ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

A Survey on Text Classification with Different Types of Classification Methods

Monica Bali, Deipali Gore

Dept. of Computer Engineering, PES's MCOE, Savitribai Phule Pune University, Pune, Maharashtra, India

Assistant Professor, Dept. of Computer Engineering, PES's MCOE, Savitribai Phule Pune University, Pune, Maharashtra, India

ABSTRACT: Text classification approach gaining more importance because of the accessibility of large number of electronic documents from a variety of resource. Text categorization (Also called Text Categorization) is the task of assigning predefined categories to documents. It is the method of finding interesting regularities in large textual, where interesting means non trivial, hidden, previously unknown and potentially useful. The goal of text mining is to enable users to extract information from textual resource and deals with operation such as retrieval, classification, clustering, data mining, natural language preprocessing and machine learning techniques together to classify different pattern. A major characteristic or difficulty of text categorization is high dimensionality of feature space. The reduction of dimensionality by selecting new attributes which is subset of old attributes is known as feature selection. Feature-selection methods are discussed in this paper for reducing the dimensionality of the dataset by removing features that are considered irrelevant for the classification. This paper surveys of text classification, several approaches of text classification, feature selection methods and applications of text classification.

KEYWORDS: Information Retrieval, Text Classification, Text Mining, Feature Selection.

I. INTRODUCTION

Today huge amount of information are being associated with the web technology and the internet. To gather useful information from it these text has to be categorized. The task to classify a given data instance into a pre-specified set of categories is known as "text categorization" (TC). Given a set of categories (subjects, topics) and a collection of text documents, it is the process of finding the correct subject (or subjects) for each document.

The expert's knowledge about the categories is directly used to categorize the documents. Most of the recent work on categorization is concentrated on approaches which require only a set of manually classified training instances that are much less costly to produce. A classifier is built by learning from a set of pre-classified examples. One of the drawbacks of supervised approaches is that they need to be trained on predefined positive and negative test samples or predefined categories. Efficiency of these models depends on the quality of the sample sets. With the enormous amount of data and different type of applications, it is not always possible to create these training sets or contextual categories manually.

TC may be formalized as the task of approximating the unknown target function Φ: $D \times C \rightarrow \{T, F\}$ (that describes how documents should be classified, according to a supposedly authoritative expert) by means of a function called the classifier, where C = {c1… c|C|} is a predefined set of categories and D is a (possibly infinite) set of documents. If Φ (dj, ci) = T, then dj is called a positive example (or a member) of c_i, while if Φ (dj, ci) = F it is called a negative example of ci.

There are two types of approaches to text categorization: rule based and machine learning based approaches. Rule based approaches mean ones where classification rules are defined manually and documents are classified based on rules. Machine learning approaches mean ones where classification rules or equations are defined automatically using

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

sample labeled documents. This class of approaches has much higher recall but a slightly lower precision than rule based approaches. Therefore, machine learning based approaches are replacing rule based one for text categorization.

II. TEXT CLASSIFICATION

Text classification is a fundamental task in document processing. The goal of text classification is to classify a set of documents into a fixed number of predefined categories/classes. A document may belong to more than one class. When classifying a document, a document is represented as a "bag of words". It does not attempt to process the actual information as information extraction does. Rather, in simple text classification task, it only counts words (term frequency) that appear and, from the count, identifies the main topics that the document covers e.g. if in the document, cricket word comes frequently then "cricket" is assigned as its topic (or class) [1] [2]. There are two phases in the Classification. They are Training Phase and Testing phase.

A. *Training Phase:*

It is also called as Model Construction or Learning Phase; the set of documents used for model construction is called training set. It describes a set of predetermined classes. Each document/sample in the training set is assumed to belong to a predefined class (labelled documents). The model is represented as classification rules, decision trees, or mathematical formulae [2] [3].

B. *Testing Phase:*

This is the 2nd step in classification and also called Mode Usage or Classification Phase. It is used for classifying future or unlabelled documents. The known label of test document/sample is compared with the classified result to estimate the accuracy of the classifier. For e.g. the labelled documents of the training set, is used further to classify unlabelled documents. Test set is independent of training set [1] [2] [3].

