations of words on the web and to evaluate and analyse current web feature.

Latent Dirichlet Allocation (LDA)

Blei et.al, 2002 [7], proposed LDA model. They described this model as a generative probabilistic three-level hierarchical Bayesian model. They say that this model is for large collections of discrete data and tries to find short descriptions for the collection to process large collection of documents. It models the corpus as a collection of documents, each document as a distribution of multiple topics and each topic as a mixture of words. It overcomes all the drawbacks of LSA and PLSA model. This model is generative at document level and word level. The basic idea of this model is documents are represented as random mixtures over latent topics, where each topic is a distribution over words. Liang et.al, 2013 [8], discussing about a new approach that is called AS-LDA. This is a new method for sentiment classification. The paper explained words in the documents consists of two parts; sentiment words and auxiliary words. The experimental results show that this approach outperforms LDA.

B. Text Classification Methods

In this section, most popular text classifiers are discussed on the basis of existing literature work. Three text categorization techniques SVM, Naïve Bayes and KNN are explained.

Support Vector Machine (SVM):

More than eighty years ago R. A. Fisher, 1936 [9], proposed the first pattern recognition algorithm. The paper discusses the linear discriminator functions and recommends the construction of linear decision surfaces. Based on that, lots of discussions and algorithms came. In 1962 Rosenblatt [10] coined another learning algorithm on neural networks. The approach was on connected neurons, where each neuron implements a separating hyperplane. The idea to minimize the error on a set of vectors by adjusting all the weights of the network was not found at that time, so Rosenblatt suggested a method that is, only the weights of the output units are adaptive. Vapnik et.al, 1995[11], proposed a new type of learning machines called Support Vector Networks. The idea is it maps the input vectors into a high dimensional feature space. And they have constructed a linear decision surface .Here they tried to find a solution for two group classification problems. To separate a set of objects a hyperplane is introduced.

K-Nearest Neighbour (KNN):

Bhatia and Vandana, 2010 [12], have done a survey on NN techniques. Main weighted KNN, Model based KNN, and Condensed NN etc. are the main techniques included in the study. They have compared the structured and structure less NN techniques with their advantages and disadvantages. Based on the study, it is clear that structure les NN technique overcome the memory limitations and structured techniques reduces the computational complexity. Trstenjak et.al, 2013 [13], proposed a possibility of using a KNN with TF-IDF method for text classification. The evaluation is based on the speed, accuracy and quality of classification. The results include both good and bad features. The main motivation for this paper was to develop concept frameworks with emphasis on KNN&TF-IDF module. The evaluation of framework was performed in an online environment.

Naïve Bayes Classifier

Rennie et.al, 2003 [14], proposed heuristic solutions to the problems of Naïve Bayes classifiers, addressing systematic issues and the problems related to the multinomial model also. In this paper they have conducted a detailed study on the reasons behind the poor performance of Naïve Bayes. Also, recommended some corrections like conversion of text, solving issues of uneven training data and normalizing the classification weights. Yousef et.al, 2006 [15], discuss about a model for micro-RNA gene prediction. This technique is based on machine learning and uses naïve Bayes classifier. The proposed model discusses a new technique, which is applicable for many species in predicting the miRNA genes. The study is based on machine learning using Naïve Bayes to generate a model from training data.

III. DETAILED STUDY OF TOPIC MODELLING METHODS

This section discusses with some of the topic modeling methods which explored based on the

characteristics and features. A general idea has given for each of the model and some of the applications are also mentioned.

1. Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) is also called as Latent Semantic Indexing when used in the context of information retrieval. Deerwester et.al, [3] proposed this method in 1990. The basic idea is the analysis of hidden semantics in text corpora. Collection of documents can be represented as a term-document matrix. Model will infer how close the documents are and how close the document to the query. But the model suffers from two problems they are polysemy and synonymy. Sometimes a word has different meanings in different contexts. This is polysemy and synonymy is two or more words pointing same concept. LSA solves this problem by mapping the same documents or words into a different space and doing the comparison in the space. This is done by Singular Value Decomposition (SVD) of term document matrix. LSA uses Singular Value Decomposition (SVD) which uses a matrix to rearrange and calculate all the dimensions of vector space. The mapping of high dimensional document vector to low dimension space must be linear and based on SVD of co-occurrence matrix. LSA is performed by some steps. First step is collecting all the relevant texts then divide it by documents. Second step involves the creation of a terms-document matrix and each column represented by documents number and each row by words. Next distance will be calculated on semantic space. LSI is one of the dimension reduction methods on texts.

