International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 3
ISSN: 2454-4248
551 – 555

Text Classification Using Machine Learning Techniques: A Comparative Study

Neha Rani
Apex Institute of Technology
Chandigarh University, Mohali
nhrajput.24@gmail.com

Aanchal Sharma
Apex Institute of Technology
Chandigarh University, Mohali
aanchal08107@gmail.com

Dr. Sudhir Pathak
Apex Institute of Technology
Chandigarh University, Mohali
drspathak007@gmail.com

***Abstract*—** Text mining is drawing enormous attention in this era as there is a huge amount of text data getting generated and it is required very hardly to manage this data to grasp maximum benefit out of it. Text classification is an essential sub-part of text mining where the related text data is assigned to a particular predefined category. In our study, we discussed different classifier techniques which are popularly used in recent years. There is comparison between different classifiers like SVM, Naïve Bayes, Neural Networks etc. which is expressed in a tabular form in this paper.

Keywords- *Text classification, SVM, Naïve Bayes, Neural Networks, K-NN, Decision Tree*

*****

I. INTRODUCTION

Text mining is flourishingly increasing trend in the research field as there is huge amount of text data getting generated in different format. In this increasing trend of managing text data, text classification plays an essential role in the management of data. Text classification is an exercise of assigning different type of text data into pre-defined cluster or category. The main aim of a classifier is to classify the related text document to their respective category. When it comes about text data, a huge amount of data is being generated all over the world on daily basis. The text miners dive deep inside the sea of data-tombs & come with pearls out of it. This huge text documentary is very difficult to manage if the data is not categorized. Arranging text data into their related category somehow make the purpose of handling and using text data much easier. Assigning text documents into their specific pre defined category defines the term of text classification.
Classification of text data into different categories is not that short process. It follows a number of steps for its successful execution. The next is given the brief text classification process consisting of some essential steps:

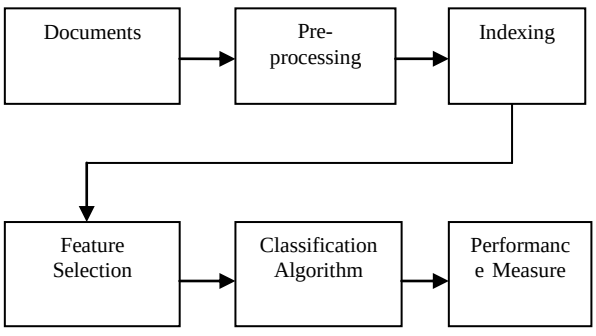

Figure 1: Text Classification Conceptual View

1. *Documents Collection*

This is very initial step of the process where text data is being collected in various forms for example .doc file, .txt file, html tags, web content etc.

2. *Pre-processing*

When the documents are collected successfully, further step is to convert every format to word document. There are some steps which take place in pre-processing, they are:

Tokenization: It is the process of converting each word in the document into different token. It splits a sentence into number of pieces called tokens.
Stop-words Removal: Stop words like connecters or prepositions are removed. For example: is, am, are, the, and, but, for, if etc.
Stemming: It is task of converting different words into the root word. For example:
Examination → Examine → Exam.

3. *Indexing*

To reduce the complexity of the documents, documentation representation is the one of the major technique of pre-processing where the full text version of the document is transformed to document vector. The vector space model is the most useful model for the representation of words.

4. *Feature Selection*

After pre-processing and indexing the significant pattern of written textual class, is highlight decision to unite vector space, which enhances the adaptability, execution and precision of a content classifier. The main idea of feature selection is to select subset of capacities from the specific archives.

5. *Classification*

Feature extracted files then passed to get classified text files. Different classification algorithms are used to automatically categorize text documents into their specific class, like Naïve Bayes classifier, Support-vector machine (SVM) classifier, neural networks, decision tree etc.

6. *Performance measure*

The running performance of the classier is analyzed on the basis of precision & recall values, F-measure and accuracy of the classier.
Precision = TP/ (TP+FP) (1)

 ISSN: 2454-4248
551 – 555

Recall = TP/ (TP+FN) (2)

F-measure = 2* (Precision*Recall)/ (Precision + Recall) (3)

Accuracy = (TP+TN)/ (TP+TN+FP+FN) (4)

Where TP is True positive, TN is True Negative, FP is False positive & FN is False negative.

II. LITERATURE SURVEY

Shugufta F. et al. (2017) [1] presented the work of document classification using support vector machine algorithm as a classifier. Their approach was to automatically categorize the text documents based on SVM machine learning algorithm. This is a supervised learning algorithm. They took reuter-corpora text documents as dataset and applied SVM method on it. They came up with 86.39% of accuracy in their work.