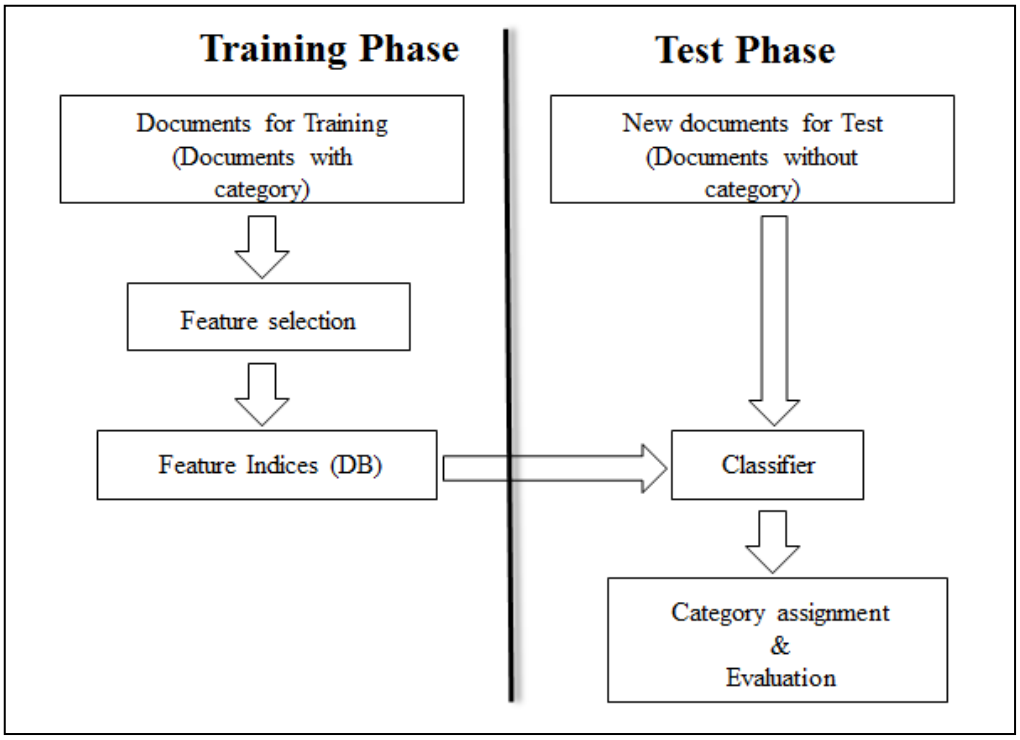

Fig. 1. Flow Diagram of Text Classification

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

Using supervised learning algorithms [4], the objective is to learn classifiers from known examples (labelled documents) and perform the classification automatically on unknown examples (unlabelled documents). Figure 2 shows the overall flow diagram of the text classification task. Consider a set of labelled documents from a source D= [d1, d2, d3…dn] belonging to a set of classes C = [c1, c2, c3…cp]. The text classification task is to train the classifier using these labelled documents, and assign categories/classes to the new, unlabelled documents. In the training phase, the n documents are arranged in p separate folders, where each folder corresponds to one class. In the next step, the training data set is prepared via a feature selection process [1] [2]. Text data typically consists of strings of characters, which are transformed into a representation suitable for learning. It is observed from previous research that words work well as features for many text categorization tasks. In the feature space representation, the sequences of characters of text documents are represented as sequence of words. Feature selection involves tokenizing the text, indexing and feature space reduction. Text can be tokenized using term frequency (TF), inverse document frequency (IDF), term frequency inverse document frequency (TFIDF) or using binary representation. Using these representations the global feature space is determined from entire training document collection.

III. CLASSIFICATION METHODS

A. *K-Nearest Neighbor:*

KNN is a classification algorithm where objects are classified by voting several labelled training examples with their smallest distance from each object. It was first described in 1950 but initially applied to classification of news articles by Massand et al. in 1992 [5]. Yang compared 12 approaches to text categorization with each other, and judged that KNN is one of recommendable approaches, in 1999 [6]. Sebastiani in 2002 [7] evaluated as a simple and competitive algorithm with Support Vector Machine for implementing text categorization systems. The Major disadvantage of KNN is that it uses all features in computing distance and costs very much time for classifying objects.

