A Recapitulation of Different Text Classification Algorithms

Riya Bajpai

Abstract: ***Text classification process has gained a lot of importance in recent years and is still one of the most popular topics of discussion because of the presence of a huge outsized range of electronic documents from diverse resources. The text categorization process assigns predefined classes to documents. It finds noteworthy similarities in large textual data were interesting, hidden, previously unknown and extremely useful patterns and information can be discovered. Text classification helps in analysis of large textual data. Text mining intends facilitating customers extract informations from resources, and deals with operations such as retrieving, classifying ,cluster formation, mining of data, processing of natural language and techniques of machine learning together to classificate unalike patterns. Inside the process of content arrangement, terms gauging strategies configuration fitting loads to the offered terms to improvise content grouping execution. This paper overviews content order, the procedure of content grouping, distinctive term gauging techniques and correlations between various characterization calculations.***

Keywords: ***Naïve Bayes, SVM, Text Mining, Text Classification, Random Forest Classifier***

I. INTRODUCTION

With the technology growing at fast pace, the expanse of digital information engendered are massive and amalgamating the facts is very crucial. Classification of text is the process of assigning textual documents to single or multiple categories that are for selecting the correct class label for the input that's given. Each information is considered in separation from every single diverse info, and the arrangement of marks is characterized ahead of time. This enables clients to discover wanted data snappier via looking through just the important classifications and not the whole data space. The significance of content arrangement is considerably increasingly detectable when the data space is enormous, for example, the World Wide Web[10]. Be that as it may, such arrangement administrations are completed by human specialists, and they don't scale up well with the development pace of website pages on the Internet. To robotize and facilitate the characterization procedure, AI strategies have been presented. In a book grouping strategy dependent on AI, classifiers are assembled (prepared) with a lot of preparing records. The prepared classifiers can, accordingly, dole out records to their appropriate categories[2][3].

Ordering the news stories as per their substance is profoundly required as it can empower instinctual labelling articles for news repositories that are online and the aggregating news resources by theme (for example google news), just like given premise to new's suggestion frameworks. Content arrangement of new's stories is helpful in disclosing designs in articles and furthermore gives better acumen into the substance of the articles. This paper intends to exhibit an assessment of different famous classifiers , on different data-sets to think about classifier qualities.

II. METHODOLOGY OF TEXT CLASSIFICATION

The various phases of text classification process are : document collection , data pre-processing, indexing of features , filtering of features, using classifiers and measurement of performance [2][5][6].

A. Documents collection:

Assembling information from unalike sorts of configurations is done initially, for example, pdf, doc, HTML, and so on.

B. Pre-processing of data:

Information(data) mining is the way toward extracting concealed patterns in a tremendous dataset. Genuine information is frequently insufficient, conflicting and ailing in a specific conduct and is probably going to contain numerous mistakes. Series of stages data passes through are as follows : Removing the Stopping words: Stopping words are conventional words showing up in content which convey little hugeness, they serve just syntactic significance however don't uncover subject matter[3][4]. It is all around perceived among the compliance recovery specialists where practical words of english (eg. "a", "the", "that", "and", "has") are inefficient . These words have approximately no discrimination esteem because they show up on each English archive. Consequently, these words are not helpful in segregating records that have substance about differing subjects. Way toward wiping out the arrangement of such utilitarian words from collection of datas manufactured by extraction of words is a concept known as stopping words expulsion. So evacuation of the stopping words, that includes first making an rundown of stopping words being expelled, called the stopping words list[5]. Then the arrangement of words created by word extraction is then examined with the goal that each word showing up in the stop rundown is expelled.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author

Riya B, Department of Computer Science and Engineering, SRM Institute of Science and Technology, KTR Campus, Chennai, Tamilnadu, India. Email: riyabajpai_vi@srmuniv.edu.in.

Stem word formation is a methodology to diminish a word to its root form and is utilized generally for data recovery assignments to expand the review rate and give most applicable outcomes, for example, cheerful >happi.