While applying LSA, it is necessary to have big text corpora. The dataset should be large enough. LSA has widely used in a variety of languages, educational technology applications and mainly in information retrieval process. LSA suffers from few drawbacks such as computational cost in getting SVD, large memory resources required and re-calculation of the whole decomposition for inclusion of new documents.

2. Probabilistic Latent Semantic Analysis (PLSA)

The heart of PLSA is a statistical model called aspect model. An aspect model is nothing but a latent variable model for co-occurrence of data. PLSA was introduced by Jan Puzicha and Thomas Hofmann in 1999 [5].This method is based on the probabilistic methodology with a latent layer and a strong statistical background. In 1999, they proposed PLSI to fix some problems found in the LSA model [5]. The main idea in this method is recognizing and distinguishing between different context of word usage without recourse to a dictionary or thesaurus. PLSA is recommended as the first probabilistic methodology and considers as the probabilistic version of LSA. It has two important implications: First one is it allows us to disambiguate polysemy. Second it exposes topic similarities by grouping together words that are part of a common context. The graphical model of PLSA is shown in Fig. 1. It includes a latent variable z and each document is considered as mixture of latent k levels. Compute z according to the conditional probability d. D is the total number of documents. Next step is generating the word w according to the conditional probability z.

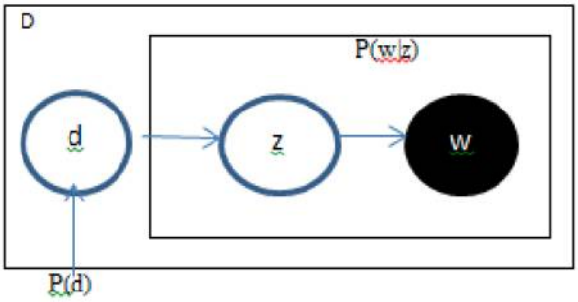

Fig. 1 PLSA model

One of the important applications is image retrieval. PLSA uses its visual features to represent image as a collection of visual words from a finite visual vocabulary. There are some other successful real-world applications; they are computer vision and recommender systems.

3. Latent Dirichlet Allocation (LDA)

LDA provides generative models that explain how documents are created and how each document obtains its words. LDA is a generative probabilistic model and it is a three-level hierarchical Bayesian model. In 2002, D. M. Blei et.al [7] proposed this model. It is based on PLSI, but compared with PLSI; LDA is a complete probabilistic generative model. The algorithm considers each document is a mixture of multiple topics, and each topic is a collection of words.In LDA each document is models as a collection over k topics, and each topic describes a multinomial distribution over a word w. The joint formula for parameters α and β, the combined distribution of θ, a set of N topics z and a set of N words w is given

$$p(\theta,z,w|\alpha,\beta)=p(\theta|\alpha)\sum_{n=1}^{N} p(zn|\theta)p(wn|zn,\beta)$$

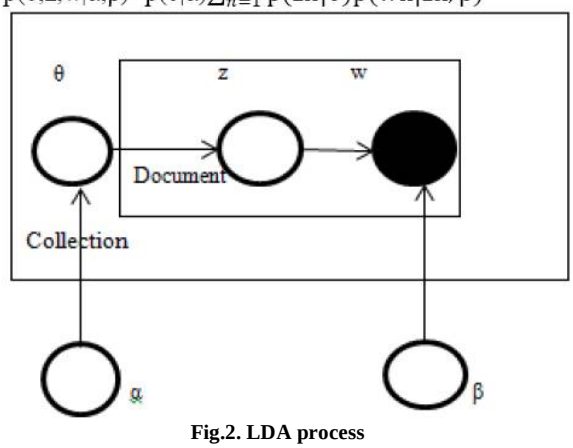

Fig.2. LDA process

In Fig. 2 it is clear that the LDA model has three layers such as document collection layer, document layer and the word layer. The three parameters α and β are arranged in these layers. Where α represents latent topics in the document collection and β represents probability distribution of latent topics. Θ represents topic distribution and the basic key idea of LDA model is to select topic vector θ. The generative process of LDA is given as follows.

