Survey Paper On Classifiers To Classify The Document

Suruthi N[1], Kavitha R[2], Vinitha K[3]
[1, 3] Dept of Computer Science and Engineering
[2]Associate professor, Dept of Computer Science and Engineering
[1, 2, 3] Parisutham Institute of Technology and Science, Thanjavur, India

Abstract- *Nowadays the machines are used to create data massively. The online documents are continuously growing. With high accessibility of information from various sources, classification task have attained to provide vital importance to the documents. Automated Text document classification is the essential method to manage and classify a massive amount of documents. This paper provides the depth of the information into the text classification process and it's phases and various classifiers. It is also motivated to comparing and complementaring various presented classifiers on the basis of few criteria like time complexity and performance.*

Keywords- Data Mining, Classifier, Text classification, Machine Learning.

I. INTRODUCTION

Data Mining is defined as the procedure of extracting information from huge sets of data. In other words, we can say that data mining is mining knowledge from data. The rapid growth of online documents in the World Wide Web has raised an urgent demand for efficient and effective classification algorithms to help people achieve fast navigation and browsing of online documents [1]. There are plenty of algorithms are available to do the classification process using data mining. 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. Evolution of information through internet and its swift access, on the negative side has embarked various data security and ethical issues. Plagiarism is defined as the act of using another person's words or ideas without giving credit to that person‖[15]. The documents are classified based on their contents and a substantial portion of the information on these digital documents is stored as text. The word is the smallest constituents of text and play a vital role in an Automatic Text Document Classification (ATDC) process [6]. There are six classification methods, which is Rocchio, k-Nearest Neighbors (k-NN), Regression Model, Naïve Bayes and Bayesian nets, Decision Trees and Decision Rules [5]. Manifold matching works to identify embeddings of multiple disparate data spaces into the same low dimensional space [10]. Feature selection is the method of how to select the best subset of the document occurring in data core for using it in purposes of data mining or applications [4]. The task of text/non-text classification in online handwritten documents is to classify handwritten strokes into two categories: text and non-text, where a stroke is a time sequence of pen-tip or finger- tip points recorded from pen-down to pen-up [3]. With increasing growth of the Internet and information technologies, the massive volume of electronic text documents are given through web pages, the news feeds, electronic emails and digital libraries. To handle such massive information, text categorization has become a key technology to discover and classify text documents [7]. The VSM challenges such as high dimensions and sparsity pose to carefully consider the textual features. In general, there are two methodologies for handling features; features selection and feature extraction[8]. With a good document clustering method, computers can automatically organize a document corpus into several hierarchies of semantic clusters. As a result, users can browse and navigate documents efficiently [11]. Text classification consists in the algorithmic assignation of text documents to predefined classes. It has multiple applications, such as sentiment analysis and spam filtering [12]. A rapid increase in digital documents due to heavy use of Internet technologies and electronic devices necessitates efficient techniques to efficiently classify the digital documents. The documents are classified based on their contents and a substantial portion of the information on these digital documents is stored as text. The word is the smallest constituents of text and play a vital role in an Automatic Text Document Classification (ATDC) process. For efficiently classifying the document they used variable Global Feature Selection Scheme[6]. The feature selection techniques broadly fall into three categories: filters, wrappers, and embedded. Cross-Language Concept Matching (CLCM) are used to convert concept-based representations of documents from one language to another using Wikipedia correspondences between concepts in different languages and thus not relying on automated full-text translations. Because the state of art model sometimes leads to misclassification[12].

II. CLASSIFICATION PROCESS

2.1 Document Collection

Text classification starts with collecting the data from various types of documents including different formats. Document collection means collecting the Documents from various fields. Previously the keyword matching techniques were used to retrieve the related information. Sometimes it leads to some misclassification. For reducing this type of complexity lot of algorithms and techniques were used. Using this techniques we can semantically match the document anywhere at anytime. Text mining is used to explore the interesting information from the structured data and nor information from unstructured text data. Text mining is the important field which drives on data mining, machine learning, information retrieval, computational linguistics and statistics. Important text mining processes are exploring information from the database, retrieve the information from the database, natural language processing, classifying the text, analyzing the content and clustering the documents. All these processes are required to execute the processing steps before doing their particular task. Preprocessing steps significantly reduces the size of the input text documents and some actions may involve. Those actions are sentence boundary determination, natural language specific stop word elimination, tokenization and stemming.

