Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Text Categorization and Machine Learning Methods: Current State of the Art

Durga Bhavani Dasari[1] and Dr. Venu Gopala Rao. K[2]

[1] Jawaharlal nehru university - Hyderabad

Received: 9 December 2011 Accepted: 1 January 2012 Published: 15 January 2012

Abstract

In this informative age, we find many documents are available in digital forms which need classification of the text. For solving this major problem present researchers focused on machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of pre classified documents, the characteristics of the categories. The main benefit of the present approach is consisting in the manual definition of a classifier by domain experts where effectiveness, less use of expert work and straightforward portability to different domains are possible. The paper examines the main approaches to text categorization comparing the machine learning paradigm and present state of the art. Various issues pertaining to three different text similarity problems, namely, semantic, conceptual and contextual are also discussed.

Index terms— Text Mining, Text Categorization, Text Classification, Text Clustering.

1 Introduction

ext categorization, the activity of labeling natural language texts with thematic categories from a set arranged in advance has accumulated an important status in the information systems field, due to because of augmentation of availability of documents in digital form and the confirms need to access them in easy ways.. Currently text categorization is applied in many contexts, ranging from document indexing depending on a managing vocabulary, to document filtering, automated metadata creation, vagueness of word sense, population of and in general any application needs document organization or chosen and adaptive document execution. These days text categorization is a discipline at the crossroads of ML and IR, and it claims a number of characteristics with other tasks like information/ knowledge pulling from texts and text mining [39,40]. "Text mining" is mostly used to represent all the tasks that, by analyzing large quantities of text and identifying usage patterns, try to extract probably helpful (although only probably correct) information. Concentrating on the above opinion, text categorization is an illustration of text mining. Along with the main point of the paper that is (i) the automatic assignment of documents to a predetermined set of categories, (ii) the automatic reorganization of such a set of categories [41], or (iii) the automatic identification of such a set of categories and the grouping of documents under each categories [42], a task generally called text clustering, or (iv) any activity of placing text items into groups, a task that has two text categorization and text clustering as certain illustrations [43]. The agile developments of online information, text categorization become one of the key techniques for dealing and arranging text data.

Text categorization techniques are helpful in to classifying news stories, discovering intriguing information on the WWW, and to guide a user's search through hypertext. Since constructing text classifiers manually is difficult and time-taking so it is beneficial of learning classifiers through instances.

2 II.

3 Text categorization

The main aim of text categorization is the classification of documents into a fixed number of predetermined categories. Every document will be either in multiple, or single, or no category at all. Utilizing machine learning, the main purpose is to learn classifiers through instances which perform the category assignments automatically. This is a monitored learning problem. Avoiding the overlapping of categories every category is considered as a isolated binary classification problem.

Coming to the process the first step in text categorization is to transform documents, which typically are strings of characters, into a representation opt for the learning algorithm and the classification task. The research in information retrieval advices that word stems performs like representation units where their ordering in a document is not a major for many tasks which leads to an attribute value representation of text. Every distinct word has a feature, with the number of times word occurs in the document as its value. For eliminating dispensable feature vectors, words are taken as features only if they occur in the training data at least 3 times and if they are not "stop-words" (like "and", "or", etc.).

The representation scheme giuides to very highdimensional feature spaces consisting of more than 10000 dimensions. Many have recognized that the need for feature collection and choice is to make the use of conventional learning methods possible, to develop generalization accuracy, and to avoid "over fitting". The recommendation of [11], the information accumulated(D D D D)

C criterion are used in the paper to choose a subset of features.

Subsequently, from IR it is clear that scaling the dimensions of the feature vector with their inverse document frequency (IDF) [8] develops performance. At present the "tfc" variant is used. To abstract from different document lengths, each document feature vector is reduced to unit length.

4 III.

5 Taxonomy of Text Classification process

Sebastiani discussed a wonderful review of text classification domain [25]. Hence, in the present work along with the brief description of the text classification a few recent works than those in Sebastiani's article including few articles which are not mentioned by Sebastiani are also discussed. In Figure **??** the graphical representation of the Text Classification process is shown.

6 Fig. 1: Taxonomy of the Text Classification Process

The task of building a classifier for documents does not vary from other tasks of Machine Learning. The main point is the representation of a document [16]. One special certainty of the text categorization problem is that the number of features (unique words or phrases) reaches orders of tens of thousands flexibly. This develops big hindrances in applying many sophisticated learning algorithms to the text categorization, so dimension reduction methods are used which can be used either in choosing a subset of the original features [3], or transforming the features into new ones, that is, adding new features 10]. We checked the two in turn in Section 3 and Section 4. Upon completion of former phases a Machine Learning algorithm can be applied. Some algorithms have been proven to perform better in Text Classification tasks is often used as Support Vector Machines. In the present section a brief description of recent modification of learning algorithms in order to be applied in Text Classification is explained. Most of the methods that are using to examine the performance of a machine learning algorithms in Text Classification are expatiated in next section.