The classification itself is usually performed by comparing the category frequencies of the k nearest documents (neighbors). The evaluation of the closeness of documents is done by measuring the angle between the two feature vectors or calculating the Euclidean distance between the vectors. In the latter case the feature vectors have to be normalized to length 1 to take into account that the size of the documents (and, thus, the length of the feature vectors) may differ. A doubtless advantage of the k-nearest neighbor method is its simplicity. It has reasonable similarity measures and does not need any resources for training. K -nearest neighbor performs well even if the category-specific documents from more than one cluster because the category contains, e.g., more than one topic. This situation is badly suited for most categorization algorithms [8]. Its disadvantage is that KNN requires more time for classifying objects when a large number of training examples are given. KNN should select some of them by computing the distance of each test objects with all of the training examples.

B. *Decision Trees:*

Decision tree methods reconstruct the manual classification of the training documents by constructing well-defined true/false queries in the form of a tree structure where the nodes represent questions and the leaves represent the corresponding category of documents. After having created the tree, a new document can easily be classified by putting it in the root node of the tree and let it run through the query structure until it reaches a certain leaf. The main advantage of decision trees is the fact that the output tree is easy to interpret even for persons who are not familiar with the details of the model [9]. The tree structure generated by the model provides the user with a consolidated view of the classification logic and is therefore useful information. A risk of the application of tree methods is known as "over fitting": A tree over fits the training data if there exists an alternative tree that classifies the training data worse but would classify the documents to be categorized later better. This circumstance is the result of the algorithm's intention to construct a tree that categorizes every training document correctly; however, this tree may not be necessarily well suited for other documents. This problem is typically moderated by using a validation data set for which the tree has to perform in a similar way as on the set of training data. Other techniques to prevent the algorithm from building huge trees (that anyway only map the training data correctly) are to set parameters like the maximum depth of the tree or theminimum number ofobservations in a leaf. If this is done, Decision Trees show verygood performance even for categorization problems with a verylarge number of entries in the dictionary.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

C. *Bayesian Approaches:*

There are two groups of Bayesian approaches in document categorization: Naïve [10] and non-naïve Bayesian approaches. The naïve part of the former is the assumption of word (i.e. feature) independence, meaning that the word order is irrelevant and consequently that the presence of one word does not affect the presence or absence of another one. A disadvantage of Bayesian approaches [11] in general is that they can only process binary feature vectors and, thus, have to abandon possibly relevant information.

D. *Neural Networks:*

Neural networks consist of many individual processing units called as neurons connected by links which have weights that allow neurons to activate other neurons. Different neural network approaches have been applied to document categorization problems. While some of them use the simplest form of neural networks, known as perceptions, which consist only of an input and an output layer, others build more sophisticated neural networks with a hidden layer between the two others. In general, these feed-forward -nets consist of at least three layers (one input, one output, and at least one hidden layer) and use back propagation as learning mechanism. The advantage of neural networks is that they can handle noisy or contradictory data very well. The advantage of the high flexibility of neural networks entails the disadvantage of very high computing costs. Another disadvantage is that neural networks are extremely difficult to understand for an average user; this may negatively influence the acceptance of these methods.

E. *Vector-based Methods:*

There are two types of vector-based methods: The centroid algorithm and support vector machines. One of the simplest categorization methods is the centroid algorithm. During the learning stage only the average feature vector for each category is calculated and set as centroid-vector for the category. A new document is easily categorized by finding the centroid-vector closest to its feature vector. The method is also inappropriate if the number of categories is very large. Support Vector Machines (SVM) need in addition to positive training documents also a certain number of negative training documents which are untypical for the category considered. SVM is then looking for the decision surface that best separates the positive from the negative examples in the n-dimensional space. The document representatives closest to the decision surface are called support vectors. The result of the algorithm remains unchanged if documents that do not belong to the support vectors are removed from the set of training data. An advantage of SVM [12] is its superior runtime-behaviour during the classification of new documents because only one dot product per new document has to be computed. A disadvantage is the fact that a document could be assigned to several categories because the similarity is typically calculated individually for each category.

IV. FEATURE SELECTION METHODS

Feature-selection methods play a very important role in the reduction of the dimensionality of the dataset by removing features that are considered irrelevant for the classification [13]. These feature selection methods possess a number of advantages such as smaller dataset size, smaller computational requirements for the text categorization algorithms (especially those that do not scale well with the feature set size) and considerable shrinking of the search space. The goal is the reduction of the curse of dimensionality to yield improved classification accuracy. Another benefit of feature selection is its tendency to reduce over fitting, i.e. the phenomenon by which a classifier is tuned also to the contingent characteristics of the training data rather than the constitutive characteristics of the categories, and therefore, to increase generalization. Best Individual Features can be performed using some of the measures, for instance, document frequency, term frequency, mutual information, information gain, odds ratio, χ^2 statistic and termstrength [13], [14], [15], [16], [17]. What is common to all of these feature scoring methods is that they conclude by ranking the features by their independently determined scores, and then select the top scoring features.

