ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

AN EMPIRICAL STUDY: TEXT CLASSIFICATION ALGORITHMS

Jekkala Chandra Sekhar[*1] & Dr. K. Subba Rao[2]
[*1]Assistant Professor, BVRIT, Narsapur, Medak, India
[2]Professor[1,] BVRIT, Narsapur, Medak, India

DOI: 10.5281/zenodo.1165796

ABSTRACT

Now a days, it is very risky to filter the unwanted data in social networks. Data is generally in the form of text majority in the social networks. There are different algorithms available for classify the text in the social networks. Machine Learning based algorithms can be applied to text for filtering unwanted text in Social Networks very accurately than existing algorithms. Machine Learning based Algorithms provides best text categorization and labelling the text through efficient feature selection. Text Categorization is the important step in machine learning algorithms. In this paper, a review on various machine learning text classification techniques has been presented. Different supervised classification techniques of text mining have been discussed in this paper.

KEYWORDS: Text Mining, Machine Learning based algorithms, unwanted data, Social Networks.

I. INTRODUCTION

With the fast advancement of data innovation and the broad use of system, the Internet has step by step turned into a crucial piece of individuals' life. Pages and informal organization destinations will produce a lot of unstructured content information, for example, sites, discussion posts, specialized documentation, and so on. These information demonstrating individuals' conduct and thought naturally, contains a ton of data, which is to a great degree hard to manage as a result of the immense number and different structures. Be that as it may, the request of breaking down content information is rising. Consequently, how to secure the data individuals require from extensive quantities of unstructured content information turns into the exploration hotspot in the field of information mining and data. Content mining appeared

Text mining [1], otherwise called information disclosure in literary database (KDT) [2] or content information mining [3], of which new intriguing learning is made, is characterized as the way toward removing already obscure, reasonable, potential and functional examples or information from the gathering of gigantic and unstructured content information or corpus.

As a branch of text mining, text mining is accepted to have higher business esteem than text mining in light of the fact that 80% of an organization's data is contained in text reports [4].

Be that as it may, text mining is more perplexing as the unstructured content information. Text mining is an exhaustive research territory, which includes in the fields of manmade brainpower, machine learning, scientific insights, database framework, et cetera.

This paper presents the historical backdrop of text mining and research status. At that point some broad models are portrayed in Section III. The fourth part is to arrange text mining work as indicated by application. At last, it is summary

II. TEXT MINING CLASSIFICATION SYSTEM

Text classification [5] sorts documents into a settled number of predefined classifications. The documents can be different, exceptional or may not fit into a class by any means. All things considered, dealing with countless can get to be distinctly entangled. In this way, a text classifier puts these documents into gatherings which are important to their substance and makes it less demanding to sort them when a scan for a particular report is done. The arrangement of classifications for the documents is called Controlled Vocabulary [6]. A decent similarity would be that of an understudy sorting an arrangement of endorsements, travel permit photocopies, exam check sheets and a couple frames into various organizers and marking every envelope as indicated by its substance for simplicity of recovery later. A decent text classifier however, would work productively for expansive preparing sets with a few elements. Include Selection frames a vital piece of any arrangement errand and it is particularly critical on account of text classification in view of the high dimensionality and nearness of commotion of elements, so it is important to choose just the most basic components. A typical stride of highlight choice is stop-word evacuation and stemming. [7] Stop-word evacuation includes erasing words which are normal and don't make a big deal about a distinction for grouping. Stemming includes lessening words which are bent to their ―stem, the root word from which they determine. As per Basu [8], Text arrangement requires, as a premise, the distinguishing proof of elements inside the records that can be utilized to separate among the reports and partner them to individual classes