7 a) Tokenization

The process of breaking a stream of text up into tokens that is words, phrases, symbols, or other meaningful elements is called Tokenization where the list of tokens is input to the next processing of text classification.

Generally, tokenization occurs at the word level. Nevertheless, it is not easy to define the meaning of the "word". Where a tokenize process responds on simple heuristics, for instance:

All contiguous strings of alphabetic characters are part of one token; similarly with numbers. Tokens are divided by whitespace characters, like a space or line break, or by punctuation characters. Punctuation and whitespace may or may not be added in the resulting list of tokens. In languages like English (and most programming languages) words are separated by whitespace, this approach is straightforward. Still, tokenization is tough for languages with no word boundaries like Chinese. [1] Simple whitespacedelimited tokenization also shows toughness in word collocations like New York which must be considered as single token. Some ways to mention this problem are by improving more complex heuristics, querying a table of common collocations, or fitting the tokens to a language model that identifies collocations in a next processing.

8 b) Stemming

In linguistic morphology and information collection, stemming is the process for decreasing deviated (or sometimes derived) words to their stem, original form. The stem need not be identical to the morphological root of the

word; it is usually enough if it is concern words map of similar stem, even if this stem is not a valid root. In computer science algorithms for stemming have been studied since 1968. Many search engines consider words with the similar stem as synonyms as a kind of query broadening, a process called conflation.

9 c) Stop word removal

Typically in computing, stop words are filtered out prior to the processing of natural language data (text) which is managed by man but not a machine. A prepared list of stop words do not exist which can be used by every tool. Though any stop word list is used by any tool in order to support the phrase search the list is ignored.

Any group of words can be selected as the stop words for a particular cause. For a few search machines, these is a list of common words, short function words, like the, is, at, which and on that create problems in performing text mining phrases that consist them. Therefore it is needed to eliminate stop words

10 d) Vector representation of the documents

Vector denotation of the documents is an algebraic model for symbolizing text documents (and any objects, in general) as vectors of identifiers, like, for example, index terms which will be utilized in information filtering, information retrieval, indexing and relevancy rankings where its primary use is in the SMART Information Retrieval System.

A sequence of words is called a document [16]. Thus every document is generally denoted by an array of words. The group of all the words of a training group is called vocabulary, or feature set. Thus a document can be produced by a binary vector, assigning the value 1 if the document includes the feature-word or 0 if there is no word in the document.

11 e) Feature Selection and Transformation

The main objective of feature-selection methods is to decrease of the dimensionality of the dataset by eliminating features that are not related for the classification [6]. The transformation procedure is explained for presenting a number of benefits, involving tiny dataset size, tiny computational needs for the text categorization algorithms (especially those that do not scale well with the feature set size) and comfortable shrinking of the search space. The goal is to reduce the curse of dimensionality to yield developed classification perfection. The other advantage of feature selection is its quality to decrease 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 augment generalization.

Feature Transformation differs considerably from Feature Selection approaches, but like them its aim is to decrease the feature set volume [10]. The approach does not weight terms in order to neglect the lower weighted but compacts the vocabulary based on feature concurrencies.

IV.

12 Assortment of Machine learning algorithms for Text Classification

After feature opting and transformation the documents can be flexibly denoted in a form that can be utilized by a ML algorithm. Most of the text classifiers adduced in the literature utilizing machine learning techniques, probabilistic models, etc. They regularly vary in the approach taken are decision trees, na?ve-Bayes, rule induction, neural networks, nearest neighbors, and lately, support vector machines. Though most of the approaches adduced, automated text classification is however a major area of research first due to the effectiveness of present automated text classifiers is not errorless and nevertheless require development.

Naive Bayes is regularly utilized in text classification applications and experiments due to its easy and effectiveness [14]. Nevertheless, its performance is reduced due to it does not model text. Schneider addressed the problems and display that they can be resolved by a few plain corrections [24]. Klopotek and Woch presented results of empirical evaluation of a Bayesian multinet classifier depending on a novel method of learning very large tree-like Bayesian networks [15]. The study advices that tree-like Bayesian networks are able to deal a text classification task in one hundred thousand variables with sufficient speed and accuracy.

When Support vector machines (SVM), are applied to text classification supplying excellent precision, but less recollection. Customizing SVMs means to develop recollect which helps in adjusting the origin associated with an SVM. Shanahan and Roma explained an automatic process for adjusting the thresholds of generic SVM [26] for improved results. Johnson et al. explained a fast decision tree construction algorithm that receives benefits of the sparse text data, and a rule simplification method that translates the decision tree into a logically equivalent rule set [9].

Lim introduced a method which raises performance of kNN based text classification by utilizing calculated parameters [18]. Some variants of the kNN method with various decision functions, k values, and feature sets are also introduced and evaluated to discover enough parameters.