A. χ^2 *statistic:*

In experimental sciences, chi-square statistics is frequently used to measure how the observation results differ from the expected results. In other words, it measures the independence of two random variables.

$$CHI = \sum_{ij} \frac{(Observed_{ij} - Expected_{ij})^2}{Expected_{ij}}$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

Chi-square statistics is also widely used in text categorization [7]. In text categorization, the two random variables are occurrence of term t_k and occurrence of class c_i and chi-square statistics measures the independence between t_k and c_i. The formula for chi-square score is:

$$CHI(t_k, c_i) = N\, x\, \frac{[P(t_k, c_i)P(\overline{t_k},\ \overline{c_i}) - P(\overline{t_k}, c_i)P(t_k,\ \overline{c_i})]^2}{P(t_k)P(\overline{t_k})P(c_i)P(\overline{c_i})}$$

Where $P(t_k)$is the percentage of documents in which term t_k occurs, $P(\overline{t_k})$is the percentage of documents in which term t_k does not occur, $P(c_i)$is the percentage of documents belonging to class c_i, $P(\overline{c_i})$is the percentage of documents not belonging to class c_i, $P(t_k, c_i)$is the percentage of documents belonging to class $c_i$in which term t_k occurs, $P(\overline{t_k},\ \overline{c_i})$is the percentage of documents not belonging to class $c_i$in which term t_k does not occur, $P(\overline{t_k}, c_i)$is the percentage of documents belonging to class c_i in which term t_k does not occur and $P(t_k,\ \overline{c_i})$ is the percentage of documents not belonging to class $c_i$in which term t_k occurs.

If chi-square score of a term t_k is low value, this means is independent from the class $c_i$and if chi-square score of a term t_k is high value, this means t_k is dependent of the classc_i. Thus the chi-square feature selection method selects the terms with the highest chi-square score which are more informative for classification.

B. *Information Gain:*

Another popular feature selection method in text categorization is information gain [7]. It is measure the decrease in entropy by existence or absence of the term in a document. Information gain score will be null for two independent variables and it will be high because of the dependence between two variables. The information gain score of a term t_k is calculated by the formula:

$$IG(t_k, c_i) = \sum_{C \in \{c_i, \overline{c_i}\}} \sum_{t \in \{t_i, \overline{t_i}\}} P(t, c).\, log(\frac{P(t, c)}{P(t).\, P(c)})$$

Information gain feature selection method selects the terms with the highest information gain scores which contains much information about the classes.

C. *Document Frequency:*

Document frequency is a very simple and popular method that measures the number of documents in which the term appears without class labels [7]. Purpose of the method is to eliminate the rare words which are assumed non-informative and misleading for classification. Document frequency feature selection method selects the terms with the scores. Its formula is:

$$DF(t_k, c_i) = P(t_k, c_i)$$

D. *Mutual information:*

Mutual information is a criterion commonly used in statistical language modelling of, word associations and related applications [18].This is able to provide a precise statistical calculation that could be applied to a very large corpus to produce a table of association of words. If one considers a two way contingency table of a term t and a category c. where A is number of times c and t co-occur. B is the number of times t occur without c. C is number of times c occur without t, N is the total number of documents, then mutual information criterion between t and c is defined to be

$$I(t, c) = log \frac{P_r(t \hat{} c)}{P_r(t)P_r(c)}$$

I (t, c) has a natural value of zero if t and s are independent. A weakness of the mutual information is that score is strongly influenced by the marginal probabilities of the terms.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

E. *Term strength:*

Term strength is originally proposed and evaluated by Wilbur and Sirotkin [19] for vocabulary reduction in text retrieval and later applied by Yang and Wilbur to text categorization. This method estimates term importance based on how commonly a term is likely to appear in closely related documents. It uses a training set of documents to derive document pairs whose similarity is above threshold. Let x and y be arbitrary pair of distinct but related documents and t may be a term therefore the term strength may be defined as

$$S(t) = P_r(t \in y | t \in x)$$

V. APPLICATIONS OF TEXT CLASSIFICATION

There are many potential applications of text categorization. The following are a few examples of its applications.

