International Journal of Academic Research and Development
ISSN: 2455-4197
Impact Factor: RJIF 5.22
www.academicsjournal.com
Volume 3; Issue 2; March 2018; Page No. 384-389

Machine learning algorithms for text-documents classification: A review

[1] R Manikandan, [2] Dr. R Sivakumar
[1] Research Scholar, PG and Research Department of Computer Science, A.V.V.M. Sri Pushpam College Autonomous, Poondi, Thanjavur, Tamil Nadu, India
[2] Associate Professor & Head, PG and Research Department of Computer Science, A.V.V.M. Sri Pushpam College Autonomous, Poondi, Thanjavur, Tamil Nadu, India

Abstract
The classification and clustering of e-documents, online news, blogs, e-mails and digital libraries need text mining, machine learning and natural language processing techniques to get meaningful knowledge. This paper provides a review of the principles, advantages and applications of document classification, Document clustering and text mining, focusing on the existing literature.

Keywords: text mining, web mining, documents classification, document clustering information retrieval

1. Introduction
Machine learning is used to teach machines how to handle the data more efficiently. The text mining studies are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources. The resources of unstructured and semi structured information include the world wide web, governmental electronic repositories, news articles, biological databases, chat rooms, digital libraries, online forums, electronic mail and blog repositories. Therefore, proper classification, clustering and knowledge discovery from these resources is an important area for research. Natural Language Processing (NLP), Data Mining, and Machine Learning techniques work together to automatically classify and discover patterns from the electronic documents. The main goal of text mining is to enable users to extract information from textual resources and deals with the operations like, retrieval, classification and clustering (supervised, unsupervised and semi supervised). However how these documented can be properly annotated, presented, classified and clustered. So it consists of several challenges, like proper annotation to the documents, appropriate document representation, dimensionality reduction to handle algorithmic issues [1], and an appropriate classifier function to obtain good generalization and avoid over-fitting. Extraction, Integration and classification of electronic documents from different sources and knowledge discovery from these documents are important for the research communities.
One of the purposes of research is to review the available and known work, so an attempt is made to collect what's known about the documents classification and representation.
This paper covers the overview of syntactic and semantic matters, domain ontology, tokenization concern and focused on the different machine learning techniques for text classification using the existing literature. The motivated perspective of the related research areas of text mining are:

2. Machine Learning Techniques
The documents can be classified by three ways, unsupervised, supervised and semi supervised methods. Many techniques and algorithms are proposed recently for the classification of electronic documents. Machine Learning algorithms are classified as

2.1 Supervised Machine Learning Algorithms
Machine learning algorithms that make predictions on given set of samples. Supervised machine learning algorithm searches for patterns within the value labels assigned to data points.

2.2 Unsupervised Machine Learning Algorithms
There are no labels associated with data points. These machine learning algorithms organize the data into a group of clusters to describe its structure and make complex data look simple and organized for analysis.

2.3 Reinforcement Machine Learning Algorithms
These algorithms choose an action, based on each data point and later learn how good the decision was. Over time, the algorithm changes its strategy to learn better and achieve the best reward.

Fig 1: Machine learning algorithm classification

This section focused on the supervised classification techniques, new developments and highlighted some of the opportunities and challenges using the existing literature. The automatic classification of documents into predefined

categories has observed as an active attention, as the internet usage rate has quickly enlarged. From last few years, the task of automatic text classification have been extensively studied and rapid progress seems in this area, including the machine learning approaches such as Naïve Bayes Classifier Algorithm, Support Vector Machine Learning Algorithm, Decision Tree Machine Learning Algorithm, Rocchio's Algorithm, K-Nearest Neighbor (K-NN), Decision Rules Classification, Artificial Neural Network, Fuzzy correlation and Genetic Algorithm

Normally supervised learning techniques are used for automatic text classification, where pre-defined category labels are assigned to documents based on the likelihood suggested by a training set of labeled documents. Some of these techniques are described below.

i) Naïve Bayes Classifier Algorithm

It would be difficult and practically impossible to classify a web page, a document, an email or any other lengthy text notes manually. This is where Naïve Bayes Classifier machine learning algorithm comes to the rescue. A classifier is a function that allocates a population's element value from one of the available categories. It is the most popular learning method grouped by similarities that works on the popular Bayes Theorem of Probability. To build machine learning models particularly for document classification. It is a simple classification of words based on Bayes Probability Theorem for subjective analysis of content.

