A Survey on Text Classification using Machine Learning Algorithms

Harshitha C P[1], Ramya K[2], Agni Hombali[3], Ranjana S Chakrasali[4]
Department of CSE, BNMIT, Bangalore, India
[4]Assistant Professor, Department of CSE, BNMIT, Bangalore, India

Abstract— **In today's world, the usage of digitalized text documents has drastically increased. The reason behind this is the growing need for portability of text related files and the greater need to eliminate the dependence on paper. Previously, the task of document classification was handled by very experienced experts who are capable of classifying large text documents into their corresponding category. Overtime, it was realized that this task extremely time consuming. Therefore the need for automatic text document classification came into the big picture. Corresponding research has shown the involvement of various classification algorithms to create an automated text document classification system. The major tasks involved in creating this type of automated system is handling large amount of texts, selecting the features from a wide range of availability and eventually selecting the classification algorithm which is best suited for classification text files. Initially the predefined classes are considered and then only the document is classified into the correct category. The system proposed to be created is significantly depending on meaningful words and efficient features while classifying the documents. The proposed methodology explores different algorithms in order to select the best algorithm possible for classifying text documents[1].**

Keywords—Pre-processing, Feature extraction, Feature Selection, K-nearest Algorithm

I. INTRODUCTION

The task of classifying documents is significant and prominent task in the IT industry. The main task is of basically analyzing the content in the given document and then classifying that particular document into the correct category. Today's generation is mostly based on technological advancements. The used of hard paper and physical copies are slowly being eradicated and it is getting close to extinction[1]. The new techniques acquired is of the digital copies and soft copies. The reason for this advancement is because of portability and size issues. In the previous generation the task of classifying documents was being done only by experts and even experts used to find it difficult and time consuming. In order to overcome this particular issue, research on automated text document categorization is a fast forwarding field of interest. The evolution of the internet is one of the biggest factors for this type of advancement. Keeping in mind the demand for information form the internet and also the degree of the unstructured data present in it, the need for automated text document classification becomes even more dire. Due to this, the method of text categorization has become a very efficient method in organizing unstructured data. In the recent years, the use of search engines were predominantly used to get information from the internet. Search engines like Yahoo and Google were widely used. These search engines used to filter out the required information and then present it to the user.

The issue with these search engines were they were not very experienced in filtering out information[2]. Often these search engines used to filter out irrelevant information adding to the manpower of the user. In the present times, these search engines are well trained by using machine learning. Now the present states of the search engines are efficient and effective. Feature Selection and Feature Weighting is done and classified into pre-defined classes. In order to categorize the document into the proper category, the classification system has to be trained properly and not to mention accurately. Large number of training document sets have to be proposed and be fed to the classifier for standards. In a way, the more training sets are provided, more will accurate will be the classification of the text document. In the spirit of finding a more efficient method to classify documents, a new mechanism is proposed to classify the documents[3]. The basic function of document classifier is of assigning a particular document to its corresponding category by using machine learning. The classification is basically done by selecting particular features from the text document and then prioritizing them by providing them with weights. Based on this kind of prioritization, the document will be classified effectively.

The features that are being extracted are differentiated by using algorithms of machine learning like KNN, Naïve Bayes, Support Vector System etc.

The following are the applications of document classification :

1. E-mail addressing: Depending on the topic of the e- mail, a particular e-mail is routed to a general address, a specific address or the mailbox.
2. Language identification: In recognizing a language, the text document classification can be used effectively to categorize a document into a particular language. In a way, it will be user to select his/her

language without much time consumption. Most of the languages are considered form left to right and top to bottom. There are few exceptions where the languages are written form right to left. Keeping such sensitive information in mind, the classifier can narrow its search down to some percentage of the total languages.

3. Readability assessment: it is basically categorizing to what appropriate readers can this particular document be provided to in order to read.

Due to the increase in the dependence on the internet, the population keeps track on the websites and blogs and the digital version of the text files[1].

II. LITERATURE SURVEY

Literature survey is the list of researches done by the user to acquire certain information about the particular topic. It’s the initial foundation of the proposed system.

In this paper [1] proposed by Vrusha U.Suryawanshi et al. the methodology explained proposes that the text document that is being considered to be classified should be assigned to some pre-defines classes before classification. The best and the most efficient method to classify any particular document is the k- nearest method. The classification done is basically by extracting a set of the most important key words out of the text document and then analyzing them. There are many algorithms for classifying documents. The best suited algorithm for classifying documents in this particular proposed method is the k-nearest method. The proposed method uses Virtual Private Networks for security reasons. the provisions of this systems enables the user to store large amounts of information so that it can be used for training the classifier an classify the documents as well. The system is completely automated and eliminates the need for human interaction. The security systems keep the process protected. This methodology eliminates the overhead processing.