For immediate document classification, Corner classification (CC) network, feed forward neural network is used. A training algorithm, TextCC is introduced in [34]. The complexity of of text classification tasks generally

varies. As the number of different classes augments as of complexity and hence the training set size is required. In multi-class text classification task, unavoidable some classes are a bit harder than others to classify. Reasons for this are: very few positive training examples for the class, and lack of good forecasting features for that class.

When training a binary classifier per category in text categorization, we use all the documents in the training corpus that has the category as related training data and all the documents in the training corpus that are of the other categories are non related training data. It is a regular case that there is an overwhelming number of non related training documents specially when there is high number of categories with every allotted to a tiny documents, which is an "imbalanced data problem". This problem gives a certain risk to classification algorithms, which can accomplish perfection by simply classifying every example as negative. To resolve this problem, cost sensitive learning is required [5].(D D D D) C 2012

13 Year

A scalability analysis of a number of classifiers in text categorization is shown in [32]. Vinciarelli introduces categorization experiments performed over noisy texts [31]. With this noisy that any text got through an extraction process (affected by errors) from media other than digital texts (e.g. transcriptions of speech recordings extracted with a recognition system). The performance of the categorization system over the clean and noisy (Word Error Rate between ~10 and ~50 percent) versions of the similar documents is compared. The noisy texts are got through Handwriting Recognition and simulation of Optical Character Recognition where the results show less performance which is agreeable. Other authors [36] also presented to parallelize and distribute the process of text classification. With such a procedure, the performance of classifiers can be developed in two ways that is accuracy and time complexity.

Of late in the area of Machine Learning the concept of combining classifiers is introduced as a new path for the development of the performance of single classifiers. Numerous methods advised for the creation of ensemble of classifiers. Mechanisms utilized to construct ensemble of classifiers consists of three issues. They are 1) Using various subset of training data with a one learning method, ii) Using various training parameters with a one training method (e. g. using different initial weights for each neural network in an ensemble), iii) Using various learning methods. In the context of combining multiple classifiers for text categorization, a number of researchers said that combination of various classifiers develops classification perfection [1], [29].

Comparison between the best individual classifier and the combined method, it is find that the performance of the combined method is greater [2]. Nardiello et al. [21] also presented algorithms in the family of "boosting"-based learners for automated text classification with good results.

V.

14 Current State of the art

Frunza, O et al [44] applied machine learning based text categorization for disease treatment relations titled "A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts". With the reference of their proposal the authors debated that The Machine Learning (ML) field has won place in almost any domain of research and of lately become a reliable tool in the medical field. The empirical domain of automatic learning is used in tasks like medical decision support, medical imaging, protein-protein interaction, extraction of medical knowledge, and for total patient management care. ML is pursued as a tool by which computer-based systems can be combined with healthcare field in order to get a better, more efficient medical care.

The two tasks that are undertaken in presented model [44] supplied the basis for the design of an information technology framework has capacity to find and separate healthcare information. The first task made to find and extracts informative sentences on diseases and treatments topics, while the second one prepared to perform a finer grained classification of these sentences according to the semantic relations that presents between diseases and treatments.

The task of sentence selection discovers sentences from Medline published abstracts that talk about diseases and treatments. The task is sameto a scan of sentences contained in the abstract of an article in order to present to the user-only sentences that are found as including related information (diseasetreatment information).

The task of relation identification has a deeper semantic dimension and it emphasized on finding disease-treatment relations in the sentences already choosen as being informative (e. g., task 1 is applied first). The training set is utilized to train the ML algorithm and the test set to test its performance.

Separately from the work of Rosario and Hearst [49], introduces [44] the annotations of the data set are utilizes to generate a hard task (task 1). It finds informative sentences that include information about diseases and treatments and semantic relations, versus non informative sentences. This permits to observe the excellence NLP and ML techniques can mingle with the task of discovering informative sentences, or in other words, they can remove out sentences that are related to medical diseases and treatments.

In this present model [44] the authors pointed on a few relations of interest and tried to find how the predictive model and representation technique work out good results. The task of discovering the semantic relations is as follows: Three models are constructed. Every model is focused on one relation and can distinguish sentences that contain the relation from sentences that do not. This setting is similar to a twoclass classification task in which instances are labeled either with the relation in question (Positive label) or with non relevant information

(Negative label);One model is built, to differentiate the three relations in a three-class classification task so that every sentence is named with one of the semantic relations. Utilizing the pipeline of tasks, we avoid some faults that can be proposed because of the truth that is considered uninformative sentences as potential data during classifying sentences directly into semantic relations. It is believed that this is a solution for discovering and separating related information made to a special semantic relation due to the second task is endeavoring to a finer grained classification of the sentences that already include information about the relations of interest.