Applications of Naïve Bayes Classifier: Sentiment Analysis, Document Categorization, Classifying news articles about Technology, Entertainment, Sports, Politics, etc. and Email Spam Filtering.

Advantages: Performs well when the input variables are categorical, Converges faster, requiring relatively little training data than other discriminative models like logistic regression, when the Naïve Bayes conditional independence assumption holds, It is easier to predict class of the test data set. A good bet for multi class predictions as well and it requires conditional independence assumption

ii) Support Vector Machine Learning Algorithm

Support Vector Machine is a supervised machine learning algorithm for classification or regression problems where the dataset teaches SVM about the classes so that SVM can classify any new data. It works by classifying the data into different classes by finding a line (hyperplane) which separates the training data set into classes.

As there are many such linear hyperplanes, SVM algorithm tries to maximize the distance between the various classes that are involved and this is referred as margin maximization. If the line that maximizes the distance between the classes is identified, the probability to generalize well to unseen data is increased.

SVM's are classified into two categories: Linear SVM's-In linear SVM's the training data, classifiers are separated by a hyperplane and Non-Linear SVM's, it is not possible to separate the training data using a hyperplane.

Advantages of Using SVM: Best classification performance (accuracy) on the training data, Renders more efficiency for correct classification of the future data, It does not make any strong assumptions on data and it does not over-fit the data.

Applications of Support Vector Machine: Stock market forecasting by various financial institutions.

iii) Decision Tree Machine Learning Algorithm

A decision tree is a graphical representation that makes use of branching methodology to exemplify all possible outcomes of a decision, based on certain conditions. In a decision tree, the internal node represents a test on the attribute, each branch of the tree represents the outcome of the test and the leaf node represents a particular class label i.e. the decision made after computing all of the attributes. The classification rules are represented through the path from root to the leaf node.

Types of Decision Trees: Classification Trees and Regression Trees, Decision trees can also be classified as Continuous Variable Decision Trees and Binary Variable Decision Trees.

Advantages: Decision trees are very instinctual and can be explained to anyone with ease. People from a non-technical background, can also decipher the hypothesis drawn from a decision tree, as they are self-explanatory. Data type is not a constraint as they can handle both categorical and numerical variables. Do not require making any assumption on the linearity in the data and hence can be used in circumstances where the parameters are non-linearly related. These machine learning algorithms do not make any assumptions on the classifier structure and space distribution.

Decision trees implicitly perform feature selection which is very important in predictive analytics. When a decision tree is fit to a training dataset, the nodes at the top on which the decision tree is split, are considered as important variables within a given dataset and feature selection is completed by default; it helps to save data preparation time, as they are not sensitive to missing values and outliers.

Drawbacks: The more the number of decisions in a tree, less is the accuracy of any expected outcome; the outcomes may be based on expectations. It do not fit well for continuous variables and result in instability and classification plateaus; it easy to use when compared to other decision making models but creating large decision trees that contain several branches is a complex and time consuming task; it consider only one attribute at a time and might not be best suited for actual data in the decision space; Large sized decision trees with multiple branches are not comprehensible and pose several presentation difficulties.