In this paper [2] proposed by Aiman Moldagulova and Rosnafisah Bte. Sulaiman, the proposed method uses the k- nearest algorithm to classify the documents. The value of the k in this particular system is a very crucial part in this system. The determinations of the value of k represents to which all categories the selected document is tested against. K-nearest is easy to implement and efficient than other classification algorithms. The two criterion kept in mind while determining the value of k is the validation error rate and the training error rate. The results show best efficiency when the value of k is between 1 and 50.

In this paper [3] proposed by Ari Aulia Hakim et al, the proposed system employs the term-frequency and inverse document frequency to prioritize the keywords by weighting it. The keywords are weighted by using term frequency (TF) in order to find the number of times the keyword has occurred in the document. Then Inverse Document Frequency (IDF) is used to find the number of documents in which the keywords are found.

This method is found to be very efficient but there seems to be a small defect. Since the method considers all the keywords and weights them, the keyword with the highest frequency might not relate to the corresponding category. Sometimes this system may give a wrong output.

In this paper [4] proposed by Seyyed Mohamma Hossein Dadgar et all, the proposed methodology suggests 3 parts to the system:

1. Pre-processing
2. Feature extraction using TF-IDF
3. Classification based on Support Vector Machine (SVM)

SVM is a very good classifier. It classifies the documents based on the lowest structural risk principle and creation of hyperplanes. The main disadvantage of this system is that the SVM uses all the keyword regardless of whether they are important or not. Sometimes this might give a wrong output.

In this paper [5] proposed by Akshita Bhandari, the proposed idea suggests an improvement in the Apriori algorithm. This algorithm is used to extract frequent itemsets from the database and then get the association rule for discovering the knowledge. There are 2 requirements needed for this algorithm namely minimum support and minimum confidence. Minimum support is used find frequent itemsets and the minimum confidence is used to find the association rules.

In this system, the size of the database is drastically reduced. Due to this the results of the Apriori algorithm will give promising results.

In this paper [6] proposed by Yiming Yang and Jan O. Perderson, the feature selection is one of the most important part in the document classification. There are many methods of feature selection namely

1. Information Gain
2. Mutual Information
3. Chi square statistic measure
4. Document Frequency Thresholding

Out of all these methods, Document Frequency Thresholding proves to be the most efficient method.

III. STATE OF ART

From the literature survey, it has been observed that there are problems with all the methodologies involved. To overcome the existing problems, a comparative analysis is done by using one algorithm “k-nearest neighbour” for a machine learning where one more step called Feature Selection is included which selects top Ranked keywords for each category that reduces the dimensionality of feature space and gives accurate results compared to previous results[2].

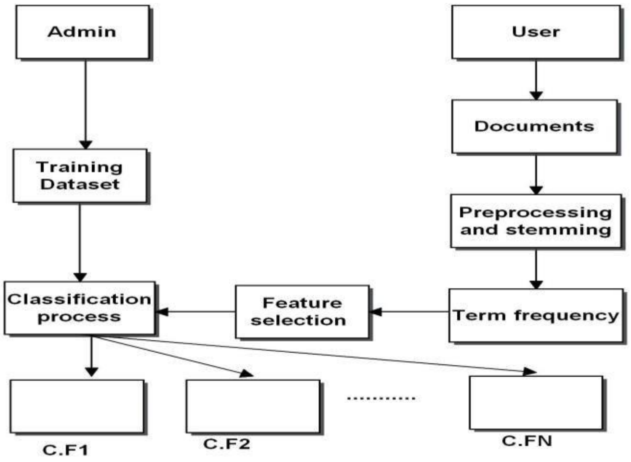

Fig 4.1. System Architecture

1. Pre-processing Technique: Pre-processing is the method used to eliminate all the unwanted words from the text document[6].

- **Stop List Removal:** This is the process of removing all that unwanted words and conjuctions.
- **Stemming:** This is the process of removing all the white spaces.

2. Feature Extraction: In this step Feature Weighting is done, which helps in assigning weights to words in a text document.

Bag of words : The bag of words is the most common and the simplest among all the other feature extraction methods; it forms a word presence feature set from all the words of an instance. It is known as a “bag” of words, since the method doesn't care about how many times a word occurs or the order of the words, all what matters is whether the word is present in a list of words. The features can be used in modelling with machine learning algorithms. This method is very flexible and simple. It is usually used for extracting features from text data in various ways. A bag of words is the presentation of text data. It specifies the frequency of words in the document. It includes: 1. A lexicon of known words 2. A frequency of the existence of those known words. The complexity of bag of words model is both in determining how to score the presence of familiar words and how to design the vocabulary of familiar words[6].

