IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 20, Issue 2, Ver. III (Mar. - Apr. 2018), PP 51-56
www.iosrjournals.org

A Survey on Machine Learning Based Text Categorization

Ayesha Priyambada Das[*1], Dr.Ajit Kumar Nayak [2], Mamata Nayak[3]
(Scholar, Computer Science and Information Technology, ITER, Email:dasayesha.55@gmail.com)[1]
(Head of the Department, Department of Computer Science and Information Technology, ITER, (Associate Professor,Department of Computer Science and Information Technology,ITER,
*Corresponding Author: Ayesha Priyambada Das**

Abstract: *Due to the availability of documents in the digital form becoming enormous the need to access them into more adjustable way becoming extremely important. In this context,document management tasks based on content is called as IR or Information Retrieval. Thishas achieved a noticeable position in the area of information system.For faster response time of IR,it is very important and essential to organize,categorize and classify texts and digital documents according to the definitions,proposed by Text Mining experts and Computer scientists.Automatic text Categorization or Topic Spotting,is a process to sorta document set automatically into categories from a predefined set.According to researchers the superior access to this problem depends on machine learning methods in which,a general posteriori process builds a classifier automatically by learning pre-classified documents given and the category's characteristics.The acceptance of automatic text categorization is done becauseit is free from the need of organizing manually the bases of the documents, which can be too costly,and clearly not possible because of the given time limit of the applications or the number of documents included. Thanks to the technology including both Information Retrieval (IR) and Machine Learning (ML) for the nicety of modern text classification system.Our aim here is to work in Text Documents Classification.It also aims towards the comparison and construction ofvarious available classifiers depending on few benchmark such as time complexity and performance.*

Keywords : *Text Classification, Text Mining, Machine Learning, Support Vector Machine.*

Date of Submission: 09-04-2018 Date of acceptance: 23-04-2018

I. Introduction

Textual data involves very important information in the document forms.Collection oftextual data becomes more beneficial when the extraction of important information contained by it, is done. The extraction of different valuable information from a vast document set efficiently is known as text mining. Text mining involves four steps, they are: Collection of text data, Analysis of the data, Interpretation and extraction of information. Information Retrieval (IR) isinvolving a set of procedure tofind documents which includes the answers to the questions. For achieving this objective, statistical standards, techniques are used for process the textual data automatically and compare the given question. Information Retrieval process is the allotment of the complete range of information processing, from document retrieval to knowledge retrieval. NLP or Natural Language Processing is the process to carry out a better understanding to the natural language by the use of computer. It represents the documents lexicographically to improve the Classification process.

Categorization of text associates to the part of data analysis within the mechanism which manages the data collections. It structures the unstructured or semi-structured textual data by giving tags to the documents. As the fast growing Internet, leads to the rapid growth of the document collection, managing such a huge collection of documents is becoming a difficult process. For analyzing the complex data the most common theme is the classification or categorization of elements. The whole mechanism is to classify a given document or information into pre-set categories. This crucial task is known as Text Categorization (TC) or Topic Spotting or Text Classification belongs to the domain of document management. In this task there is given a set of categories (subject and topics) andset of text documents and the process is involving, choosing the correct and exact subject for eachdocuments. To establish any kind of automatic system which contains text document, to be used as data, it is important to draw out the high level of excellent information effectively. Byusing different techniques from Statistics, Information Retrieval, Machine Leaning system, Text Mining extracts quality information from the textual data. As I mentioned before, Text Categorization is the conversion of set of documents into different categories by considering a pre-classified set. Different researches say that the governing approach to this task deal withMachine Learning where, a preliminary process creates a classifierautomatically by learning characteristics of a pre-classified document sets.

Before until late '80s one of the most popular approach to TC was the *Knowledge Engineering (KE)* approach which was based on defining manually a set of rules,encoding expert knowledge for the categorization