Applications: Finance for option pricing; Remote sensing is an application area for pattern recognition based on decision trees; Banks to classify loan applicants by their probability of defaulting payments; Gerber Products, a popular baby product company, used decision tree machine learning algorithm to decide whether they should continue using the plastic PVC (Poly Vinyl Chloride) in their products; Rush University

Medical Centre has developed a tool named Guardian that uses a decision tree machine learning algorithm to identify at-risk patients and disease trends.

iv) Rocchio's Algorithm

Rocchio's Algorithm [2] is a vector space method for document routing or filtering in informational retrieval, build prototype vector for each class using a training set of documents, i.e. the average vector over all training document vectors that belong to class and calculate similarity between test document and each of prototype vectors, which assign test document to the class with maximum similarity.

When given a category, the vector of documents belonging to this category is given a positive weight, and the vectors of remaining documents are given negative weight. The positively and negatively weighted vectors, the prototype vector of this category is obtained.

This algorithm [3] is easy to implement, efficient in computation, fast learner and have relevance feedback mechanism but low classification accuracy. Linear combination is too simple for classification and constant α and β are empirical. This is a widely used relevance feedback algorithm that operates in the vector space model [4].

The researchers have used a variation of Rocchio's algorithm in a machine learning context, i.e., for learning a user profile from unstructured text [5, 6], the goal in these applications is to automatically induce a text classifier that can distinguish between classes of documents.

v) K-Nearest Neighbor (K-NN)

The k-nearest neighbor algorithm (k-NN) [7] is used to test the degree of similarity between documents and k training data and to store a certain amount of classification data, thereby determining the category of test documents.

This method is an instant-based learning algorithm that categorized objects based on closest feature space in the training set [8]. The training sets are mapped into multi-dimensional feature space. The feature space is partitioned into regions based on the category of the training set. A point in the feature space is assigned to a particular category if it is the most frequent category among the k nearest training data. Usually Euclidean Distance is typically used in computing the distance between the vectors.

The key element of this method is the availability of a similarity measure for identifying neighbors of a particular document [9]. The training phase consists only of storing the feature vectors and categories of the training set. In the classification phase, distances from the new vector, representing an input document, to all stored vectors are computed and k closest samples are selected. The annotated category of a document is predicted based on the nearest point which has been assigned to a particular category.

Calculate similarity between test document and each neighbour, and assign test document to the class which contains most of the neighbors. This method is effective, non parametric and easy to implement.

As compare to Rocchio algorithm more local characteristics of documents are considered, however the classification time is long and difficult to find optimal value of k. i.e., to analyze the k-NN and the Rocchio algorithm, some shortcomings of each are identified in [9]. A new algorithm is proposed in [10] which incorporating the relationship of concept-based thesauri into document categorization using a k-NN classifier, while [11] presents the use of phrases as basic features in the email classification problem and performed extensive empirical evaluation using large email collections and tested with three text classification algorithms, namely, a naïve Bayes classifier and two k-NN classifiers using TF- IDF weighting and resemblance respectively. The k-nearest neighbor classification method is outstanding with its simplicity and is widely used techniques for text classification. This method performs well even in handling the classification tasks with multi-categorized documents.

The major drawback: It uses all features in distance computation, and causes the method computationally intensive, especially when the size of training set grows. Besides, the accuracy of k-nearest neighbor classification is severely degraded by the presence of noisy or irrelevant features.

vi) Decision Rules Classification

Decision rules classification method uses the rule-based inference to classify documents to their annotated categories [12, 13]. The algorithms construct a rule set that describe the profile for each category. Rules are typically constructed in the format of "IF condition THEN conclusion", where the condition portion is filled by features of the category, and the conclusion portion is represented with the category's name or another rule to be tested. The rule set for a particular category is then constructed by combining every separate rule from the same category with logical operator, typically use "and" and "or". During the classification tasks, not necessarily every rule in the rule set needs to be satisfied. In the case of handling a dataset with large number of features for each category, heuristics implementation is recommended to reduce the size of rules set without affecting the performance of the classification. The [14] presents a hybrid method of rule based processing and back-propagation neural networks for spam filtering, Instead of using keywords, this study utilize the spamming behaviours as features for describing emails.