1) Randomly choose θ ~ Dir(α)
2) Choose φ ~Dir(β)

For each word in the text
a) Choose a topic z ~Multinomial (θ)
b) Choose a word Wn~ Multinomial (φz)

The main idea of the algorithm is to choose a topic vector θ and the word vector φ.θ topic is a distribution over a vocabulary. In LDA, a document can generate multiple topics, and it is possible to assign probability to documents outside the corpus by using some inference methods. There are two inference methods; variational inference and Gibbs sampling.

IV. DETAILED STUDY OF TEXT CLASSIFICATION TECHNIQUES

A detailed analysis has done on three text categorization techniques based on the characteristics, advantages and disadvantages. Some of the applications are also outlined in this section.

1. K-Nearest Neighbour (KNN)

KNN is a widely used classification algorithm that is because of its simplest nature. But it gives highly competitive results. It can be used not only for classification but also for regression. KNN is a non-parametric lazy learning algorithm [12]. Non parametric algorithm does not make any assumptions on the underlying data distribution. Lazy learning means that there is no explicit training phase. i.e., the training phase is very fast. KNN makes all of its decisions based on the entire training data. Whenever we have a new point to classify, we find its K nearest neighbors from the training data. It assumes that the data is in feature space, means that data points are in a metric space. Similarity of two points is the distance between them in a metric space. This can be calculated by any of the following measures, Euclidean, Minkowski, Hamming Distance and Mahalanobis Distance. We are assuming a value to k; the value decides how many neighbors influence the classification. The choice of parameter k plays very important role in this algorithm. The process of KNN involves some steps. To calculate the number of K nearest neighbors, find the parameter k. In the next step calculate the distance between the query-instance and the training data. Next, find the nearest neighbors based on the Kth minimum distance and gather all the nearest neighbors together and avoid all the other nearest neighbors which do not satisfy the k value. Finally use this majority nearest neighbors for the predictive purpose. When the number of classes is two, typically k is odd. If the training data is large, KNN is effective.It is based on the dataset that if the dataset is huge KNN works well and gives effective classification results.

2. Naïve Bayes

Naïve Bayes classifier is based on Bayes theorem. This classification model is very useful for very large datasets. It is one of the most popular, widely used model and easy to build [14]. This is a simple algorithm for modeling classifiers but not a single algorithm, a family of algorithms. This is based on the main idea that is it assumes the value of a particular feature is independent from the value of any other feature. Naïve Bayes are linear classifiers.The main function is to classify new cases as they arrived or to decide to which class label they belong, based on the currently existing objects. It requires only small amount of training data to estimate parameters. It can handle an arbitrary number of independent variables whether continuous or categorical. The algorithm gives a way to compute the posterior probability. Bayes theorem forms the backbone of Naïve Bayes classification. Spam filtering is one of the best applications of Naïve Bayes text classifier. It is very easy and fast to predict class of test data set.

Naïve Bayes main strength is its efficiency and it combines efficiency with good accuracy. This is used as the main baseline in text categorization research. Along with the pros there are some cons also. If category variable has a category in the test data, which was not there in the training data, then the model will assign a zero probability and will be unable to make a prediction. To overcome this there are smoothing techniques that will solve this problem. On the other side this model is also known as a bad one. So the probability results are not to be taken as serious.

3. Support Vector Machine (SVM)

Support Vector Machine is a vector space based learning methods. They are used for both classification and regression process. This model is based on the idea of decision plane that separates a set of objects having multiple class boundaries. The process is segregating the two classes with a hyper-plane. The data points that lie close to the hyperplane are called support vectors. SVM's maximizes the space around the hyperplane. The distance between nearest data point and hyper plane will help to decide the optimal hyper-plane and have to select a hyperplane with high margin [11]. This is known as the margin. In Fig.1, The small rounds are separated from stars by an optimal hyperplane B. SVM is effective in high dimensional space. But it doesn't

perform well with large noisy data set and it needs more training time.