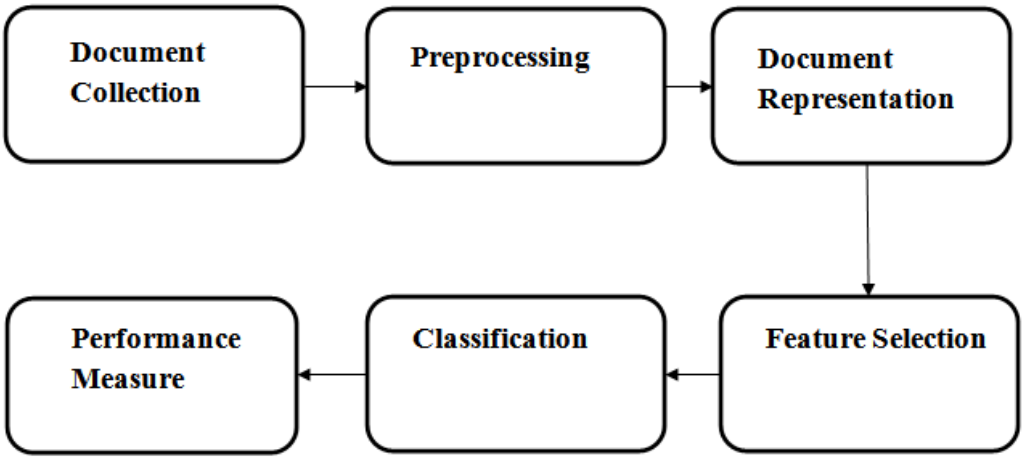
of different informationwith the given set categories to govern the process. In the early '90s this mechanism has wayward acceptability and Machine Learning approach came into consideration. For learning and generalizing an input to output mapping is the work of supervised machine learning method.In case of text categorization, the input will be a set of documents and the output will be their categories respectively.

II. Literature Survey

Anuag Sarkar, Saptarshi Chatterjee, Writyan Das, Debabratta Datta, "Text classification using Support Vector Machine". In this paper the author has defined the process of Text-based Classification technique from the application's point of view. A text-based classifier has been implemented in this paper that can be used to classify input text into one of the two categories. Here as described by the author, the classifier is first trained with an initial dataset using supervised learning. After the training process, classifier uses the trained data to classify any new input text. The classifier according to them achieves a reasonable rate of accuracy even though it has been implemented using simple techniques.

Mita K. Dalal, Mukesh A. Zaveri "Automatic Text Classification: A Technical Review". In this paper the author has described text classification can be automated using machine learning techniques. Preprocessing and feature selection steps in the process play vital roles in the quantity of the training input data which affects the classification accuracy. Here the author has also mention that, Text Classifiers are not yet available for several regional languages and it would be useful for several commercial and government projects, if it is available for the several different languages. *Franeesca Possemato and Antonello Rizzi "Automatic TC by Granula Computing approach: facing unbalanced data sets".* A system is proposed by the authors in this paper by relying on a Granular evaluation mechanism to consider a document as series of words. The authors proposed a system here, which is able to mine frequent terms automatically, for the consideration of a term rather than a single word. The entire process of Classification is able to process sequence of automatic elements by means of an embedded method depending on clustering mechanism. A well-known data set with a extraordinary less anatomical complication of the combination of models of classification can be evaluate effectiveness of the system. *Dai Li and Yi L. Murphey "Automatic Text Categorization using a system of High-precision and High-recall models".* The paper represents a system of a Document Categorization system which operates automatically, known as "HPHR", which includes high recall, high precision and noise-filtered Text Categorization models. Here specific models are described by the authors that are generated based on machine learning algorithms, a cluster algorithm which is speedy, efficient and effective in grouping the documents to various sub categories and an algorithm for text category generation that assigns text to sub-categories automatically which represents high recall, high precision and noise filtered models for Text Categorization using the given set of training data. It is also mentioned that, the performance of HPHR displayed superiority over the systems generally used in Text or Document Categorization.

III. Process Of Automatic Text Categorization



(**Figure 1**: Overall process of Automatic Document Categorization task)

3.1.Collection of Document:

It is the very first step of Categorization method. Different categories of documents such as doc, html, web based, .pdf documents are collected in this method.

3.2.Preprocessing:
A lot of text mining preprocessing techniques are there. The documents are structured by all the following steps. They are as follows:
Tokenization*:*Here in this step, document is divided into a list of tokens by treated as a string.
Removing stop words*:*The useless words for the process like, "a", "and", "the", can be eliminated without affecting the process.
Stemming word*:* A stemming algorithm can be applied to convert various word forms into similar form.
3.3.Document Representation:
Documents need to convert into more manageable representation during preprocessing step because the text documents can't be processed into their original form directly by some classifiers and learning algorithms. The most common model"*Bag-of-word"* issued for the conversion of the text into more controllable form. All the words in a document can be used by this model simply as the features. A feature can be defined as a individual unit without any internal structure, which can be simply described as a dimension in the feature space. The representation of the document is a vector in the space, which is the order of features and their weights.This method of assigning weights to the feature may be different. The simplest mechanism is the *Binary* one.The most acceptable *TF-IDF* technique which gives the word(w) a weight in the document (d) can be expressed by the equation:

$$TF\text{-}IDF\ weight(w,d) = TermFreq(w,d).log(N/DocFreq(w))$$

3.4.Feature selection:
In a large document collection of the number of different words can be huge. This kind of wordcan be dropped without doing any harm to the classification method or to the accomplishment of the classifier because these are extraneous to classification task. The step of preprocessing that eliminates the extraneous words is known as *Feature selection*. A lot of feature evaluation metrics have been considered, they are - Expected cross entropy, Information gain (IG), the weight of evidence, Gini index, Term frequency, Odds ratio and Chi-square.