C. Indexing of features :

The document is converted from complete content into vector of documents. A general practiced Documentation portrayal is known as a vector space model where record is described as word-vector. Typically, a corpus of reports is signified as word--by--word matrix of document. V.S.M. portrayal plan has detriments of itself. To give examples are: high dimensionality's portrayal, connection that fails with contiguous words and losing semantic relationship occurs amongst the words as a report to conquer such issues, to allocate appropriate loads(weights) to the terms terms-weighing technique can be exploited.

D. Weighing of terms :

In content portrayal, terms such as phrases, words, 0r another ordering units used to perceive the substance of a text. Every single term present inside report vector(document-vector) has to be connected with a weight, estimating the importance of this term and connotes whether the terms add-up any value to arrangement errand. This section centers around various conventional terms weighting techniques. Table I contains six diverse term gauging techniques dependent on managed and unaided strategies[1]. Unaided strategies commonly known as unsupervised learning don't have prior information about the category to which a document should belong to.

Frequency of terms and Inverse Document Frequency are 2 most significant contemplations in the customary feature algorithm of wieght calculation. Utilized generally in recovery of datas and healthier outcomes have been accomplished. TF alludes to the frequency of a term shows up in the text document. The highlights might be: word, expression, short forms of words. Beforehand, the document vectors were built fixated on TF. In various classes of records, be that as it may, there might be enormous contrasts between the T.F. of featured items. Therefore, the recurrence data is of significance to the content order . Inverse Document Frequency is the amount of the component thing conveyance in the set of document. The I.D.F. factor changes inversely with the number. Unaided strategies don't have known data on classification of training archives.

Table I: Different term weighing methods

Methods	Term weighing factors	Denoted by	Description
Supervised term weighing methods	Chi square	χ^2	Multiply *tf* by χ^2 funtion
	Information gain	*ig*	Multiply *tf* by *ig* funtion
	Odd ratio	*OR*	Multiply *tf* by OR funtion
	Relevance factor	*rf*	Multiply *tf* by *rf* funtion
Unsupervised term weighing methods	Term frequency	*Tf*	Number of times term occur in adocuments
	Inverse document frequency	*Idf*	Multiply *tf* by idf funtion

III. DIFFERENT ALGORITHM FOR TEXT CLASSIFYING

Data-base is lavish with shrouded data valuable for decision making intelligently. Order calculations should be utilized for model extraction portraying significant information classes. Reports classification into supervised, unsupervised and semi-supervised techniques. There are a few strategies implemented for group content, for example, K Nearest Neighbor (KNN), Artificial Neural Networks (ANN), Support Vector Machine (SVM), Naive Bayes Classifier, and Decision Trees. A few procedures are depicted as following:

A. Naïve Bayes Classifier

A celebrated & customary way to deal with content classification is Naïve Bayes (NB). It gets the hang of preparing models in priori likelihood given inconspicuous models. Essential idea is to ascertain the likelihood it arranges reports dependent on learning before given concealed instances of classes and probabilities which ascribe qualities have a place with classifications. The suspicion that properties are autonomous of one another underlies on this methodology. Despite the fact that this hypothesis disturbs the way that traits are reliant on one another, its exhibition is achievable. In text classification, for vectorization execution of Naïve Bayes is second rate when highlights are co-identified with one another. It is utilized prominently for content order, yet in addition for someother challenges in classification, as its learning is quick & straightforward.

B. K-Nearest Neighbour classifier

KNN is an arrangement algorithm wherein articles are assembled by casting a ballot a few marked preparing models with their littlest good ways from every other item. It was applied first by Massand in 1992 for arranging news articles though it was introduced in 1950 .Yang looked into twelve ways to deal with content order with one another, and made a decision about KNN that it is one of advisable methodologies, in 1999 . In 2002 Sebastiani assessed an uncomplicated & aggressive calculation with Support Vector Machine for executing content classification frameworks.K.N.N. Utilizes every highlights in registering separation & costs a lot time for characterizing objects is its drawback.