The main advantage: of the implementation of decision rules method for classification tasks is the construction of local dictionary for each individual category during the feature extraction phase [12]. Local dictionaries are able to distinguish the meaning of a particular word for different categories. However, the drawback of the decision rule method is the impossibility to assign a document to a category exclusively due to the rules from different rule sets is applicable to each other. Besides, the learning and updating of decision rule methods need extensive involvement of human experts to construct or update the rule sets. Like the decision trees classification method, the decision rules method does not work well when the number of distinguishing features is large.

vii) Artificial Neural Network

Artificial neural networks are constructed from a large number of elements with an input fan order of magnitudes larger than in computational elements of traditional architectures [15, 16].

These elements, namely artificial neuron are interconnected into group using a mathematical model for information processing based on a connectionist approach to computation. The neural networks make their neuron sensitive to store item. It can be used for distortion tolerant storing of a large number of cases represented by high dimensional vectors. Different types of neural network approaches have been implemented to document classification tasks. Some of the researches use the single-layer perceptron, which contains only an input layer and an output layer due to its simplicity of implementing [17]. Inputs are fed directly to the outputs via a series of weights. In this way it can be considered the simplest kind of feed-forward network.

The multi-layer perceptron which is more sophisticated, which consists of an input layer, one or more hidden layers, and an output layer in its structure, also widely implemented for classification tasks [15]. The main advantage of the implementation of artificial neural network in classification tasks is the ability in handling documents with high-dimensional features, and documents with noisy and contradictory data. Furthermore, linear speed up in the matching process with respect of the large number of computational elements is provided by a computing architecture which is inherently parallel, where each element can compare its input value against the value of stored cases independently from others [16].

The drawback of the artificial neural networks is their high computing cost which consumes high CPU and physical memory usage. Another disadvantage is that the artificial neural networks are extremely difficult to understand for average users. This may negatively influence the acceptance of these methods.

In recent years, neural network has been applied in document classification systems to improve efficiency. Text categorization models using back-propagation neural network (BPNN) and modified back-propagation neuralnetwork (MBPNN) are proposed in [18] for documents classification. An efficient feature selection method is used to reduce the dimensionality as well as improve the performance. New Neural network based document classification method [19], was presented, which is helpful for companies to manage patent documents more effectively.

The ANN can get Inputs xi arrives through pre-synaptic connections, Synaptic efficacy is modelled using real weights wi and the response of the neuron is a nonlinear function f of its weighted inputs. The output from neuron j for pattern p is Opj where

Neural network for document classification produce good results in complex domains and suitable for both discrete and continuous data (especially better for the continuous domain). Testing is very fast however training is relatively slow and learned results are difficult for users to interpret than learned rules (comparing with Decision tree), Empirical Risk Minimization (ERM) makes ANN try to minimize training error, may lead to overfitting.

viii) Fuzzy correlation

Fuzzy correlation can deal with fuzzy information or incomplete data, and also convert the property value into fuzzy sets for multiple document classification [20]. In [21] the authors explores the challenges of multi-class text categorization using one-against-one fuzzy support vector machine with Reuter's news as the example data, and shows better results using one-against-one fuzzy support vector machine as a new technique when compare with one-against-one support vector machine [3] presented the improvement of decision rule and design a new algorithm of f-k-NN (fuzzy k-NN) to improve categorization performance when the class distribution is uneven, and show that the new method is more effective. So the researchers shows great interest recently to use the fuzzy rules and sets to improve the classification accuracy, by incorporating the fuzzy correlation or fuzzy logic with the machine learning algorithm and the feature selection methods to improve the classification process.

ix) Genetic Algorithm

Genetic algorithm [22] aims to find optimum characteristic parameters using the mechanisms of genetic evolution and survival of the fittest in natural selection. Genetic algorithms make it possible to remove misleading judgments in the algorithms and improve the accuracy of document classification. This is an adaptive probability global optimization algorithm, which simulated in a natural environment of biological and genetic evolution, and is widely used for their simplicity and strength. Now several researchers used this method for the improvement of the text classification process. In authors in [23] introduced the genetic algorithm to text categorization and used to build and optimize the user template, and also introduced simulated annealing to improve the shortcomings of genetic algorithm. In the experimental analysis, they show that the improved method is feasible and effective for text classification.