Observation: Probabilistic models are standard and reliable for tasks performed on short texts in the It is find potential developments in results when more information is brought in the representation technique for the task of classifying short medical texts. The second task that mentioned can be seen as a task that could get advantage from solving the first task first. Also, to perform a triage of the sentences (task 1) for a relation classification task is paramount step. Probabilistic models mixed with a representation technique bring the best results. This work seems to be quite effective text classification using machine learning to extract the relations semantically between the treatments. And it is quite clear that the model is not considering the context and conceptual issues to derive the relations between treatment relations.

For the preparation of text classifiers a new methodology which combines the distribution clustering of words and a learning technique was proposed by Al-Mubaid et al [45]. Al-Mubaid et al [50] opines that task of categorization becomes difficult if the content of the document has high dimensionality. He proposes that, this difficulty of high dimensionality can be resolved by feature clustering which is more effective than the current technique i. E feature selection. Thus the new method utilizes distributional clustering method (IB) to This new model follows a good feature clustering techniques and a learning algorithm Lsquare which is logic based. This approach depends on the methodology where the text is presented by forming different clusters from the input data set and text classifiers are developed by using the Lsquare [51].

Word Features and Feature Clustering: In the vector representation every word in the text corresponds to a feature, henceforth leading to the high dimensionality of the document. By forming the clusters alike words i.e word clustering, high dimensionality of a text is minimized. Distributional clustering of words [52], [53], [54], [55], [56] is said to be the most successful to get the word clustering for TC. Every feature is a cluster alike words. For word feature techniques [53], feature clustering is more effective and useful when compared to the feature selection.

Since big quantity of lexis is brought into a group in the word clusters the necessity for feature selection automatically gets reduced. Since large number of words is brought into a group in the word clusters the necessity for feature selection automatically gets reduced. As lexis of text is brought into a cluster whole information of the text gets carried. Where as in feature selection there is a possibility to miss any information of the text.

15 Distributional Clustering Using the IB Method:

Lexis Clusters formed by the clustering alike words is more efficient and easier when compared to feature selection [56]. In this new proposed model the common structure of Bottleneck a new technique is utilized to form the word clusters [53]. IB method traces the fully developed pertinent coding or the compact version of one variable X, given the joint distribution of two random variables P(X, Y), while the mutual information about the other variable Y is saved to the extent feasible. In the technique used in [53], X denotes the input lexis and variable Y denotes the class labels. In addition, they give a hierarchical top-down clustering process for generating the distributional IB clusters [53]. Initiating with one cluster that consists all the input data, the clusters divides in iterations with incrementing the annealing parameter .

Observation: Recent developments in the techniques of feature clustering and dimension reduction are well utilized in the proposed in new model. The proposed TC approach combines these new advancements with logic-based learning techniques. The proposed method is experimented on all trainingtesting settings utilizing WebKB data set and on 0NG data set. These experiments proved that TC approach is more effective than that of SVM-based system. This technique of machine learning doesn't consider the semantic, theoretical and relative relations of the texts and the new model is tested under the same parameters. This is a disadvantage of the new approach and the feature research will be done in such a way that it recognizes all the semantic, theoretical and relative relations of the texts.

Sun, A. et al [46] opines that classification techniques that are utilizing top-down approach are competent enough to deal with changes to the category trees in text mining. Though these approaches are effective one common problem in all these methods is Blocking. It means rejection of the texts by the classifiers which cannot be sent to the classifiers at lower-levels. Thus Sun, A. et al [57] Extended Multiplicative Method (EMM): The extended multiplicative method is an extension of the multiplicative method projected by Dumais and Chen [58]. The proposed new model will be able to handle category trees with more levels, where as the source method is limited only to the 3 level category trees. Like STTD, EMM links a local classifier with each leaf node and a sub-tree classifier with each non-leaf node. Let n c be a leaf node at level n and the parent node be c n-1 . Observation: The challenge of Blocking in hierarchical text classification is mainly targeted in the proposed new model. Top-down approach is used to resolve the blocking problem. To differentiate the degree of blocking, we have established blocking factor as a new kind of classifier-centric performance measure. As a solution to the blocking challenge three methods were put forward namely, threshold reduction, restricted voting, and extended multiplicative methods. Of all the techniques restricted voting model is effective in bringing down the Blocking

problem and has proved to be the best in terms of F 1 M measure too. But the disadvantage of this technology is it requires more classifiers thus demanding more time for training. Though they are few advantages, all the said models are not effective in summing-up the given document. Furthermore even these new models depend on term and document frequency and are unable to consider the contextual and semantic relations of the text. Thus further research will be focused on developing a model which recognizes semantic, conceptual and contextual relations of the texts thus enabling an effective precision. Text categorization methods that are utilizing machine learning techniques to bring on manuscript classifiers face a problem with very high computational costs that sometimes rise exponentially in the number of features because of the usage of the example manuscripts those can be part of the multiple classes. As a remedy to these raising costs, Sarinnapakorn, K et al [47] proposed a "baseline induction algorithm" which will be exclusively used for sub sets of features, where a set of classifiers are united. Along with the above said solutions Sarinnapakorn, K et al [47] proposed one more technique i. e alternative fusion techniques for the classifies that send back both class labels and confidences in these labels. This technique is developed from the Dempster-Shafer Theory. Every run of BIA stimulates a sub classifier that, for article x and class label l, returns f(x; l) (-,) that measures the sub classifier's confidence in l (higher f(x; l) designates higher confidence). A fusion methodology is required to unite these suggestions and confidence values. The instruction standardizes the function f(x;1) so as to ensure that its values commence in between the range of [0, 1]. If suppose range 1 is the alteration between f(x;1) and the least belief of the classifier of any random label is elucidated, the resulting solution is then partitioned based on the high count obtained in the outputs of the sub-classifier. This is particularly done to ensure that the changed values can be considered as degrees of confidence, where values nearing 1 replicate their confidence in 1 while values nearing 0 replicate their robust incredulity in 1.