2.2 Tokenization

A document is considered as a string, and then partitioned into a list of tokens. Stop words such as “the”, “a”, “and”, etc. are frequently occurring; therefore the insignificant words need to be removed. Tokenization is the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. Tokens can be individual words, phrases or even whole sentences. In the process of tokenization, some characters like punctuation marks are discarded. The tokens become the input for another process like parsing and text mining. Tokenization is used in computer science, where it plays a large part in the process of lexical analysis. Tokens or words are separated by whitespace, punctuation marks or line breaks. White space or punctuation marks may or may not be included depending on the need. All characters within contiguous strings are part of the token. Tokens can be made up of all alpha characters, alphanumeric characters or numeric characters only.

2.3 Feature Extraction

Feature extraction is the process of selecting a subset of the terms occurring in the training set and using only this subset as features in text classification. Feature extraction serves two main purposes. First, it makes training and applying a classifier more efficient by decreasing the size of the effective vocabulary. Second, feature selection often increases classification accuracy by eliminating noise features. A noise feature is one that, when included in the document representation, increases the classification error on new data. Additional features can be mined from the classifiable text; however nature of such features should be highly dependent on the nature of classification to be carried out. If web sites need to be separated into spam and non spam websites, then the word frequency distribution or the ontology is of little use for the classification, because of widespread tactics by the spammers to copy and paste mixture of texts from legitimate web sites in creation of their spam web sites [4]. A feature extraction, feature selection method is used to achieve the classification by selecting appropriate set of features as the input to machine learning based classifiers.

2.4 Natural Language Processing

Feature extraction and reduction phases of text classification process are performed with the help of Natural Language Processing techniques. Linguistic features can be extracted from texts and used as part of their feature vectors [3]. For example parts of the text that are written in direct speech, use of different types of declinations, length of sentences, proportions of different parts of speech in sentences (such as noun phrases, preposition phrases or verb phrases) can all be detected and used as a feature vector or in addition to word frequency feature vector [4]. Many different classes of machine learning algorithms have been applied to natural language processing tasks. These algorithms take as input a large set of "features" that are generated from the input data. Some of the earliest-used algorithms, such as decision trees, produced systems of hard if-then rules similar to the systems of hand-written rules that were then common. Increasingly, however, research has focused on statistical models, which make soft, probabilistic decisions based on attaching real-valued weights to each input feature. Such models have the advantage that they can express the relative certainty of many different possible answers rather than only one, producing more reliable results when such a model is included as a component of a larger system.

2.5 Feature Reduction

Feature reduction a.k.a. Dimensionality reduction is about transforming data of very high dimensionality into data of much lower dimensionality such that each of the lower dimensions manifest much more information. The computational complexity of any operations with such feature vectors will be proportional to the size of the feature vector (Yang & Pedersen, 1997), so any methods that reduce the size of the feature vector while not significantly impacting the classification performance are very welcome in any practical

application. Additionally, it has been shown that some specific words in specific languages only add noise to the data and removing them from the feature vector actually improves classification performance. The set of feature reduction operations involves a combination of three general approaches [5]: 1. Stop words; 2. Stemming; 3. Statistical filtering.

2.6 Feature selection

Feature selection is the process that leads to the reduction of dimensionality of the original data set. The selection term set should contain enough or more reliable information about the original data set. To this end, many criteria are used. For apply the feature selection there are two ways to select it. The first is forward selection starts with no terms and adds them one by one, at each one adding the one that reductions the mistakes. The second is the backward selection that starts with all the terms and eliminates them one by one. Hence, eliminate the one that reductions the most error, in hopes no further elimination up to the error.

2.7 Classification

With each passing day, automatic classification of documents in predefined categories is gaining active attention of many researchers. Supervised, unsupervised and semi supervised are the methods used to classify documents. The last decade has seen the unprecedented and rapid progress in this area, including the machine learning approaches such as Bayesian classifier, Decision Tree, Knearest neighbor(KNN), Support Vector Machines(SVMs), Neural Networks, Rocchio's.

III. CLASSIFIERS

3.1 K-Nearest Neighbour

K nearest neighbors is an elegant supervised machine learning algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).K-NN works on a principle that the points (documents) which are close in the space belong to the same class. The algorithm assimilates all training samples and predicts the response for a new sample by analyzing a certain number (K) of the nearest neighbors of the sample by using some similarity measure such as Euclidean distance measure etc. A major demerit of the similarity measure used in k-NN is that it uses all features in computing distances which degrades its performance. In myriad document data sets, only smaller number of the total vocabulary may be useful in categorizing documents. A probable approach to tackle this problem is to learn weights for different features (or words in document data etc.) [11]. Proposed Weight Adjusted k-Nearest Neighbor (WAKNN) classification algorithm is based on the k-NN classification paradigm which can enhance the performance of text classification [12].