C. Neural Network

Neural system is a group associated information and yield units where every association consist of weight related with it. Neural system learning is likewise alluded to as connectionist learning because of associations amongst a few units. Neural system includes lengthy preparing procedure where requirement for number of parameters for grouping of classifications. Back engendering neural-system is a feed-forward, multilayer neural system comprising of an info , a concealed and a yield layer. In input layer and output layer the existent neurons are having predispositions, that aree association from units whose activation function capacity is constantly 1. The inclination terms are likewise go about like wieghts[7][8]. BPNN receives data sorces and yields from the net could be 1 /0/ bipolar (- 1, +1). In 1995, Wiener first applied this concept to content arrangement.

He utilized information Reuter 21578 for assessing the way to deal with content classification and presume that exhibition of back spread is superior to K.N.N. In 2002, consistent back proliferation to content categorization was applied by Ruiz and Srinivasan [8].

D. Decision Trees

Another classifier called Decision tree is a recursive segment of the case space. A decision tree is comprised of numerous nodes structuring an established tree, which means it is a coordinated tree with a node called "root". Root node has nil approaching edges. Every other node has atleast a single approaching or incoming edge. Node with active edges is alluded to as interior or test hub. Other nodes are recognized as leaves (otherwise called terminal or decison hubs). In a decision tree, each inside node isolates and separates case space into a minimum of two sub-spaces relating to specific discrete capacity of the information qualities values[9]. The decision tree grouping technique is better than other decision helping devices and has a few points of interest like its effortlessness in comprehension and deciphering, in any event, for non-master clients : decison trees are constructed utilizing avaricious pursuit calculations, likewise utilizing some heuristic that measures "debasement". Immaterial characteristics may severely influence the development of a decision tree. . Little varieties in the information can demonstrate that altogether different looking trees are produced.

E. Support Vector machines

SVM strategy applicable in grouping of non-linear(non straight) and linear information. SVM streamlines the loads of the inward results of preparing models and its info vector, called “Lagrange multipliers”, rather than those of its information vector, itself, as its learning procedure. It gives a consolidated portrayal model which has learned . Calculation utilizes non-linear mapping to change lower dimension training information into higher dimension and after that it scan for direct ideal isolating hyper plane. A significant exploration objective is SVM improve speed in preparing and testing to make practical alternative for enormous informational index[6]. It was initially applied in 1998 to content classification by Joachims . Joachims clarifies the S.V.M. in content order using contrasting of svm ,knn and naive bayes. Dr Drucker utilized S.V.M. for actualizing a spammed email separating framework & after that contrasted it and Naïve Bayes in executing the framework in 1999 . Finishing up this work after exhaustive perception that SVM was improved way in dealing with spammed emails separating ,than Naïve Bayes. Cristianini and Shawe-Taylor in 2000, showed an instance of applying Support Vector Machine to content order in a course book[4][5] . SVM can be used for classification of linearly separable classes and for non linearly separable classes .Linearly separable classes are those which can be distinguished easily whereas non linearly separable classes are those that can not be separated easily.For classification of non linearly separable data set we use Kernel function which converts an input in low dimensional feature space to output into high dimensional feature space.s

F. Random Forest Classifier:

This is a method which follows ensemble learning. Ensemble learning is the technique in which a master classifier is there which summarises the results of the base learners. Each Base learner takes the data set and perform training and testing. Each base learner can use different algorithms like one can use linear regression another can use naïve bayes another can use logistic regression etc to maintain diversity of their output. This is called heterogeneous ensemble learning. If all learners use same algorithms then diversity might reduce therefore to bring out diversity we can use different data set for the base learners during training time. So there are two aspects in ensemble learning : either have different algorithms for same dataset for each base learner or have different training data set and same algorithm. Base learners can also be called as weak learners and we can consider them as models or classifiers. The result of these weak classifiers are combined and summarised by a strong classifier. The predictive power and accuracy of the strong classifier is way more than any of the weak classifier. Its precision rate is high and error rate is lower than the weak learners. Here we are not dependent on just one classifier / algorithm , that’s why the output given by strong classifier is high.It is done in two ways : bagging (bootstrap aggregation) and boosting. In Random Forest classifier a lot of different decision trees are generated and their outputs are summarised and result is given .