3.5.Classification:
A Classifier is created automatically by a process of learning the assets of categories from bunch of pre-classified training data. In the terminology of machine learning method, the learning approach deals with*supervised machine learning*. The cause behind this is, the process is accompanied by utilizing the true category assignment function which is known on the training set. There are a lot of methods that aretaken into consideration for classifier learning. Some classification algorithmshave been built particularly for categorization process.
They are:

1) Probabilistic Classifier
2) Bayesian Logistic Regression
3) Decision Tree Classifier
4) Decision Rule Classifier
5) Neural Network
6) Support Vector Machine

3.5.1. Probabilistic Classifier:
By considering the Probabilistic classifiers, the status ofCategorization value $CSV_i(\vec{d}_j, c_i)$ as probability $P(c_i|\vec{d}_j)$ which shows that document d belongs to the category c and the probability is calculated by the application of Byes' theorem can be implemented as below.

$$P(c_i|\vec{d}_j) = \frac{P\left(\vec{d}_j|c_i\right) . P(c_i)}{P(\vec{d}_i)}$$

Where,
$P(\vec{d}_j)$*:* Describes margina
a randomly picked docume
is d_j may be binary or weig
$P(c_i)$*:* Probability of a docu

Thecalculation$P(\underset{d_j}{\rightarrow}|c_j)$ is ambiguous. As the possibility of number for $d_i$vector is high. Tosolve this problem, some assumptions we have to make on the structure of the *d*. Most common assumption for this problem involving that all co-ordinates are independent. This independent assumption can be encoded as the given equation below,

$$\mathbf{P}(\underset{d_j}{\rightarrow}|c_i)=\prod_{k=1}^{|I|} P(w_{kj}|c_i)$$

The result from this assumption is the p
Usually this particular assumption is n
independence classifier is one of the best known naïve based method, which results by the use of vector of binary-valued representations for the documents. A paper regarding Naïve Bayes classifiers by Lewis(1998) is prove to be an excellent approach on various form that is being researched on Naïve Bayes classifier. Some of the advantages of this algorithm is given below:

- Methods of probabilistic are quantitative in nature.
- The performance of the system is improved by this.

3.5.2. Bayesian Logistic Regression:

BLR or Bayesian Logistic Regression is an old statistical method, which is quickly gaining popularity in the area of TC, due to its apparently very high performance. It can be described as:

$$\mathbf{P(c|d)= \phi\ (\beta,d)=\phi(\sum_i \beta_i . d_j)}$$

c : value of Category membership, whose value can be +1 or -1

d : that is ($\boldsymbol{d_1}, \boldsymbol{d_2}$, ...): Represented as of document in the feature space.

β ($\boldsymbol{\beta_1},\boldsymbol{\beta_2}$,……): Parameter vector.

Φ : Logically link function.

Some of the issues regarding this algorithm,

- For a Logistic Regression Model, Proper protection should be considered that isnot to over-fit the training data.
- As different priors can be possible, the most common used priors are Laplace and Gaussian.
- B -approach can be used as a prior dispersion for parameter vector β, it will accreditprobability that is high in value to each β_i's being zero or near to that.

3.5.3. Decision Tree Classifier:

Probabilistic mechanisms are determinable in nature so they are not easily understandable by human being. Some of the algorithmswhich are not suffer from this type of problem and can be easily understandable by humans are Symbolic algorithms. Decision Tree Classifiers fall under symbolic classifier. A DT or Decision Tree Classifier cab be defined to be a tree where, internal nodes are represented as features, edge joining nodes (leaving) is representing as by tests on feature's weight, leafs are representing categories. Starting from the root of the tree DT categorize a document and by moving downward until a leaf node is reached, through branches whose conditions are satisfied by the documents. Most of the classifiers usethe representations of binary document and binary trees. Some of the standard packages for DT-based systems are*ID3,C4.5* and *CART*.Most acceptable method is depend on*Divide and Conquer* strategy in order to learning a DT for category c_i, which involves in checking if all the training examples have the same level of importance or not. If not, then it will select a term t_k then it places each of the class in a separate sub-tree.