3. Discussion and Conclusion

This paper provides a review of machine learning approaches and documents representation techniques. Analyses of classification algorithms were presented. Several algorithms or combination of algorithms as hybrid approaches was proposed for the automatic classification of documents, among these algorithms, SVM, NB and kNN classifiers are shown most appropriate in the existing literature.

Concept base or semantically representation of documents requires more research. Better classification will be performed when consider the semantic under considerations, semantically and ontology base documents representation opportunities were discussed. With the addition of the ontology and semantic to represent the documents will be more improve accuracy and the classification process. So the identification of features that capture semantic content is one of the important areas for research. The general multiple learning issues in the presence of noise is a tremendously challenging problem that is just now being formulated and will likely require more work in order to successfully develop strategies to find the underlying nature of the manifold.

Several algorithms or combination of algorithms as hybrid approaches were proposed for the automatics classification of documents. Among these algorithms, SVM, NB, kNN and their hybrid system with the combination of different other algorithms are shown most appropriate in the existing literature. However the NB is perform well in spam filtering

and email categorization, requires a small amount of training data to estimate the parameters necessary for classification. Naive Bayes works well on numeric and textual data, easy to implement comparing with other algorithms, however conditional independence assumption is violated by real-world data and perform very poorly when features are highly correlated and does not consider frequency of word occurrences. SVM classifier has been recognized as one of the most effective text classification method in the comparisons of supervised machine learning algorithms [24]. SVM capture the inherent characteristics of the data better and embedding the Structural Risk Minimization (SRM) principle which minimizes the upper bound on the generalization error (better than the Empirical Risk Minimization principle) also ability to learn can be independent of the dimensionality of the feature space and global minima vs. local minima, however, the SVM has been found some difficulties in parameter tuning and kernel selection. If a suitable pre-processing is used with k-NN, then this algorithm continues to achieve very good results and scales up well with the number of documents, which is not the case for SVM [25, 26]. As for naive Bayes, it also achieved good performance with suitable preprocessing.

k-NN algorithm performed well as more local characteristic of documents are considered, however the classification time is long and difficult to find optimal value of k.

More works are required for the performance improvement and accuracy of the documents classification process. New methods and solutions are required for useful knowledge from the increasing volume of electronics documents. The following are the some of opportunities of the unstructured data classification and knowledge discovery.

To reduce the training and testing time of classifier and improve the classification accuracy, precision and recall. The use of semantics and ontology for the documents classification and informational retrieval. Mining trend, i.e. marketing, business, and financial trend (stock exchange trend) form e-documents (Online news, stories, views and events).

Automatic classification and analysis of sentiment, views and extraction knowledge from it. The sentiments and opinion mining is the new active area of text mining. Classification and clustering of semi-structured documents have some challenges and new opportunities.

To identify or match semantically similar data from the web is an important problem with many practical applications. So web information, integration and schema matching needs more exploration.