IV. RESULT

The accuracy of the various text classification algorithms were tested on the Reuters-21578 dataset and is summarized in the table II. This dataset is a collection of documents that appeared in 1987 in reuters newswire.

Table II: Results of accuracy test done on Reuters-21578.

Reference	KNN	Naïve bayes	SVM	Neural Network
[11]	85.0	71.5	85.9	82.0
[12]	86.3	73.4	86.3	-
[13]	82.3	72.0	86.0	-

V. CONCLUSION

This paper overviews content arrangement calculations. This study fixated on the overarching writing and investigated the archives portrayal and an examination of highlight choice systems and characterization calculations.Term-weighing is the most imperative step in developing a book-classifier. Diverse term-weighing draws near, including supervised & unsupervised term weighting, is seriously explored by past examinations. This paper likewise talks about concise prelude to the different content portrayal plans. The current grouping strategies are thought about dependent on advantages and disadvantages. All the different algorithms have their own advantages and disadvantages. After the discussion above its comprehended that just a single classifier can be generalised as the best as various calculations accomplish diversely relying upon information accumulation.

.

REFERENCES

1. 1.Kohei watanabe "Newsmap-A semi-supervised approach to geographical news classification "_– *Digital journalism volime 6, 2018 issue 3*
2. Anuradha Patra, Divakar Singh- "A Survey Report on Text Classification with Different Term Weighing Methods and Comparison between Classification Algorithms"- *International Journal of Computer Applications (0975 – 8887) Volume 75– No.7, August 2013.*
3. Mazhar Iqbal Rana, Shehzad Khalid, Muhammad Usman Akbar –" News Classification Based On Their Headlines: A Review_" - *17th IEEE INMIC 2014 Karachi*.
4. Lilima Pradhan, Neha Ayushi Taneja, Charu Dixit ,Monika Suhag, "Comparison of Text Classifiers on News Articles "- *IRJET, Volume: 04 Issue: 03 , Mar -2017*
5. Zach CHASE Nicolas GENAIN Orren KARNIOL-TAMBOUR , "Learning Multi-Label Topic Classification of News Articles"
6. 6 .Inoshika Dilrukshi, Kasun De Zoysa,Amitha Caldera,"*Twitter News Classification Using SVM* ", The 8th International Conference on Computer Science & Education (ICCSE 2013) April 26-28, 2013. Colombo, Sri Lanka
7. Ashish Agarwal, Ankita Mandal, Matthias Schaeld, Fangzheng Ji, Jihao Zhang, Yiqi Sun, "*Good, Neutral or Bad news classification*", NewsIR'19 Workshop at SIGIR, Paris, France, 25-July-2019,published at http://ceur-ws.org
8. 8.Mingyang Jiang,Yanchun Liang,Xiaoyue Feng,Xiaojing Fan,Zhili Pei,Yu Xue,Renchu Guan," *Text classification based on deep belief network and softmax regression " S*pringer - Neural Computing and applications ,January 2018, Volume 29, issue 1, pp 61–70
9. 9 . Sushilkumar Kalmegh ," *Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News* ", International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 2, February 2015.
10. Chee-Hong Chan, Aixin Sun,Ee-Peng Lim "*Automated Online News Classifcation with Personalization*", 4th International Conference of Asian Digital Library (ICADL2001), Pages 320--329, Bangalore, India, December, 2001.
11. Yang Yi-Ming, Liu Xin." A re-examination of text categorization methods " Proceeding of ACM SIGIR Conference on Research and development in Information Retrieval, Berkeley , California ,1999.42
12. Weiss S M, Damerau CFJ. "maximizing Text-mining performance, IEEE Intelligent systems, new York, IEEE Press , 1999,121
13. 13.Joachins T. Text Categorization with Support Vector Machine: learning with many relevant features, machine learning ,1998,137-142,11398.

AUTHORS PROFILE

Riya B. ,is currently pursuing Master's degree in Computer Science and Engineering in SRM Institute of Science and Technology Kattankulathur, Chennai, India. Pursued her Bachelor's degree in SRM Institute of science and technology, Chennai, India. Her area of interest is machine learning, Image Processing, Database, Data Structures, and Network Security and Cryptography