Ladislav L. et al. (2017) [10] focused on classification of text document by automatic labeling. They eliminated the step of preprocessing in order to escape from loss of information. They used deep neural networks in their work for text classification purpose. They compared two different networks one of which is standard multilayer perceptron and the other is convolutional network. Among which they concluded that convolutional networks given more efficient results. They worked on Czech newspaper text dataset. They have used BoW representation for FDNN and series of word indexes for CNN as the inputs.

June Ling O. et al. (2017) [12] came up with an idea of analyzing & determining the relevant news content on the basis of sentiment-based classifier. They took 250 English news documentaries which are in text form as their dataset. These text documents were labeled with different sentiments on the basis of which classification is done. They used knn approach in their work to classify the new content.

Roopesh S. et al. (2017) [14] proposed a work to classify multiclass documents in the case of text documents. They used Naïve Bayes classifier to solve the problem of text classification. The approach is first applied in linear and then in hierarchical manner to get the efficient results. They concluded that hierarchical approach is more effective as compared to linear one. They improved the accuracy & efficiency of the classifier.

Seyyed M. H. Dadgar et al. (2016) [3] aimed to classify news to different categories, using SVM classification technique. They used TF-IDF and SVM classifier in their work. They followed the process of text preprocessing, feature extraction on the basis of TF-IDF and at the end classification by using SVM classifer. They evaluated their results using two different data sets os BBC news and 20newsgroup dataset. The precision was 97.84% & 94.93% for both the datasets respectively.

Adel M Hamdan et al. (2016) [4] came up with a comparative study for text classification. They took Arabic language textual documents as their data set. They performed the task of text classification using three different techniques which are SVM classifier, Multilayer perceptron Neural Networks and Naïve Bayes classifier. They end up with the result that SVM technique is the best classifier for text classification while using Arabic text documents.

R Benkhelifa et al. (2016) [5] proposed the classification of lexicon based textual data of social media. The dataset is taken from facebook. They classified the facebook data into their respective categories using three different techniques of machine learning. The techniques used in this work are SVM, K-NN and Naïve-Bayes. They introduced a new approach of preprocessing which is based on stop words and internal slang, for improving the classification.

Sheetal Ashok R. S. et al (2016) [6] proposed a hybrid approach for SMS classification in order to detect spam and ham out of it. They used a hybrid approach of naïve bayes with apriori algorithm for their work. They concluded 60 to 68% of accuracy while testing 50 to 200 of untrained sms data set.

Sandeep Kaur et al. (2016) [8] proposed a method of classification of online news using neural networks which increased the accuracy of the classification up to 99% which is considered a very good result in case of text classification into multiple categories. They classified the news into four different categories.

Aaditya Jain et al. (2016) [15] introduces a modified form of maximum entropy based classier in combination with Naïve bayes classifier. This approach provided a lot of flexibility in their work. The combination of naïve bayes and maximum entropy classier is done by using operators which combine both the results linearly. Naïve bayes is used because of its simplicity on the other hand maximum entropy is used for its flexibility.

Omar Al-Momani et al. (2015) [13] gave a comparative study in the field of text classification. They used three techniques of text-classification and then compared the result for better performance. The algorithms used were K-NN, decision tree (C4.5) and rochhio classifier. They concluded that rocchio and K-NN worked well in case of Arabic data as compared to C4.5 decision tree.

P. Jotikabukkana et al. (2015) [17] proposed a technique used to classify the social media text by utilizing the initial keywords from well formed sources of data, like as online news. Authors used Term frequency inverse document frequency and WAM (Word article matrix). The experiment has been performed on Twitter message.

III. GENERIC CRETERIA OF CLASSIFYING TEXT DOCUMENTS

There are certain set of rules which are followed by every classifier for the purpose of text classification. It gives a generic strategy which is used to classify text data into pre-defined categories. Text classification process comprised of two sub-step processes:

Training Phase:

Every document is trained where every document is made to belong to a particular pre-defined category. These categories are defined on the basis of their content. It classifies the unlabelled documents to their suitable category.

Testing Phase:

 ISSN: 2454-4248
551 – 555

This is the second phase where the documents from the dataset, which are not used in training phase, are tested on the basis of the features extracted to get into a particular category.