3.2 Support Vector Machine

Initially, Support vector machines (SVM) was developed for building an optimal binary (2-class) classifier but thereafter the technique was extended to regression and clustering problems. The working principle of SVM is to find out a hyper plane (linear/non-linear) which maximizes the margin. SVM is a partial case of kernel-based methods. It binds feature vectors into a higher-dimensional space using a kernel function and builds an optimal linear discriminating function in this space or an optimal hyper-plane that is congruent with the training data [15]. The kernel is not explicitly defined in case of SVM. Instead, a distance between any 2 points in the hyper-space needs to be defined. The key features of SVMs are the use of kernels, the absence of local minima, the sparseness of the solution and the capacity control obtained by optimizing the margin. Besides the advantages of SVMs - from a practical point of view they have some drawbacks. An important practical question that is not entirely solved, is the selection of the kernel function parameters - for Gaussian kernels the width parameter [sigma] - and the value of [epsilon] in the [epsilon]-insensitive loss function. Support Vector Machine is a supervised learning method used for classification. There are three papers that have used Support Vector Machine as their method to predict students performance. A powerful Support Vector Machine (SVM) which was first proposed by Vapnik and it has a great potency of interest in the machine learning research community. Several past studies have reported that the SVM generally has a proficient of delivering the high accuracy in classification when compared to other data classification algorithms. Though, for certain datasets, the achievement of SVM is very subtle in determining the cost parameter and kernel parameters. As in the case of closure, to figure out the most encouraging condition environment the user normally needs to conduct extensive cross validation. Basically this technique is baited to as a model selection. A superior asset of this SVM technique is that, concurrent miniaturize the projected classification error and make best use of the geometric margin, So SVM is also named as paramount Margin Classifiers. It is found on the Structural Risk Minimization (SRM), SVM can be used for both classification and prediction. There are several advantages of SVM such as it uses maximum marginal hyper plane for classifying linearly separable data, Data can be separated clearly into rations, extends by itself in order to classify the linearly inseparable data. In machine learning, support vector machines (SVMs, also support vector

networks[1]) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. When data are unlabelled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. The support vector clustering[2]algorithm created by Hava Siegelmann and Vladimir Vapnik, applies the statistics of support vectors, developed in the support vector machines algorithm, to categorize unlabeled data, and is one of the most widely used clustering algorithms in industrial applications

3.3 Naïve Bayes

The Naive Bayes classifier is a probabilistic classifier based on Bayes theorem with strong and naïve independence assumptions[5]. It is supposed to be one of the most basic text classification techniques with various applications in email spam detection, personal email sorting, document categorization, sexually explicit content detection, language detection and sentiment detection. Experiments witness that this algorithm performs well on numeric and textual data. Though it is often outperformed by other techniques such as boosted trees, random forests, Max Entropy, Support Vector Machines etc., Naive Bayes classifier is quite efficient since it is less computationally intensive (in both CPU and memory) and it necessitates a small amount of training data. The assumption of conditional independence is breached by real-world data with highly correlated features thereby degrading its performance. The advantages of Naïve Bayes classification methods are easy to understand and implement, computationally short time in training process and noise resistance. Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong independence assumptions between the features. naive Bayes models are known under a variety of names, including simple Bayes and independence Bayes. All these names reference the use of Bayes' theorem in the classifier's decision rule, but naive Bayes is not a Bayesian method. Naive Bayes classifiers are a collection of classification algorithms based **on** Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

3.4 Neural Networks

Neural networks can be used to model complex relationships between inputs and outputs to find patterns in data. By using neural networks as a tool, data warehousing firms are gathering information from datasets in the process known as data mining. A neural network classifier is a network of units, where the input units usually represent terms, the output unit(s) represents the category. For classifying a text document, its term weights are assigned to the input units; the activation of these units is propagated forward through the network, and the value that the output unit(s) takes up as a consequence determines the categorization decision. Suitability for both discrete and continuous data makes neural network a popular choice for text classification purpose. The term neural network was traditionally used to refer to a network or circuit of neurons.[1] The modern usage of the term often refers to artificial neural networks, which are composed of artificial neurons or nodes. Thus the term may refer to either biological neural networks, made up of real biological neurons, or artificial neural networks, for solving artificial intelligence (AI) problems.The connections of the biological neuron are modeled as weights. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. This activity is referred as a linear combination. Finally, an activation function controls the amplitude of the output.