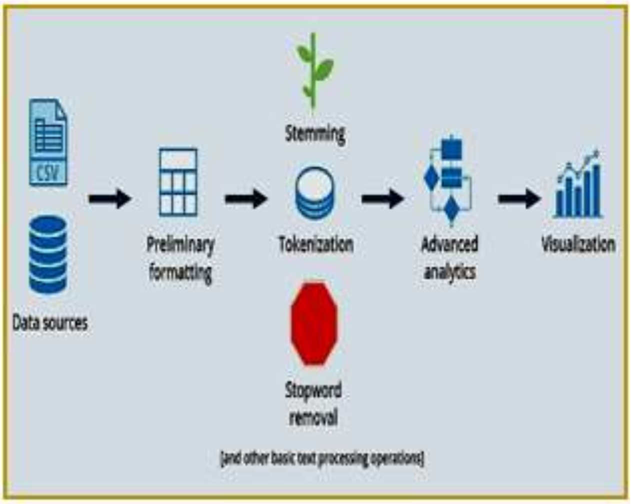

Fig 1: Text Mining Classification Process

III. TEXT MINING CLASSIFICATION ALGORITHMS

A. Naive Bayes According to Patra [8], Naive Bayes first learns preparing cases in priori likelihood when given concealed cases. The components are thought to be autonomous importance the nearness of one element does not influence the nearness of another element. Due to this supposition that qualities are free of each different underlies on this approach, it is called 'Naive'. Despite the fact that this hypothesis damages the way that traits are subject to each other, its execution is doable. It is the most generally utilized classifier in light of its straightforwardness and furthermore on the grounds that it is ceaselessly adjusting in the event that a client distinguishes an inaccurately characterized case, in this way enhancing its productivity. NB depends on the Bayes administer of restrictive likelihood [9] given by formula (1). h is the hypothesis and x is the attribute.

i. $$P(hi/xi) = \frac{P(xi/h)*P(hi)}{P(xi)} \qquad \text{----------------(1)}$$

B. C4.5 could be a modification of the ID3 algorithmic rule that focuses on making a choice tree, employing a fastened set of attributes, to classify a coaching example into a set set of categories as expressed by Macskassy et al [10]. C4.5 is AN entropy based mostly algorithmic rule. it's a wide used call tree learning algorithmic rule. At each step, if the remaining instances all belong to identical category, it predicts that exact category, otherwise, it selects the attribute with the very best data gain and creates a choice supported that attribute to separate the coaching set into one or 2 subsets. If the feature is distinct then the coaching set is split into one set supported its distinct worth. within the case of continuous options, 2 subsets are created on the idea of threshold comparison. The on top of steps area unit recurrent recursively until all the nodes area unit final, or till the edge limit is met. the edge limit are specific by the user. Once the choice tree is made, C4.5 prunes the tree so as to avoid over fitting, once more supported a setting specific by the user.

C. Support Vector Machine SVMs are efficient binary classifiers that is based on structural risk minimization, meaning that it describes a general model of capacity control [11] and provides a tradeoff between hypothesis space complexity (the VC dimension of approximating functions) and the quality of fitting the training data (empirical error). They are learning machines which are based on statistical learning theory. Any SVM would try to maximize the boundary between the positive and negative examples in a dataset. SVMs non-linearly map their n-dimensional input space into a higher-dimensional feature space. Using this high-dimensional feature space a linear classifier is then constructed with the help of quadratic programming, though this step can potentially be very costly. So to optimize this step, SVMs make use of different kernel methods which might improve the computation of inner numerical products.

The Datasets

Diabetes

This dataset consists of 768 instances with 9 attributes and the training examples are taken from a larger database which recorded the biological statistics of women, all around 21 years of age, and of Pima Indian origin. Given these training examples to a text classifier, the classifier will predict whether the patient has been tested positive/negative with diabetes mellitus based on the criteria set forth by the World Health Organization that a reading of 200 mg/dl, 2 hours post lunch shows signs of diabetes.

Calories

The dataset consists of 40 food items and 4 attributes. Some of them claim to be ―lite, ―low-fat, ―no-fat, or ―healthy foods. These foods are classified based on their distribution i.e., nationally advertised, regionally distributed or locally prepared. Using the above three algorithms, the dataset is trained and correctly/incorrectly classified instances are determined by the Wekatool.