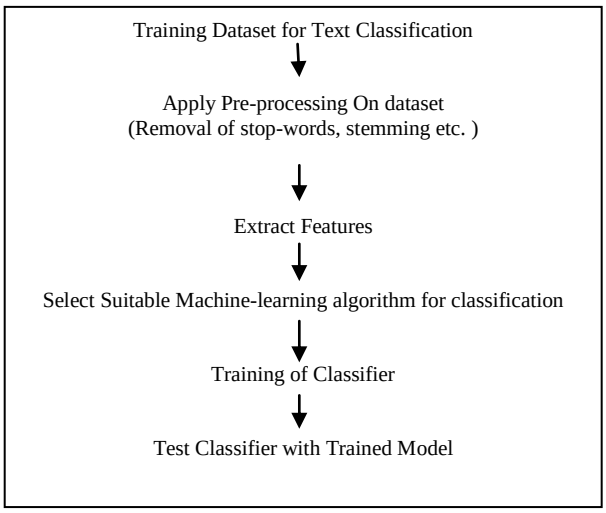

Figure 2: Strategy for Text classification

IV. Different Text Classifying Techniques

A. Naïve Bayes Classifier:

Naïve bayes is a probabilistic classifier which do not work on any single algorithm rather it work on a family of algorithm that all work on a single principle of classification. All of the extracted features using this classifier are independent of each other. The advantage of using this classifier is that it work good on both numeric as well as textual data and moreover it is easier to implement. The disadvantage of this classifier is that its performance gets poorer when the extracted features are correlated to each other.

B. Support vector machine (SVM):

SVM is an algorithm of supervised learning which is used for both classification and regression. In most of cases it is used as a classifier. It is used for many NLP tasks in text classification. Here each data item is plot as n-dimensional space where n represents no. of features extracted. SVM comes with a unique feature that it includes both types: positive & negative training sets. It represents each text-document as a vector where its dimensions are the no. of different keywords. But if the size of text document is large then there will be a number of dimensions in hyper-space which may increase computational cost of the process.

C. Artificial Neural Networks:

Artificial neural-networks work on the concept of human brain consisting neurons. It consists of a layered arrangement of neurons where the input vectors are converted into the some form of output. ANN is considered to be a good classifier because it can better handle multiple categories and work well on it. It supports fast testing phase.ANN is when combined with naïve bayes algorithm it comes up with a new idea which is called knowledge based neural networks and this is much efficient in handling noisy data.

D. Decision Tree:

Decision tree is another classifier algorithm which is widely used for the purpose of classification. It works on a series of some test questions & conditions applied on it. It is represented in a tree form of structure where the branches of tree represents 'weight' and each leaf is a different 'class'. Decision-tree is good in learning disjunction expressions and can handle noisy data. But training a decision-tree may act as an expensive process. If there is a mistake in very higher level then there will be flaws in whole sub-tree and whole structure may act as invalid.

E. K-nearest neighbor:

KNN is one of the easiest and simplest algorithms of machine learning. KNN classify text documents by calculating distance between the documents and neighbors those who are having similar classes are most probable of being from that class. But KNN is a slow learning algorithm and the complexity of calculating sample similarity is quite high.

V. Comparison Table Of Different Classifiers:

Table 1: Comparison table of text classifiers

Classifier	Research work and year	Approach used	Dataset used	Precision (average %age)	Recall (average %age)	Accuracy (average %age)	F1 (average %age)
SVM	[1] 2017	Support vector machine	Reuter dataset	-	-	78.3	-
	[2] 2016	SVM-RBF kernel	Unstructured data	95.92	96.35	97.6	-
	[3] 2016	TF-IDF and SVM classifier	BBC news	97.84	-	99.22	-
			20newsgroup	94.93	-	97.34	-
	[4] 2016	SVM classifier	Arabic dataset	77.8	77.4	-	77.5
	[5] 2016	SVM classifier	Facebook	77.9	77.2	77.28	77.15

 ISSN: 2454-4248
551 – 555

			dataset				
Naïve bayes	[5] 2016	NB classifier	Facebook dataset	78.63	77.65	77.25	77.66
	[4] 2016	NB classifier	Arabic dataset	75.45	75.98	-	75.62
	[6] 2016	Naïve bayes with apriori algorithm	SMS dataset	59	63	62.5	60.93
	[7] 2017	Naïve bayes classifier	20newsgroup dataset	-	-	94	-
Neural Networks	[10] 2017	FDNN	Czech text data	83.7	83.6	-	83.9
		CNN		86.4	82.8	-	84.7
	[11] 2017	Deep Neural Networks	WOS datasets	-	-	82.9	-
	[8] 2016	Neural Networks	News dataset	76.12	53.75	99.28	63
	[4] 2016	MLP- NN classifier	Arabic dataset	72.83	68.53	-	69.81
KNN	[12] 2017	Sentiment based	News content	37.3	37.7	-	30.5
		Polarity based	News content	78	73.4	-	66.8
	[5] 2016	K-nn classifier	Facebook dataset	63.43	56.43	56.42	53.63
	[13] 2016	K-nn classifier	Arabic dataset	26.44	26.55	-	26
Decision Tree	[13] 2016	C4.5 DT	Arabic dataset	67	64.1	-	65