A. *Document Organization:*

An instance of document organization is document indexing with a controlled dictionary, such as the ACM Classification Scheme [20]. This is an automatic indexing of scientific articles by means of a controlled dictionary, where the categories are the entries of the controlled dictionary. In the case of digital libraries, documents are usually indexed by thematic metadata that describe their semantics with controlled vocabulary (e.g. keywords, key phrases, bibliography codes). Another example of document organization is classification of News articles and ads. This type of problem can be tackled by TC techniques.

B. *Spam filtering:*

A persistent problem for Internet service providers and users is the deluge of spam, the unsolicited bulk messages indiscriminately sent by spammers. Because of the huge volume of junk mail, extra capacity or cost has to be added to handle the flood. The most widely recognized form of spam is e-mail spam. Text classification systems [21] can classify incoming e-mail as negative (non-spam) or positive (i.e. spam) and reject those that they finds to be spam. A challenge with spam filtering applications is the lack of negative examples. While spam messages are everywhere, non-spam messages are hard to collect because of the privacy issues. The unbalanced distribution of data examples should be addressed by TC algorithms.

C. *Hierarchical categorization of Web pages:*

Due to the tremendous increase of the amount of Web pages or sites, it is more and more difficult to find the information we are interested in. Classifying Web pages or sites under hierarchical catalogues can make a Web search easier by restricting the search to a particular category of interest. While manual categorization of Web pages is infeasible and costly to maintain, TC methods [22] can be employed to do the job automatically.

D. *Text filtering:*

Text filtering is the activity of classifying a stream of incoming documents dispatched in an asynchronous way by an information producer to an information consumer [Belkin and Croft 1992]. A typical case is a newsfeed filter [23], where the producer is a news agency and the consumer is a newspaper. In this case, the filtering system should block the delivery of the documents the consumer is likely not interested in (e.g., all news not concerning sports, in the case of a sports newspaper). Filtering can be seen as a case of single-label TC, that is, the classification of incoming documents into two disjoint categories, the relevant and the irrelevant additionally, a filtering system may also further classify the documents deemed relevant to the consumer into thematic categories; in the example above, all articles about sports should be further classified according to which sport they deal with, so as to allow journalists specialized in individual sports to access only documents of prospective interest for them. Similarly, an e-mail filter might be trained to discard “junk” mail [24] and further classify non junk mail into topical categories of interest to the user. A filtering system may be installed at the producer end, in which case it must route the documents to the interested consumers only, or at the consumer end, in which case it must block the delivery of documents deemed uninteresting to the consumer.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

VI. CONCLUSION AND FUTURE WORK

Text Classification is an important application area in information retrieval, text mining and machine learning why because classifying millions of text document manually is an expensive and time consuming task. Therefore, automatic text classifier is constructed using pre classified sample documents whose accuracy and time efficiency is much better than manual text classification. If the input to the classifier is having less noisy data, we obtain efficient results. So during mining the text, efficient pre-processing algorithms must be chosen. The test data also should be pre-processed before classifying it. Text can be classified better by identifying patterns. Once patterns are identified we can classify given text or documents efficiently. Identifying efficient patterns also plays major role in text classification. Text classification techniques need to be designed to effectively manage large numbers of elements with varying frequencies. Various methods for text classification are discussed in this paper. Feature selection methods are able to successfully reduce the problem of dimensionality in text classification applications.Future work is required to improve the performance and accuracy of the text classification process. From the above discussion it is understood that no single classifier and feature selection method can be mentioned as a general model for any application. Different algorithms perform differently depending on data collection