TF-IDF : A problem with bag of words approach is that the words with higher frequency becomes dominant in the data. These words may not provide much information for the model. And due to this problem domain specific words which does not have larger score may be discarded or ignored. To resolve this problem, the frequency of the words is rescaled by considering how frequently the words occur in all the documents. Due to this, the scores for frequent words are also frequent among all the documents are reduced. This way of scoring is known as Term Frequency – Inverse Document Frequency. • Term Frequency (TF) is the frequency of the word in the current document. • Inverse Document Frequency (IDF) is the score of the words among all the documents. These scores can highlight the words that are unique that is the words that represent needful information in a specified document. Therefore the IDF of an infrequent term is high, and the IDF of a frequent term is low[4].

Word2Vec: Word2Vec is used to construct word embeddings. The models created by using word2vec are shallow meaning two-layer neural networks. Once trained, they reproduce semantic contexts of words. The model takes a huge corpus of text as an input. It then creates a vector space which is usually of hundreds of dimensions. Each distinctive word in the corpus is alloted with corresponding vector in the space. The words with common contexts are placed in near proximity in vector space. Word2vec can use one of the two architectures: continuous skip gram or continuous bag of words (CBOW). In the continuous skip gram, the current word is considered to predict the neighbouring window of context words. In this architecture the nearby context words are considered more heavily than words with distant context. In the continuous bag of words architecture, the sequence of context words does not impact the prediction as it is based on bag of words model[4].

3. Feature Selection: When all the keywords recorded contribute to the classification of the documents, it imposes a negative effect on the classification. All the words cannot be considered for classification. Therefore there are only few keywords selected which can be significantly contributing to the accurate classification of the text document. This process is called feature selection[6].

Document Frequency Threshold : This method is widely used. In this method. The document frequency of the word is obtained. Document Frequency of the word suggests that in how many documents, this particular keyword is found. If the Document Frequency of the keyword exceeds a certain threshold then it is considered as a selected keyword for further classification[4].

Information Gain(IG): Information Gain is strongly based on entropy of a particular term. It is calculated by analysing the presence and absence of a term in a particular document. It is a type of category prediction which involves the development of a decision tree[6].

Mutual Information: This method is used to get a co-relation between 2 or more chance variables. The first step is to find the entropy of a single discrete chance variable. Then the joint and conditional entropy of 2 or more chance variable is obtained. Eventually the rate of transmission between the chance variables is found[6].

CHI Statistics: This method is very essential in classification where there is a flavor of statistics involved. The chi square basically determines how different is a particular term form the expected results. It is basically finding out the divergence of an attribute from a specified expectations of an attribute[6]. **Term Strength:** Term Strength is basically a degree of how much a term is found in commonly related documents. It doesn't need a pre-defined list of Stop Words - it discovers them automatically[6].

1. So it's a technique for vocabulary reduction in text retrieval.
2. This method estimates term importance based on how often a term appears in "related" documents.

Strength of a term t

1. Measures how informative a word is for identifying two related documents.
2. s(t)=P(t∈y|t∈x)
3. For two related documents x,y what's the probability that t belongs to y given it belongs to x?
4. Estimate s(t) on training data using Maximum Likelihood Estimation.

Classification Algorithm:

KNN : KNN stands for k-nearest neighbor. The classification based on KNN algorithm is strongly dependent on the determination of the value of k. The classification is most accurate when the value of k is between 1 and 50. It's basically finding the nearest neighbours of a particular term[6].

Support Vector System : This is an algorithm that uses non- linear mapping to transform the original training data to higher dimension. It basically revolves around developing hyperplanes for the classification of documents. The main disadvantages of SVM is they are slow. The main advantage of SVM is that it is very accurate[6].

Decision Tree : A Decision tree is basically a guide to make decision. It's a tree formed after analysis of the constraints. In this method the most commonly used measures is the Information Gain[6].

IV. CONCLUSION

The text document classification is an essential tool for classifying text documents into categories for further analysis. It proves to be less tedious and less time consuming than the conventional methods of document classification. The methods and algorithms used in this project is considered after strict evaluation the comparison between their performances of the existing methods and algorithms. The performance of the system is observed to be optimal and very efficient in classifying documents. The system involves a pre-processing method to eliminate all the unwanted words to filter the text document so that it can be classified in a better way. The feature extraction and feature selection proves to be an efficient method to prioritize the keywords in the text document. Eventually, the classification algorithm will consider the keywords and their weights and then classify the document under a category. This route of analyzing and classifying the document is proves to be a very good way to categorize the document. The time taken for categorization is also found to be very optimal.