4. Reference

1. Dasgupta. Feature selection methods for text classification. In Proceedings of the 13th ACMSIGKDD international conference on Knowledge discovery and data mining, 2007, 230 -239.
2. Rocchio J. Relevance Feedback in Information Retrieval", In G. Salton (ed.). The SMART System, 67-88.
3. Willian W. Cohen and Yoram Singer, "Context-sensitive learning method for text categorization", SIGIR' 96, 19th International Conference on Research and Develeoement in Informational Retrieval, 1996, 307-315.
4. Ittner D, Lewis D, Ahn D. Text Categorization of Low Quality Images", In: Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, 1995, 301-315.
5. Balabanovic M, Shoham Y. FAB Content-based, Collaborative Recommendation", Communications of the Association for Computing Machinery.1997; 40(3):66-72, 1997.
6. Pazzani M, Billsus D. Learning and Revising User Profiles", The Identification of Interesting Web Sites. Machine Learning. 1997; 27(3):313-331.
7. Tam V, Santoso A, Setiono R, A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization", Proceedings of the 16th International Conference on Pattern Recognition,2002, 235-238.
8. Eui-Hong (Sam) Han, George Karypis, Vipin Kumar. "Text Categorization Using Weighted Adjusted k-Nearest Neighbor Classification", Department of Computer Science and Engineering. Army HPC Research Centre, University of Minnesota, Minneapolis, USA. 1999.
9. Duoqian Miao, Qiguo Duan, Hongyun Zhang, Na Jiao. Rough set based hybrid algorithm for text classification", Expert Systems with Applications, 2009.
10. Bang SL, Yang JD, Yang HJ. Hierarchical document categorization with k-NN and concept-based thesauri. Information Processing and Management, 2206, 397-406.
11. Matthew Changa, Chung Keung Poon_, "Using Phrases as Features in Email Classification", The Journal of Systems and Software,doi, 2009, 10.1016/j.jss,
12. Chidanand Apte, Fred Damerau, Sholom M. Weiss. Towards Language Independent Automated Learning of Text Categorization Models", In Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1994, 23-30.
13. Chidanand Apte, Fred Damerau, Sholom M. Weiss; "Automated Learning of Decision Rules for Text Categorization", ACM Transactions on Information Systems (TOIS). 1994; 12(3):233-251.
14. Chih-Hung Wu. Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks", Expert Systems with Applications, 2009, 4321-4330.
15. Miguel E. Ruiz, Padmini Srinivasan. Automatic Text Categorization Using Neural Network",In Proceedings of the 8th ASIS SIG/CR Workshop on Classification Research,1998, 59-72.
16. Petri Myllymaki, Henry Tirri. Bayesian Case-Based Reasoning with Neural Network", In Proceeding of the IEEE International Conference on Neural Network. 1993; 93(1):422-427.
17. Hwee-Tou Ng, Wei-Boon Goh, Kok-Leong Low. Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization, In Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. 1997, 67-73.
18. Bo Yu, Zong-ben Xu, Cheng-hua Li. Latent semantic analysis for text categorization using neural network", Knowledge-Based Systems 2008, 900-904.

19. Trappey A J C, Hsu FC, Trappey CV, Lin CI. Development of a patent document classification and search platform using a back-propagation network", Expert Systems with Applications, 2006, 755-765.
20. Que HE. Applications of fuzzy correlation on multiple document classification.Unpublished master thesis", Information Engineering epartment, Tamkang University, Taipei, Taiwan, 2000.
21. Tai-Yue, Wang, Huei-Min Chiang One-Against-One Fuzzy Support Vector Machine Classifier: An Approach to Text Categorization", Expert Systems with Applications, doi: 10.1016/j.eswa, 2009.
22. Wang Xiaoping, Li-Ming Cao. Genetic Algorithm Theory, Application and Software[M].XI'AN:Xi'an Jiaotong University Press, 2002.
23. ZHU Zhen-fang, LIU Pei-yu, Lu Ran, "Research of text classification technology based on genetic annealing algorithm" IEEE, 978-0-7695-3311-7/08, 2008.
24. Yang Y, Liu X. An re-examination of text categorization", Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, 1999. 42-49.
25. Pingpeng Yuan, Yuqin Chen, Hai Jin, Li Huang "MSVM-kNN: Combining SVM and k-NN for Multi-Class Text Classification"978-0-7695-3316-2/08,2008,IEEEDOI 10.1109/WSCS.2008.
26. Fabrice Colas and Pavel Brazdil, "Comparison of svm and some older classification algorithms in text classi_cation tasks", Artificial Intelligence in Theory and Practice, 2006, 169-178.