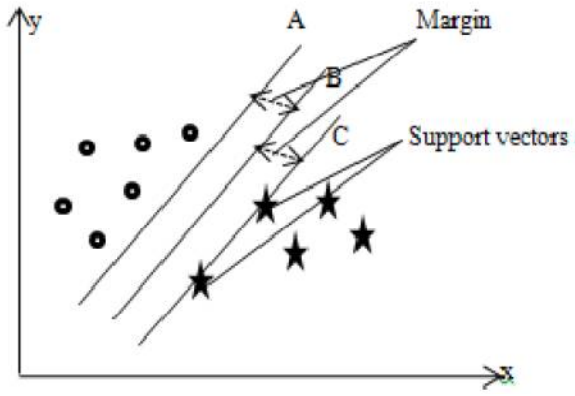

Fig. 3 Support Vectors with optimal margin

SVM has a technique called the kernel trick. This transforms low dimensional input space to higher dimensional space. It is mostly useful in non-linear classification problem. This model V. Combined approach of Topic Modeling and Text classification

The combined approach of these two categories will provide a new way to advanced text mining. First discusses the differences between these two categories. The main difference is that topic modeling is an unsupervised learning and text categorization is a supervised learning. Topic modeling is not mutually exclusive; the same document can have its probability distribution spread across many topics where the second category is mutually exclusive. Text classification involves set of classes known in advance and don't change. The first category changes its topics and gives different topics each time when run the algorithm.Text classification is a bag of words approach, where topic modeling refers the correlations between the words. Including the differences, the combination of text classifiers and topic models yields better results.

In 2011,Liu et.al, [16], evaluated the performance of three topic models such as LSA, PLSA and LDA. They explained that the performance of LDA is very good on topic extraction. Based on the experiments conducted, the paper argues that LDA overcomes all the drawbacks of LSA and PLSA and improves the topic detection and can work on huge datasets. Liu et.al, 2015[17] showed the comparison process of four text classifiers. They have compared and analysed the classifiers to find the accuracy level of each classifier. And they conclude that among the four SVM shows superiority over the others.

Based on the studies done by the researchers, it is clear that the LDA shows good performance than the other topic models and SVM yields better accuracy than the other classifiers. Some researchers have proven that the blend of these two can increase the performance level.In 2011, WongkotSriurai[18] proved that topic model approach can improve the performance of text classification. In this paper, he used topic model approach to cluster the words into a set of topics and compared this withthree text categorization algorithms, Naïve Bayes, Support Vector and Decision Tree. Based on the experiments conducted by him, the combination yielded better performance. Li et.al, 2016[19], verified that the text classification based on topic model can reduce the text dimension and get the features easily. Also they argue that traditional text classification techniques are unable to meet the needs. They proposed a LDA based SVM classification model to categorize the news articles and it gives good classification results. TABLE I and TABLE II demonstrates the basic ideas of both categories.

TABLE I. Characteristics of Topic model methods.

Topic model	Characteristics	Advantages	Disadvantages
Latent Semantic Analysis (LSA)	1. Examines words used in a document with same or similar meaning. 2. Low-dimension representation of documents and words. 3. Creating latent semantic space.	1. Can cluster the words and documents in the space, so we can retrieve 2. Trying to solve the problems of polysemy an synonymy. 3. Error reduction by dimension reduction.	1. Not giving well defined probabilities.
Probabilistic Latent Semantic Analysis (PLSA)	1. Uses probabilistic method. 2. Implemented for the automated document indexing 3. Strong statistical background.	1. Can be easily extendable and embedded in other models because of its probabilistic nature. 2. It handles polysemy.	1. Overfitting is occurring and the efficiency is poor. 2. More number of free parameters is available.
Latent Dirichlet Allocation (LDA)	1. Generative probabilistic model. 2. Need to remove stop words manually. 3. Words are grouped into topics and can exist in more than one topic. 4. Bayesian version of PLSA.	1. Works well with large corpus. 2. Does not suffer from overfitting issues like in PLSA. 3. Can be embedded in other complicated models. 4. Noise reduction is possible by dimension reduction.	1. Fixed k, the number of topics should set in advance. 2. Uncorrelated topics.

TABLE II. Characteristics of Text Classification Methods

Classifier	Characteristics	Advantages	Disadvantages
KNN	1. Assumes data is in a feature space 2. It is a non-parametric lazy algorithm	1. Simple and easy to learn. 2. Training data is a very fast process.	1. Need to find the value of k 2. Computational cost is high. 3. Knn is a lazy learner. 4. Memory space is limited.
Naïve Bayes	1. Strong statistical background 2. Assumes value of a feature is independent of the value of any other feature.	1. This algorithm is very fast. 2. Need only less training data. 3. Very easy to implement.	1. It makes very strong independence assumptions ie, class conditional independence, therefore loss of accuracy.
SVM	1. A hyperplane separates two classes with largest separation or margin. 2. Kernel trick is applied to create nonlinear classifiers.	1. Uses kernel trick. 2. Good theoretical guarantees and avoid overfitting. 3. Memory efficient. 4. Effective in High dimensional space. 5. Good classification accuracy. 6. Works well with small train sets.	1. Needs more training time. 2. If the data contains more noise, performance will be poor.