- DT classifier is the top-ranking classifiers and its performance is mixed.
- It is hardly used in the cases where, the understanding by human to the classifier is not necessary.

3.5.4. Decision Rule Classifier:

These classifiers are also *Symbolic Classifiers* as DT. In this Classifier**,** a classifier is formed by method of Inductive Rule learning for each of categoryc_i. Rules like this are alikeDNF or *Disjunction Normal Form* rules. Rules for CONSTRUE system are formed from the training set using inductive rule of learning.DNF rules are very much alike of DT's, they can encode any of the Boolean function but,the advantages on DNF learner over DT learner is that, they are more likely to createmore condensed classifiers than DT learner. Learningmethods by rules usually selects as best one from all the possible rules depend on some minimum criteria. Another difference is, DTs are created usually by *Divide and Conquer*strategy, while DNF rules are created by bottom to up strategy. In DNF every training example $d_j$is viewed as one one class initially$n_1,\ldots,n_n \rightarrow \gamma_i$where $n_1,\ldots,n_n$are terms included ind_j and γ_i equals negative example of c_i.

- Heuristic and optimality criteria define DNF learners vary in their respective methods.

- RIPPER is one of the prominent member of this family.
- The ability to bias the performance for the higher precision is an attractive features of RIPPER.

3.5.5. Neural Network (NN):

To perform text categorization mechanism, neural network can be used. A NN or Neural Network of classification of textual documents is defined as a network combination of units. Input units ofthe network represent the terms and output units represent the categories. Nodes, representing input to the network receivesvalues of features and the output nodes producesvalue of the category status. Link weight represents the relation of dependencies. For categorization process a textual datad_j, feature weights $w_{kj}$is pushed into the input units. All the units are activated by forwarding the network and output unit determines the decision of categorization process. The neural network is skilled by *Back Propagation network*where, the weight of the terms of a document meant for training are uploaded into the input units. If miss-classification happens, the error is back propagated through the network for the minimization of rate of error by modifying some of the values of parameters.

- *Perceptron*is the simplest form of neural network such networks are equivalent to linear classifiers.
- A non-linear NN in TC represents higher-order interaction in between the terms.

3.5.6. Support Vector Machine (SVM) -

Support Vector Machine (SVM) is proved to be a very fast and adequate algorithm for the problem of text categorization. In thegeometric term, SVM classifier (binary)could be termed as a *Hyper-plane.* Itseparates the elements that representpositive instances from the negative instances of category in the feature space. So the *Hyper-plane*classifier is picked during training which maintains a maximum margin between the known positive instances and the known negative instances. SVM *Hyper-plane* are completely driven with a approximately smallersubset of the training data called, *Support vectors* and rest of the training data usually have no effect on the trained classifier.

The whole aim is to completely understand in the case of positive and the negative instances that are linearly separable. In the case of 2-dimensional system, different lines could be picked as decision surfaces. According to the SVM mechanism, it chooses the middle elements from a large set of parallel strait lines in which, maximal distance within two elementsina set of element is highest. Most appropriate decision surface is found out by the help of comparatively a small set of training data examples, defined as*Support vector*.Two important advantages for TC by SVM are:

- There is no need for term selection because SVMs are likely to be adequately boisterous to use a statistical model that has too many parameters and can possibly scale up to considerable dimensionality.
- In case of parameter training, there isno need of validation for human and machine effort.
- Algorithm of SVM arises unique from different algorithm of categorization for its *Support vectors*.

Issues to be consider while using ML techniques:

There are four primary issues that are needed to be considered while using machine learning techniques for Text categorization process.

1. First of all, we have to decide on categories which are going to be used forthe classification the data instances.
2. For each of the categories we need to provide some training sets.As a rule of thumb, for each of the categories around 30 examples are needed.
3. Decision will be made on the features which represents each of the instances. Most of the algorithms should have the capacity to focus on the appropriate features.
4. Decision will be made for the algorithm that is to be used for the process of Categorization.

3.6.Performance measure:

As we are conducting performance evaluation of a text searching system, computation for text classifier is primarily conducted tentatively rather than analytically. The reason behind this is to prove a system that is correct or not analytically, we possibly need a formal specification to solve the task. The experimental calculation usually measures the effectiveness of the classifierwhich is able to take the right decision rather than the efficiency of the classifier.