3.5 Rocchio's Algorithm

The Rocchio's algorithm is based on a method of relevance feedback found in information retrieval systems which stemmed from the SMART Information Retrieval System around the year 1970. In this algorithm, a prototype vector is built for each class. A prototype vector is average vector over all training document vectors that belong to class ci[6]. Similarity between text document and each of prototype vectors is determined and text document is assigned to the class having maximum similarity. The algorithm is based on the assumption that most users have a general conception of which documents should be denoted as relevant or non-relevant. This algorithm is deemed as very fast learner and

easy to implement. Although easy to implement, this algorithm suffers from poor classification accuracy. The selection of values for the constants alpha and beta plays a vital role in its performance. Like many other retrieval systems, the Rocchio feedback approach was developed using the Vector Space Model. The algorithm is based on the assumption that most users have a general conception of which documents should be denoted as relevant or non-relevant. Therefore, the user's search query is revised to include an arbitrary percentage of relevant and non-relevant documents as a means of increasing the search engine's recall, and possibly the precision as well.

3.6Decision Tree

Decision tree is one of the most popular technique for prediction. Most of the research have used this technique because of its simplicity and comprehensibility to uncover small or large data structure and predict the values. Decision tree in data mining is one of the simplest and easiest methods which are most frequently used by the researchers on their work. A tree has many analogies in real life, and turns out that it has influenced a wide area of machine learning**,** covering both classification and regression. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. The root node of the decision tree is a top node resembles simple question also called as a posture that bears multiple branches called sub nodes with answers for the root node question. In turn each answer related to a set of questions or conditions that help us to predict the data, on which the final decision is made. ID3 and C4.5 are called as induction algorithm of decision tree developed by the researcher called Ross Quinlan. Both algorithm supports greedy method, top-down recursive in divide-and-conquer manner and they does not support backtracking.C4.5 is also known as superset of ID3.The advantages of this technique are, it doesn't require detailed knowledge, it deals with complex data, these are easy to understand, and data Classification becomes simpler, makes learning easier, it produces very accurate end result.

IV. CONCLUSION

Text classification is a widespread domain of research encompassing Data mining, NLP and Machine Learning. It has witnessed much heed owing to the high growth rate of internet and relevance of internet search engines. This review paper circumscribes existing literature and explores the document representation and analysis of feature extraction methods and broaches to different available classifiers. Various methods of classification and feature extraction have been compared and contrasted with all coeval methods based on different parameters like time complexities and performance. It is deemed that no single representation scheme and classifier can be mentioned as a general model for any application. Performance of different algorithms varies according to the data collection. However, SVM with term weighted VSM representation scheme has shown some potential results in the tasks of text classification up to some extent but still universal acceptance of this algorithm remains implausible.

REFERENCES

[1] An effcient Wikipedia semantic matching approach to text document classification, Information Sciences, 2017.
[2] Classification of text documents based on score level fusion approach, Pattern Recognition Letters, 2017.
[3] Combination of global and local contexts for text/non-text classification in heterogeneous online handwritten documents, Pattern Recognition, 2016.
[4] Feature selection for document classification based on topology, Egyptian Informatics Journal, 2018.
[5] Performance Comparison and Optimization of Text Document Classification using k-NN and Naïve Bayes Classification Techniques, Procedia Computer Science, 2017.
[6] Variable Global Feature Selection Scheme for automatic classification of text documents, Expert Systems With Applications, 2017.
[7] A novel multivariate filter method for feature selection in text classification problems, Engineering Applications of Artificial Intelligence, 2018.
[8] A link-bridged topic model for cross-domain document classification, Information Processing and Management,2013.
[9] Employing fisher discriminant analysis for Arabic text classification, Computers and Electrical Engineering, 2017.
[10] Efficiency investigation of manifold matching for text document classification, Pattern Recognition Letters, 2013.
[11] A contemporary feature selection and classification framework for imbalanced biomedical datasets, Egyptian Informatics Journal, 2018.
[12] Wikipedia-based cross-language text classification, Information Sciences, 2017.
[13] Text Plagiarism Classification using Syntax based Linguistic Features, Expert Systems With Applications, 2017.
[14] Evaluation of a rule-based method for epidemiological document classification towards the automation of systematic reviews, Journal of Biomedical Informatics, 2017.

[15] Document classification algorithm based on MMP and LS- SVM, Procedia Engineering, 2013.