CONCLUSION

Recent years the amount of unstructured text is increased. It becomes a necessary to organize the data. Text classifiers and topic models perform a vital role in organization of data. This paper surveys on topic models and text categorization. Focuses on the analysis of existing literature and explored the characteristics, pros and cons of both categories.Paper does not go into specific details.Every model has good characteristics in particular area under particular circumstances. Furthermore, it has mentioned that the combination of these two categories will achieve good text classification results. More specifically SVM-LDA combination shows better results than others.From the discussion it is understood that no topic model and classification model can be referred as a general model for any application.Different classification techniques perform differently depending on the dataset.

REFERENCE

[1] John D Lafferty,David M Blei, “Dynamic Topic Models,” International Conference on Machine Learning (ICML),2009.
[2] Blei D M,Griffiths T,Jorddan M,Tannenbaum J,“Hierarchical Topic Models and Nested Chineese Restaurant Process,” proceedings ofAdvanced in Neural Information ProcessingSystem, Cambridge, 2004.
[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K.Landauer and R. Harshman, “Indexing by latent semantic analysis,” Journal of the American Society fro Infromation Science, vol. 41, no. 6, pp. 391-407, 1990.
[4] Susan T. Dumais, Michael L. Littman, Thomas K.Landauer, “Automatic Cross-Langauage Retrieval using Latent Semantic Indexing,” AAAI TechnicalReport, pp. 18-24, 1997.
[5] T.Hofman, “Unsupervised Learning by Probabilistic Latent Semantic Analysis,” Machine Learning, vol. 42, pp. 177-196, 2001.
[6] Incheon Paik, Shinsuke Mori and Wuhui Chen, “Semantic Words similarity n triple relation usingintermediate concept by PLSA,” IEEE International Conference proceedings, 2010.
[7] Blei, D.M., Ng, A.Y., and Jordan, M.I., “Latent Dirichlet Allocation,” Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.
[8] Jiguang Liang, Ping Liu, Jianlog Tan and Sheo Bai, “Sentiment Classification based on AS-LDA model,” Procedia Computer Science, vol. 31, pp.511-516, 2014.
[9] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” The Annals ofEugenics, vol. 7, pp. 179-188, 1936.
[10] F. Rosenblatt, "principles of Neurodynamics,"Spartam Books, New York, 1962 .
[11] V. Vapnik, “The Nature of Sstatistical Learning Theory,” Springer, 1995.
[12] Nithin Bhatia and Vandana, “Survey of Nearest Neighbor Techniques,” International Journal of Computer Science and Information Security IJCSIS, vol. 8, 2010.
[13] Bruno Trstenjak, Sasa Mikac and DzenanaDonko, “KNN with TF-IDF based Framework for Text Categorization,” In 24th DAAAM InternationalSymposium on Intelligent Manufacturing and Automation, 2014.
[14] Jason D. M. Rennie, Lawrence Shih, Jaime Teevan and David R. Karger, “Tackling the poor assumptions of Naive Bayes Text Classifier,” In 20 th International Conference on Machine Learning, ICML, 2003.
[15] Malik Yousef, Michael Nebozhyn, Hagit Shatkay, Stathis Kanlerakis, Louise C. Showe and Michael K.Showe, “Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier.,”Bioinformatics, vol. 22, pp.1325-1334,2006.
[16] Wongkot Sriurai, “Improvin Text categorization by using a topic model”, Advanced Computing: An International Journal(ACIJ),Vol. 2, No.6, Nov. 2011.
[17] Zhenzhong Li, Wenqian Shang and Menghan Yan, “News Text classification model based on topic model”, IEEE, June 2016.
[18] Maofu Liu, Hezhang, Huijun Hu and Wei Wei, “Topic categorization and representation of health community generated data”, Multimedia Tools Applications, Nov,2015.
[19] Zelong Liu, Maozhen Li, Yang Liu and Mahesh Ponraj, “Performance Evaluation of Latent Dirichlet Allocation in Text Mining”, In the proceedings of Fuzzy Systems and Knowledge Discovery(FSKD). China, 2011.