16 An article

Step 2 utilizes the changed confidence values in the estimations of the BBAs that are closely related to the class labels. Refer the appendix for valid evidence that masses just estimated fulfill the requirements in (1). The Dempster-Shafer rule of arrangement is to blend the mass values restored by the various sub classifiers for all the four specified opportunities mentioned in every available class label.

Observation: Sarinnapakorn, K et al [47] designated a methodology to tackle forbidden computational charges of text-classification schemes wherein every individual file fits in the multiple classes at that point of time. The designated model specifically deals with the orientation mechanisms, whose training period increases in a linear fashion in accordance with the multiple features that are utilized for depicting critical hurdles in the case of text files. The feature called observation that the sub classifier amalgamation results in typical bursting of specialized computational reduction, exploiting the fact that the performance that was accomplished earlier can be still enhanced. The enhancement may probably occur if the chosen characteristic-selection mechanism utilizes provoked sub classifiers who harmonize amicably. The chosen box was a black one and hence the exact featured option of the BIA was not considered seriously.

Bell, D. A. et al [48] claims those results prove otherwise stating various text differentiation methodologies present various results. He also prescribed a methodology for merging the classifiers. Various techniques like support vector machine (SVM). Nearest fellow neighbors (kNN) and Rocchio were researched upon to unite the effects of two or more various categorization techniques in accordance with a sequential line of attack. A more refined version of the tactic to be employed is explained as follows:

Utilization of various confirmation techniques employs merging mechanisms like Dempster's rule or or the orthogonal sum [14] to resolve the Data Information Knowledge fusion issue. A more conventional way to substantial motive of explanation depends on the concept of statistical methodologies to present indicative assurance ie. The Dempster-Shafer (D-S) hypothesis that utilizes the quantitative data extracted from the classifiers.

Evidence Theory: The D-S hypothesis is an efficient technique realized for surviving the tentative expressions implanted in the confirmatory issues that are precariously used in the reasoning methods and it best ensembles with conclusion-based actions. This hypothesis is often considered as a simplification of Bayesian probability hypothesis by assisting in issuing a rational presentation for lack of evidence as also by abandoning the irrelevant and inadequate reasoning standards. A reasoning technique is devised as bits of evidence and specialize them to a stern formal mechanism so as to draw assumptions from a undisclosed evidence where it is expressed in the form of evidential functions. Few functions that are used frequently are mass functions, belief functions, doubt functions and plausibility functions. All these functions express the same data as the others.

Categorization-Specific Mass Function: The designated model contemplates the issue of calculating degrees of principle for the proof deduced from the text classifiers and the varied exact delineations of mass and belief terms for this specific field and then blend number of pieces of proofs to arrive at a conventional decision. The 2-Points Focused Combination Method: Suppose that there exists a set of training data and a set of algorithms, where every individual algorithm produces one or more classifiers depending on the selected training set of data and then merge various outputs of various classifiers depending on the same testing files using Dempster's rule of merging to prepare the ultimate classification verdict.

Observation: Bell, D. A. et al [48] proposes a unique mechanism for presenting outputs obtained from various classifiers. A focal element triplet can be converted to a focal element quarter by expanding it. A consequential

methodology implemented for a number of classifiers depending on the new structure was scrutinized as also modus operandi used for calculating

17 Conclusion

This paper focuses on investigating the utilization of Machine learning mechanisms for ascertaining text classifiers and tries to generalize the specific properties of the recent trends in learning techniques with text data and recognize whether any of the stipulated models cited recently in current literature are judged as text analogous in terms of semantic, conceptual and contextual format. It is apparent from the statistics obtained that least count of models has been insinuated in recent times, focusing largely on reducing the computational density of the machine learning forms to enhance competence. Concerning recent literature, no recent work has been devised to focus on managing coherency of the files already classified.