Results and Evaluation

The following below two table describes the Correctly classified instance percentage after training for Diabetes and Correctly classified instance percentage after training for Calories. For this, WEKA Tool can be considered.

Table I. Diabetes Dataset Results

% Split of Training Set	Algorithm		
	Naive B	*C4.5*	*SVM*
At 66% (261 instances)	77.01	76.24	79.31
At 90% (77 instances)	77.9	75.32	80.52
At 33% (515 instances)	73.98	70.29	75.73
Precision(Weighted Avg.)	0.767	0.756	0.787
Recall (Weighted Avg.)	0.77	0.762	0.793

Table I. Calories Dataset Results

% Split of Training Set	Algorithm		
	Naive Bayes	C4.5	**SVM**
At 66% (14 instances)	78.57	78.57	**71.52**
At 90% (4 instances)	100	75	**75**
At 33% (27 instances)	81.48	85.18	**66.67**
Precision(Weighted Avg.)	0.844	0.802	**0.81**
Recall (Weighted Avg.)	**0.786**	**0.786**	**0.714**

In the main dataset, SVM beats the staying two classifiers. In the mean time, the execution of SVM is more terrible with the second dataset. Both the datasets were part into preparing set and testing set. When we select a 66% split, it infers that 66% of the dataset is preparing information, while the rest of the cases are trying cases. It is watched that SVM performs ineffectively when the quantity of properties is less which is clear in the Calories dataset.

Algorithms Comparison Using Chart

The following ttwo charts will be described with clear explanation about the comparison of text classification algoriconsidering Y-axis as % Spilt of Training set and X-axis as Algorithms like SVM,C4.5,Naïve Bayes