REFERENCES

1. Vishal Gupta, Gurpreet S. Lehal, 'A Survey of Text Mining Techniques and Applications', Journal of Emerging Technologies in Web Intelligence, VOL. 1, NO. 1, 2009.
2. Jiawei Han, Michelin Kamber, 'Data Mining Concepts and Techniques', Morgan Kaufmann publishers, USA, pp.70-181, 2001.
3. Megha Gupta, Naveen Aggrawal, 'Classification Techniques Analysis', NCCI 2010 –National Conference on Computational Instrumentation CSIO Chandigarh, INDIA, pp. 128-131, 19-20 March 2010.
4. Setu Madhavi Namburu, Haiying Tu, Jianhui Luo and Krishna R. Pattipati, 'Experiments on Supervised Learning Algorithms for Text Categorization', International Conference , IEEE computer society, pp. 1-8, 2005.
5. Massand.B, Linoff. G, Waltz. D, 'Classifying News Stories using Memory based Reasoning', the Proceedings of 15th ACM International Conference on Research and Development in Information Retrieval. pp. 59-65, 1992.
6. Yang, 'An evaluation of statistical approaches to text categorization', Information Retrieval. pp. 67-88, 1999.
7. Sebastiani.F, 'Machine Learning in Automated Text Categorization', ACM Computing Survey. pp. 1-47, 2002.
8. S. Niharika, V. Sneha Latha and D. R. Lavanya, 'A Survey on Text Categorization', International Journal of Computer Trends and Technology- volume3, Issue1, pp. 39-45, 2012.
9. D. E. Johnson, F. J. Oles, T. Zhang, T. Goetz, 'A decision-tree-based symbolic rule induction system for text categorization', IBM Systems Journal, September 2002.
10. Kim S. B., Rim H. C., Yook D. S. and Lim H. S., 'Effective Methods for Improving Naïve Bayes Text Classifiers', LNAI 2417, pp.414-423, 2002.
11. Klopotek M. and Woch M., 'Very Large Bayesian Networks in Text Classification', ICCS 2003, LNCS 2657, pp. 397-406, 2003.
12. Joachims, T., 'Transductive inference for text classification using support vector machines', Proceedings of ICML-99, 16th International Conference on Machine Learning, eds. I. Bratko & S. Dzeroski, Morgan Kaufmann Publishers, San Francisco, US: Bled, SL, pp. 200–209, 1999.
13. Forman, G., 'An Experimental Study of Feature Selection Metrics for Text Categorization', Journal of Machine Learning Research3, pp. 1289-1305, 2003.
14. Brank J., Grobelnik M., Milic-Frayling N., Mladenic D., 'Interaction of Feature Selection Methods and Linear Classification Models', Proc. of the 19th International Conference on Machine Learning, Australia, 2002.
15. Torkkola K., 'Discriminative Features for Text Document Classification', Proc.International Conference on Pattern Recognition, Canada, 2002.
16. Sousa P., Pimentao J. P., Santos B. R. and Moura-Pires F., 'Feature Selection Algorithms to Improve Documents Classification Performance', LNAI 2663, pp. 288-296, 2003.
17. Soucy P. and Mineau G., 'Feature Selection Strategies for Text Categorization', AI 2003, LNAI 2671, pp. 505-509, 2003.
18. Kennt Ward Church and Patrick Hanks, 'Word association norms, mutual information and lexicography', in proceedings of ACL 27, Vancouver, Canada, pp. 76-83, 1989.
19. J.W.Wilbur and k.sirotkin, 'The automatic identification of stop words', pp. 45-55, 1992.
20. Rijsbergen, C. J. V. Information Retrieval, 2nd ed. Butterworths, London, UK. Available at http://www.dcs.gla.ac.uk/Keith. 1979.
21. Drucker, H. Vapnik, V. and Wu, D., 'Support vector machines for spam categorization', IEEE Transactions on Neural Networks, 10(5), pp. 1048-1054, 1999.
22. Guyon, I. and Elisseeff, A., 'An Introduction to Variable and Feature Selection (Kernel Machines Section)', JMLR, 3: pp. 1157-1182, 2003.
23. Amati, G., D'Aloisi, D., Giannini, V. & Ubaldini, F., 'A framework for filtering news and managing distributed data', Journal of Universal Computer Science, 3(8), pp. 1007–1021, 1997.
24. Weiss, S.M., Apt´e, C., Damerau, F.J., Johnson, D.E., Oles, F.J., Goetz, T. & Hampp, T., 'Maximizing text-mining performance', IEEE Intelligent Systems, 14(4), pp. 63–69, 1999.

BIOGRAPHY

Monica Ramling Balireceived the B.E degree in Information Technology Engineering from T. P. C. T's College of Engineering in 2011 from BAMU, Aurangabad. She is now pursuing M.E. degree in Computer Engineering from P.E.S.'s Modern College of Engineering, Savitribai Phule Pune University.