18 Global Journal of Computer Science and Technology

Volume XII Issue XI Version I [1]

Figure 1:

[1] © 2012 Global Journals Inc. (US) © 2012 Global Journals Inc. (US)

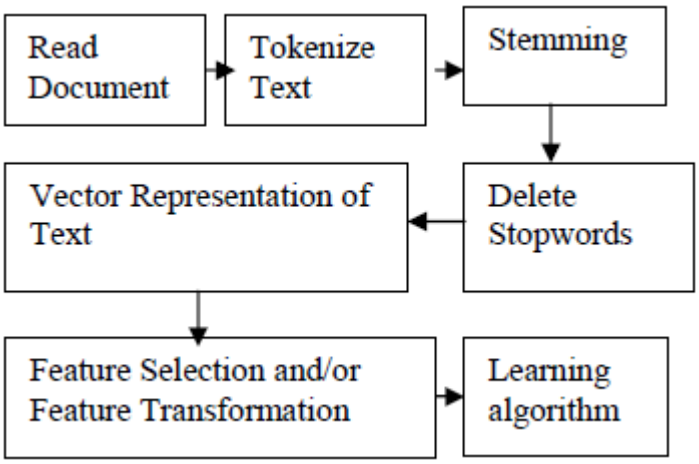

Figure 2:

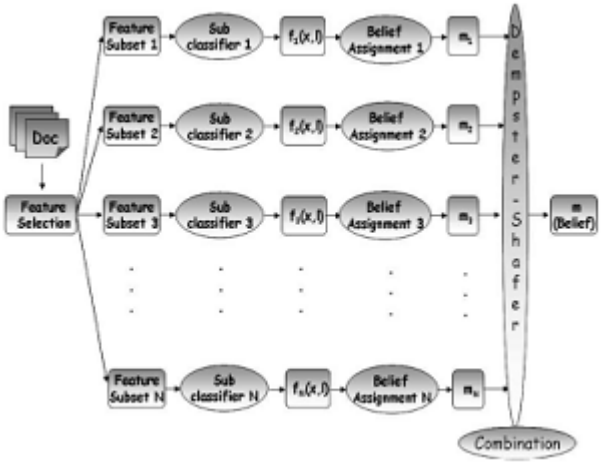

Figure 3:

[European Colloquium on Information Retrieval Research ()] , ECIR-01. *European Colloquium on Information Retrieval Research* 2001.

[Kehagias et al. ()] 'A Comparison of Word-and Sense-Based Text Categorization Using Several Classification Algorithms'. A Kehagias , V Petridis , V Kaburlasos , P Fragkou . *JIIS* 2003. 21 p. .

[Johnson et al. (2002)] 'A decision-tree-based symbolic rule induction system for text categorization'. D E Johnson , F J Oles , T Zhang , T Goetz . *IBM Systems Journal* September 2002.

[Dhillon et al. ()] 'A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification'. I Dhillon , S Mallela , R Kumar . *J. Machine Learning Research* 2003. 3.

[Frunza et al. (2011)] 'A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts'. O Frunza , D Inkpen , T Tran . doi: 10. 1109/TKDE. 2010. 152. `http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5560656&isnumber=5753264` *IEEE Transactions on* June 2011. 23 (6) p. . (Knowledge and Data Engineering)

[Felici and Truemper ()] 'A Minsat Approach for Learning in Logic Domains'. G Felici , K Truemper . *Informs J. Computing* 2002. 14 (1) .

[Al-Mubaid and Umair (2006)] 'A New Text Categorization Technique Using Distributional Clustering and Learning Logic'. H Al-Mubaid , S A Umair . doi: 10. 1109/TKDE. 2006. 135. *IEEE Transactions on* Sept. 2006. 18 (9) p. . (Knowledge and Data Engineering)

[Yang et al. ()] 'A scalability analysis of classifiers in text categorization'. Y Yang , J Zhang , B Kisiel . *ACM SIGIR'03*, 2003. p. .

[Qiang et al. (2005)] 'A Study of Semi-discrete Matrix Decomposition for LSI in Automated Text Categorization'. Wang Qiang , Wang Xiaolong , Guan Yi . Lecture Notes in Computer Science Jan 2005. 3248 p. .

[Qiang et al. (2005)] 'A Study of Semidiscrete Matrix Decomposition for LSI in Automated Text Categorization'. W Qiang , W Xiaolong , G Yi . *LNCS* Jan 2005. 3248 p. .

[Montanes et al. ()] 'A Wrapper Approach with Support Vector Machines for Text Categorization'. E Montanes , J R Quevedo , I Diaz . *LNCS* 2003. 2686 p. .

[Zu et al. ()] 'Accuracy improvement of automatic text classification based on feature transformation'. G Zu , W Ohyama , T Wakabayashi , F Kimura . *Proc: the 2003 ACM Symposium on Document Engineering*, (the 2003 ACM Symposium on Document Engineering) November 20-22, 2003. p. .

[Han et al. (2004)] 'Accuracy Improvement of Automatic Text Classification Based on Feature Transformation and Multi-classifier Combination'. X Han , G Zu , W Ohyama , T Wakabayashi , F Kimura . *LNCS* Jan 2004. 3309 p. .