VI. CONCLUSION

Text classification is an essential and highly required feature of time mining. In this study we have given a brief idea of popular text classification algorithm and compared some recent research works done using different classifiers and techniques. Most of the papers considered in our study were from 2016 and 2017 researches. We compare every technique on the basis of four parameters which are precision, recall, accuracy and f-measure. We concluded that the most frequently used techniques for text classification are SVM, Naïve Bayes and Neural networks, which give better results as compared to other techniques like K-nn or decision tree.

REFERENCES

[1] S Fatima, Dr. B. Srinivasu "Text Document categorization using support vector machine" International Research Journal of Engineering and Technology (IRJET) 2017.

[2] Nisha Gautam, Abhishek Bhardwaj "Novel Technique For Text Classification By SVM-RBF Kernel" IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555 Vol.6, No.2, Mar-April 2016.

[3] Seyyed Mohammad Hossein Dadgar et al "A Novel Text Mining Approach Based on TF-IDF and Support Vector Machine for News Classification" 2nd IEEE International Conference on Engineering and Technology (ICETECH), 17th& 18thMarch 2016.

[4] Adel Hamdan Mohammad et al "Arabic Text Categorization Using Support vector machine, Naïve Bayes and Neural Network" GSTF Journal on Computing (JOC) ,Volume 5, Issue 1; 2016 pp. 108-115.

[5] R. Benkhelifa, F. Z. Laallam "Facebook Posts Text Classification to Improve Information Filtering" 12th International Conference on Web Information Systems and Technologies (WEBIST 2016) - Volume 1, pages 202-207 .

[6] S A Rao Sable et al "SMS Classification Based on Naïve Bayes Classifier and Semi-supervised Learning" International Journal Of Innovations In Engineering Research And Technology [IJIERT] ISSN: 2394-3696 VOLUME 3, ISSUE 7, July-2016.

[7] Amey K. Shet Tilve, Surabhi N. Jain "Text Classification using Naïve Bayes, VSM and Pos Tagger" International Journal of Ethics in Engineering & Management Education (ISSN: 2348-4748, Volume 4, Issue 1, January 2017).

[8] Sandeep Kaur, Navdeep Kaur Khiva "Online news classification using Deep Learning Technique" International Research Journal of Engineering and Technology (IRJET) Volume: 03 Issue: 10 ,Oct -2016.

[9] Lea Vega and Andres Mendez-Vazquez "Dynamic Neural Networks for Text Classification" International Conference on Computational Intelligence and Applications 2016.

[10] Ladislav Lenc, Pavel Kra "Deep Neural Networks for Czech Multi-label Document Classification" arXiv:1701.03849v2 [cs.CL] 18 Jan 2017.

[11] K Kowsari, Donald E. Brown et al. "HDLTex: Hierarchical Deep Learning for Text Classification" arXiv:1709.08267v2 [cs.LG] 6Oct 2017.

[12] June Ling Ong Hui, Gan Keng Hoon et al. "Effects of Word Class and Text Position in Sentiment-based News Classification" 4th Information Systems International Conference 2017, ISICO 2017, 6-8 November 2017, Bali, Indonesia.

[13] Omar Al-Momani, Tariq Alwada et al. "Arabic Text Categorization using k-nearest neighbour, Decision Trees (C4.5) and Rocchio Classifier: A Comparative Study" International Journal of Current Engineering and Technology 2016.

ISSN: 2454-4248
551 – 555

[14] E Jadon, R Sharma et al. "Data Mining: Document Classification using Naive Bayes Classifier" International Journal of Computer Applications (0975 – 8887) Volume 167 – No.6, June 2017.

[15] Aaditya Jain, R. D. Mishra "Text Categorization: By Combining Naive Bayes And Modified Maximum Entropy Classifier" International Journal of Advances in Electronics and Computer Science, ISSN: 2393-283, Special Issue, Sep.-2016.

[16] Motaz K. Saad et al "Arabic Text Classification Using Decision Trees" Workshop on computer science and information technologies CSIT'2010, Moscow – Saint-Petersburg, Russia, 2010.

[17] P. Jotikabukkana, V. Sornlertlamvanich, O. Manabu and C. Haruechaiyasak, "Effectiveness of social media text classification by utilizing the online news category," 2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), Chonburi, 2015, pp. 1-5.