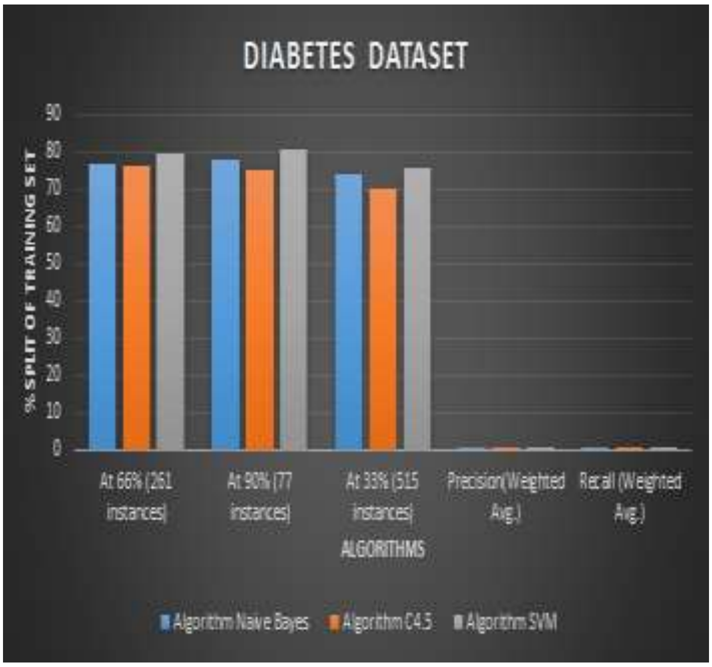

Chart 1: Diabetes Dataset Result Analysis

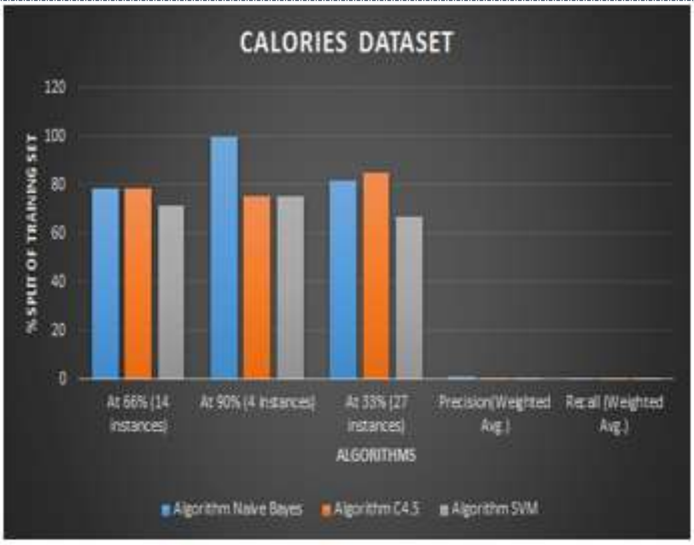

Chart 2: Calories Dataset Result Analysis

IV. CONCLUSION AND FUTURE SCOPE

In a nutshell, text classification is an essential zone of research for applications requiring steady need to mark archives and sort out information for use in further research. Utilization of Naive Bayes, C4.5 and Support Vector Machine on two or three datasets with changing preparing illustrations helped us think about execution of each of these classifiers. Support Vector Machine beats the staying two classifiers and turns out to be the best among the three. SVM may have a few inconveniences however that can be enhanced by joining SVM with different calculations. SVM has turned out to be powerful when the correct parameters are picked generally the outcomes are not ideal. Sudheer et al [12] have proposed consolidating SVM with Particle Swarm Optimization for tuning the parameters. Another approach proposed by Phung et al [13] is to isolate the Quadratic Programming issue into littler sub-issues which will lessen calculation time for expansive datasets

V. REFERENCES

[1] Tan, Ah Hwee, et al. "Text Mining: The state of the art and the challenges." Proceedings of the Pakdd Workshop on Knowledge Disocovery from Advanced Databases(2000):65--70.

[2] Feldman, Ronen, and I. Dagan. "Knowledge Discovery in Textual Databases (KDT)." In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95(1995):112--117.

[3] Hearst, Marti A. "Untangling text data mining." University of Maryland1999:3--10.

[4] S. Grimes. "Unstructured data and the 80 percent rule." Carabridge Bridgepoints, 2008.

[5] Jiménez, S. Text Classification and Clustering with WEKA, 2014.

[6] Wilcox, A. and Hripcsak, G. Classification algorithms applied to narrative reports. p.455, 1999.

[7] Pandey, U. and Chakraverty, S. A Review of Text Classification Approaches for E-mail Management. IACSIT International Journal of Engineering and Technology, 3(2), 2011.

[8] Patra, A. and Singh, D.(2013). A Survey Report on Text Classification with Different Term Weighing Methods and Comparison between Classification Algorithms. International Journal of Computer Applications Volume 75–No.7, August 2013 pp.14-18

[9] Dunham, M. (2003). Data mining introductory and advanced topics.1st ed. Upper Saddle River, N.J.: Prentice Hall/Pearson Education.9

[10] .Macskassy, S., Hirsh, H., Banerjee, A. and Dayanik, A. Converting numerical classification into text classification. Artificial Intelligence, 143(1), pp.51—77, 2003.

[11] Sewell, M. (2014). Structural Risk Minimization. [online] Svms.org.Available at: http://www.svms.org/srm/ [Accessed 4 Sep. 2014]

[12] Sudheer, C., Maheswaran, R., Panigrahi, B. and Mathur, S. (2014). A hybrid SVM-PSO model for forecasting monthly streamflow. Neural Computing and Applications, 24(6), pp.1381--1389.

[13] Phung, S., Nguyen, G. and Bouzerdoum, A. (2010). Efficient SVM training with reduced weighted samples.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

CITE AN ARTICLE

Sekhar, J. C., & Rao, K. S. (n.d.). AN EMPIRICAL STUDY: TEXT CLASSIFICATION ALGORITHMS. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY*, *7*(2), 175-180.