[Cardoso-Cachopo and Oliveira (2003)] 'An Empirical Comparison of Text Categorization Methods'. Ana Cardoso-Cachopo , Arlindo L Oliveira . Lecture Notes in Computer Science Jan 2003. 2857 p. .

[Yang ()] 'An evaluation of statistical approaches to text categorization'. Y Yang . *Journal of Information Retrieval* 1999. 1 (1/2) p. .

[Forman ()] 'An Experimental Study of Feature Selection Metrics for Text Categorization'. G Forman . *Journal of Machine Learning Research* 3 2003. p. .

[Kessler et al. ()] 'Automatic detection of text genre'. B Kessler , G Nunberg , H Schutze . *Proceedings of the Thirty-Fifth ACL and EACL*, (the Thirty-Fifth ACL and EACL) 1997. p. .

[Borko and Bernick ()] 'Automatic document classification'. H Borko , M Bernick . *J. Assoc. Comput. Mach* 1963. 10 p. .

[Sun et al. (2004)] 'Blocking reduction strategies in hierarchical text classification'. A Sun , E. -P Lim , W. -K Ng , J Srivastava . doi: 10. 1109/TKDE. 2004. 50. *IEEE Transactions on* Oct. 2004. 16 (10) p. . (Knowledge and Data Engineering)

[Bi et al. ()] 'Combining Multiple Classifiers Using Dempster's Rule of Combination for Text Categorization'. Y Bi , D Bell , H Wang , G Guo , K Greer . *MDAI* 2004. p. .

[Bao and Ishii ()] 'Combining Multiple kNN Classifiers for Text Categorization by Reducts'. Y Bao , N Ishii . *LNCS* 2002. 2534 p. .

[Sarinnapakorn and Kubat (2007)] 'Combining Subclassifiers in Text Categorization: A DST-Based Solution and a Case Study'. K Sarinnapakorn , M Kubat . doi: 10. 1109/TKDE. 2007. 190663. *IEEE Transactions on* Dec. 2007. 19 (12) p. . (Knowledge and Data Engineering)

[Nardiello et al. (2003)] 'Discretizing Continuous Attributes in AdaBoost for Text Categorization'. P Nardiello , F Sebastiani , A Sperduti . *LNCS* Jan 2003. 2633 p. .

[Torkkola ()] 'Discriminative Features for Text Document Classification'. K Torkkola . *Proc. International Conference on Pattern Recognition*, (International Conference on Pattern Recognition) 2002.

[Pereira et al. ()] 'Distributional Clustering of English Words'. F Pereira , N Tishby , L Lee . *Proc. 31st Ann. Meeting of the ACL*, (31st Ann. Meeting of the ACL) 1993. p. .

[Baker and Mccallum ()] 'Distributional Clustering of Words for Text Classification'. L D Baker , A K Mccallum . *Proc. Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, (Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval) 1998.

[Bekkerman et al. ()] 'Distribu¬tional Word Clusters vs Words for Text Categorization'. R Bekkerman , R El-Yaniv , N Tishby , Y Winter . *J. Machine Learning Research* 2003. 3.

[Kim et al. ()] *Effective Methods for Improving Naive Bayes Text Classifiers*, S B Kim , H C Rim , D S Yook , H S Lim . LNAI 2417. 2002. p. .

[Zhou and Guan (2002)] 'Evaluation and Construction of Training Corpuses for Text Classification: A Preliminary Study'. Shuigeng Zhou , Jihong Guan . Lecture Notes in Computer Science Jan 2002. 2553 p. .

[Srinivasan and Rindflesch ()] 'Exploring Text Mining from Medline'. P Srinivasan , T Rindflesch . *Proc. Am. Medical Informatics Assoc. (AMIA) Symp* 2002.

[Sousa et al. ()] 'Feature Selection Algorithms to Improve Documents Classification Performance'. P Sousa , J P Pimentao , B R Santos , F Moura-Pires . *LNAI 2663*, 2003. p. .

[Soucy and Mineau ()] *Feature Selection Strategies for Text Categorization*, P Soucy , G Mineau . 2003, LNAI 2671, 2003. AI. p. .

[Novovicova and Malik ()] 'Feature Selection Using Improved Mutual Information for Text Classification'. J Novovicova , A Malik , PudilP . *SSPR&SPR 2004*, 2004. 3138 p. .

[Manning and Sch¨utze ()] *Foundations of Statistical Natural Language Processing*, C Manning , H Sch¨utze . 1999. Cambridge, MA: MIT Press.

[Dumais and Chen (2000)] 'Hierarchical Classification of Web Content'. S T Dumais , H Chen . *Proc. ACM SIGIR '00*, (ACM SIGIR '00) July 2000. p. .

[Sun and Lim (2001)] 'Hierarchical Text Classification and Evaluation'. A Sun , E. -P Lim . *Proc. IEEE Int'l Conf. Data Mining (ICDM '01)*, (IEEE Int'l Conf. Data Mining (ICDM '01)) Nov. 2001. p. .