IV. Conclusion

As we know, Text Categorization process plays an important role in the issues of Information retrieval, Text mining and machine learning. It is found to be very much successful in tackling variety of real world problems due to involvement of Machine leaning community in Text Categorization system. Machine Learning algorithm provides many techniques for Text Categorization and Text Mining related issues. Process to Text

Categorization and many approaches have discussed in this paper. In the process of TC, the problem of dimensionality is reduced by Feature selection. Support Vector Machine learning approach is the best among the other supervised learning techniques. A lot of augmentations can be made for the feature preparation and for the classification engine. So we can say, design of Text Classification system is still more an art than exact science.

References

[1] Anurag Sarkar, Saptarshi Chatterjee, Writayan Das, Debabrata Datta, "*Text Classification using Support Vector Machine*", International Journak of Engineering Science Invention, *volume 4, Issue 11, November 2015*.

[2] Istvan Pilaszy, "*Text Categorization and Support Vector Machine*",Department of Measurement and Information System Budapest University of Technology and Economics.

[3] Fabrizio Sebastiani, "*Machine Learning in Automated Text Categorization*",ACM computing survey, Consiglio Nazionale delle Ricerche, Italy, *26th October 2001*.

[4] Mojgan Farhoodi, Alireza Yari, "*Applying Machine Learning Algorithm for Automatic Persian Text Classification*", Iran Telecommunication Research Center.

[5] Pradnya Kumbhar, Manisha Mali, Dr. Mohammad Atique, "*A Genetic-Fuzzy Approach for Automatic Text Categorization*", IEEE 7th International Advance Computing Conference, *2017*.

[6] Francesca Possemato, Antonello Rizzi, "*Automatic Text Categorization by a Granular Computing Approach: facing Unbalanced Data Sets*", IEEE conference.

[7] Dai Li, Yi L. Murphey, "*Automatic Text Categorization using a System of High-Precision and High-Recall Models*".

[8] Ahmed Faraz, "*An Elaboration of Text Categorization and Automatic Text Classification through Mathematical and Graphical Modelling*", Computer Science and Engineering: An International Journal (CSEIJ), *volume 5, June 2015.*

[9] Abdelwadood Moh'd A MESLEH, "*Chi square Feature Extraction Based SVMs Arabic Language Text Categorization System*", Journal of Computer Science, *2007*.

[10] Mita K. Dalal, Mukesh A. Zaveri, "*Automatic Text Classification: A Technical Review*", International Journal of Computer Applications(0975-8887), *volume 28, August 2011*.

[11] Pooja Bolaj, Sharvari Govilkar, "*A Survey on Text Categorization Techniques for Indian Regional Languages*", International Journal of Computer Science and Information Technology, *Volume.7, 2016*.

[12] B. Mahalakshmi, Dr. K. Duraiswamy, "*An Overview of Categorization Techniques*". 2249-6645, International Journal of Modern Engineering Research(IJMER), *Oct 2012*.

[13] Sumanta Kashyapi, Dr. Madhu Kumari, "*Research Issues in Text Categorization based on Machine Learning: A Review*", International Journal of Advanced Technology in Engineering and Science, *Volume. No.01, January 2016*.

[14] Youngjoong Ko, Jungyun Seo, "*Automatic Text Categorization by Unsupervised learning*",In Proceedings of the 18th conference on Computer Linguistics, *Volume.1, pages 453-459, 2000*.

[15] Charu C Aggarwal and ChengXiang Zhai, "*Mining Text Data*", Springer Science and Business Media, *2012*.

[16] Bhumika, Prof Sukhjit Singh Sehra, Prof Anand Nayyer, "*A Review Paper on Algorithms used for Text Clasiification*", International Journal of Application or Innovation in Engineering and Management(IJAIEM), *Volume 2, Issue 3, March 2013.*

[17] Vandana Korde et al, "*Text Classification and Classifiers*", International Journal of Artificial Intelligence and Applications(IJAIA), *Volume.3, No.2, March 2012*.

[18] Senthil Kumar B, Bhavitha Varma E, "*A survey on Text Categorization*", International Journal of Advanced Research in Computer and Communication Engineering, *Volume.5, Issue 8, August 2016*.

[19] CT.vidya, S.M.Nithya, T.Vishnu Priya, "*A Survey on Text Classifiaction Techniques and Applications*", International Journal of Advanced Research in Engineering and Technology(IJARCET), *Volume 5, Issue 1, January 2016*.

[20] Edda Leopold, "*Text Categorization with Support Vector Machines,How to Represent Texts in Input Space?*", Kluwer Academic Publisher,*423-444,2002*.

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

Ayesha Priyambada Das*. " A Survey on Machine Learning Based Text Categorization." IOSR Journal of Computer Engineering (IOSR-JCE) 20.2 (2018): 51-56.