REFERENCES

[1] Vrusha U.Suryawanshi, Pallavi Bogawar, Pallavi Patil, Priya Meshram, Komal Yadav, Prof. Nikhil S. Sakhare, "Automatic Text Classification System", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 2, February 2015.

[2] Aiman Moldagulova, Rosnafisah Bte. Sulaiman, "Using KNN Algorithm for Classification of Textual Documents", 8th International Conference on Information Technology (ICIT) 2017.

[3] Ari Aulia Hakim, Alva Erwin, Kho I Eng, Maulahikmah Galinium, Wahyu Muliady, "Automated Document Classification for News Article in Bahasa Indonesia based on Term Frequency Inverse Document Frequency (TF-IDF) Approach", 6th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, Indonesia, 2014.

[4] Seyyed Mohammad Hossein, Mohammad Shirzad Araghi, Morteza Mastery Farahani, "A Novel Text Mining Approach Based on TF-IDF and Support Vector Machine for News Classification", 2nd IEEE

[5] International Conference on Engineering and Technology (ICETECH), 17th& 18thMarch 2016, Coimbatore, TN, India.

[6] Akshita Bhandari, Ashutosh Gupta, Debasis Das, "Improvised apriori algorithm using frequent pattern tree for real time applications in data mining", International Conference on Information and Communication Technologies (ICICT 2014).

[7] Yiming Yang , Jan O. Pedersen, "A Comparitive study on Feature Selection in Text Categorization",2014.

[8] Mukesh A. Zaveri and Mita K. Dalal, "Automatic Text Classification: A Technical Review", International Journal of Computer Applications (0975 – 8887), Volume 28– No.2, August 2011.

[9] Youngjoong Ko and Jungyun Seo "Automatic Text Categorization by Unsupervised Learning" KOSEF under Grant No. 97-01-02 03-01-03.

[10] Kim S., Han K., Rim H., and Myaeng S. H., "Some effective techniques for naïve bayes text classification", IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 11, pp. 1457-1466., 2006.

[11] Huma Lodhi, Craig Saunders , John Shawe-Taylor, Nello Cristianini, Chris Watkins , "Text Classification using String Kernels" Journal of Machine Learning Research 2 (2002) 419-444, Published 2/02.

[12] Zhang W., Yoshida T., and Tang X."Text classification using multi-word features". In proceedings of the IEEE international conference on Systems, Man and Cybernetics, pp. 3519 – 3524. 2007.

[13] HaoLili., and HaoLizhu.. "Automatic identification of stopwords in Chinese text classification". In proceedings of the IEEE international conference on Computer Science and Software Engineering, pp. 718– 722. 2008.

[14] Porter M. F."An algorithm for suffix stripping". Program, 14 (3), pp. 130-137.1980.

[15] Liu T., Chen Z., Zhang B., Ma W., and Wu G. Improving text classification using local latent semanticindexing. In proceedings of the 4th IEEE international conference on Data Mining, pp. 162- 169. . 2004.

[16] M. M. Saad Missen, and M. Boughanem. "Using WordNet"s semantic relations for opinion detection in blogs". ECIR 2009, LNCS 5478, pp. 729-733, Springer-Verlag Berlin Heidelberg.

[17] Balahur A., and Montoyo A. "A feature dependent method for opinion mining and classification".In proceedings of the IEEE international conference on Natural Language Processing and Knowledge Engineering, pp. 1-7. 2009.

[18] Cho K. and Kim J. "Automatic Text Categorization on Hierarchical Category Structure by using ICF(Inverted Category Frequency)".Weighting. In Proceedings of KISS conferencepp.507-510. 1997.

[19]

[20] Tran, L. Q., Moon, C. W., Le, D. X., & Thoma, G. R.(2001). Web Page Downloading and lassification. Proceedings 14th IEEE Symposium on Computer-Based Medical Systems. CBMS 2001, 321–326. doi:10.1109/CBMS.2001.941739

[21] Wang, L., & Zhao, X. (2012). Improved KNN classification algorithms research in text categorization, i, 1848–1852.

[22] Wang, Y. U., & Wang, Z. (2007). A fast KNN algorithm for text categorization, (August), 19–22.

[23] Yan, Z. (2010). Combining KNN Algorithm and Other Classifiers, (1), 1–6.