[Lim (2004)] 'Improving kNN Based Text Classification with Well Estimated Parameters'. Heui Lim . *LNCS* Oct 2004. 3316 p. .

[Shanahan and Roma ()] *Improving SVM Text Classification Performance through Threshold Adjustment, LNAI 2837*, J Shanahan , N Roma . 2003. p. .

[Pazienza ()] 'Information Extraction'. M T Pazienza . Lecture Notes in Computer Science 1997. 1299.

[Fragoudis et al. ()] 'Integrating Feature and Instance Selection for Text Classification'. D Fragoudis , D Meretakis , S Likothanassis . *SIGKDD '02*, (Edmonton, Alberta, Canada) July 23-26, 2002.

[Brank et al. ()] 'Interaction of Feature Selection Methods and Linear Classification Models'. J Brank , M Grobelnik , N Milic-Frayling , D Mladenic . *Proc. of the 19th International Conference on Machine Learning*, (of the 19th International Conference on Machine LearningAustralia) 2002.

[Cho and Lee (2003)] 'Learning Neural Network Ensemble for Practical Text Classification'. Sung-Bae Cho , Jee-Haeng Lee . Lecture Notes in Computer Science Aug 2003. 2690 p. .

[Al-Mubaid and Truemper ()] 'Learning to Find Context-Based Spelling Errors'. H Al-Mubaid , K Truemper . *Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques* 2006.

[Springer ()] 'Little words can make a big difference for text classification'. Heidelberg Springer , GermanyRiloff E . *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval*, (SIGIR-95, 18th ACM International Conference on Research and Development in Information RetrievalSeattle, WA) 1995. 1995. p. .

[Sebastiani ()] 'Machine Learning in Automated Text Categorization'. F Sebastiani . *ACM Computing Surveys* 2002. 34 (1) p. .

[Knight ()] 'Mining online text'. K Knight . *Commun. ACM* 1999. 42 p. .

[Vinciarelli ()] 'Noisy Text Categorization, Pattern Recognition'. A Vinciarelli . *17th International Conference on (ICPR'04)*, 2004. p. .

[Bell et al. (2005)] 'On combining classifier mass functions for text categorization'. D A Bell , J W Guan , Y Bi . doi: 10. 1109/TKDE. 2005. 167. *IEEE Transactions on* Oct. 2005. 17 (10) p. . (Knowledge and Data Engineering)

[Verayuth Lertnattee and Theeramunkong (2004)] 'Parallel Text Categorization for Multi-dimensional Data'. Thanaruk Verayuth Lertnattee , Theeramunkong . Lecture Notes in Computer Science Jan 2004. 3320 p. .

[Madsen et al. ()] 'Pruning the Vocabulary for Better Context Recognition'. R E Madsen , S Sigurdsson , L K Hansen , J Lansen . *7th International Conference on Pattern Recognition*, 2004.

[Guan and Zhou ()] *Pruning Training Corpus to Speedup Text Classification*, J Guan , S Zhou . 2002. p. .

[Lewis et al. ()] 'RCV1: A New Benchmark Collection for Text Categorization Research'. D Lewis , Y Yang , T Rose , F Li . *Journal of Machine Learning Research* 2004. 5 p. .

[Chawla et al. ()] 'SMOTE: Synthetic Minority Over-sampling Technique'. N V Chawla , K W Bowyer , L O Hall , W P Kegelmeyer . *Journal of AI Research* 16 2002. p. .

[Schneider ()] 'Techniques for Improving the Performance of Naive Bayes for Text Classification'. K Schneider . *LNCS* 2005. 3406 p. .

[Ke and Shaoping ()] 'Text categorization based on Concept indexing and principal component analysis'. H Ke , M Shaoping . *Proc. TENCON 2002 Conference on Computers*, (TENCON 2002 Conference on Computers) 2002. p. .

[Leopold and Kindermann ()] 'Text Categorization with Support Vector Machines. How to Represent Texts in Input Space?'. Edda & Leopold , Jörg Kindermann . *Machine Learning* 2002. 46 p. .

[Merkl ()] 'Text classification with selforganizing maps: Some lessons learned'. D Merkl . *Neurocomputing* 1998. 21 (3) p. .

[Slonim and Tishby] 'The Power of Word Clusters for Text Classification'. N Slonim , N Tishby . *Proc. 23rd*, (23rd)

[Klopotek and Woch ()] 'Very Large Bayesian Networks in Text Classification'. M Klopotek , M Woch . *ICCS 2003, LNCS 2657*, 2003. p. .

[Zhang et al. (2005)] Zhenya Zhang , Shuguang Zhang , Enhong Chen , Xufa Wang , Hongmei Cheng . *TextCC: New Feed Forward Neural Network for Classifying Documents Instantly*, Lecture Notes in Computer Science Jan 2005. 3497 p. .