Text Categorization and Text Mining Using Different Types of Classifiers

Stalin Jose J
Research Scholar, Bharathiar University, Department of Computer Science, Coimbatore, Tamil Nadu, India.

DR.P.Suresh
Head of the Department, Department of Computer Science, Salem Sowdeswari College, Salem, Tamil Nadu, India.

Abstract – In recent times, text data mining has gained more attention in which text categorization is one of the most interesting fields. It has gained more popularity because of the rapid growth of textual documents. These documents are associated with selective large number classes such as medical, sports, Olympic Games and so on. This text categorization can provide several opportunities to develop multi-label learning techniques which particularly for text based information. Text data mining is the process of finding helpful learning patterns from the text based information, which is one of the key factors used by automatic text categorization. It is achieved by creating new machine learning techniques. In any case, the ML framework produces less expressivity. With the help of Train- Test scenario, this ML framework is deployed. In the event that the current framework is discovered insufficient, then the Train-Test-Retrain is produced that is tedious and time consuming process. In this paper, we have compared the performance of three classifiers such as Naive Bayes, Decision Tree J48, and Multi-Layer Perceptron (MLP) for a text mining dataset. The performance of all three classifiers is obtained through simulation and the experimental results are given. From the obtained results, it is known that the decision Tree J48 classifier provides better performance compared to other two classifiers.

Index Terms – Text data mining, text categorization, Train- Test scenario, Naive Bayes, Decision Tree J48, and Multi-Layer Perceptron (MLP).

1. INTRODUCTION

Text mining is the process of extracting high quality data from text, which is otherwise referred as text data mining, generally equal to content analysis. High quality data is commonly inferred through the conceiving of patterns and techniques through means like statistical pattern learning. Generally, the text data mining includes the way toward organizing the information content (commonly parsing, with the expansion of some inferred phonetic features and the expulsion of others, and resulting inclusion into a dataset), extracting patterns inside the organized information, lastly assessment and understanding of the result. In text data mining, the 'high quality' generally referred to some integration of applicability, uniqueness, and attractiveness. Conventional text data mining jobs incorporate information extraction, text summarization, text clustering, text classification, generation of granular taxonomies, relation modeling of entities (that is, learning relations among named entities), and sentiment analytics. Text analysis includes data recovery, lexical investigation to contemplate word frequency disseminations, pattern acknowledgment, labeling/annotation, extraction of information, data mining approaches incorporating connection and link investigation, representation, and predictive investigation. The major objective is, basically, to transform text into useful information for investigation, by means of use of natural language processing (NLP) and strategies for analysis.

1.1 Text analytics

Text analysis portrays a group of linguistic, measurable, and machine learning methods which frame and structure the data substance of textual data sources to obtain commercial knowledge, exploratory information investigation, research, or analysis [1]. This term 'text analysis' is generally compatible with text data mining; actually, Ronen Feldman altered a 2000 depiction of 'text data mining' [2] in the year of 2004 to depict 'text analysis' [3]. The last term is presently utilized more commonly in business context while the term 'text data mining' is utilized in few of the former application fields in 1980s,[4] prominently government intelligence and life-sciences research. Also, the term text analysis portrays that utilization of text analysis to react to commercial issues, regardless of whether freely or in conjunction with inquiry and examination of handled, numerical information. It is an adage that 80 % of commercial related information begins in unstructured format, essentially text [5]. These strategies and procedures detect and current knowledge factors, commercial principles, and connections, that is, generally secured in text based format, impervious to mechanized operating.

2. RELATED WORK

Decision tree techniques [8] recreate the non-automatic request of the training data in the kind of a tree based structure, where a node indicates the inquiries and a leaf demonstrates the specific class of dataset. The negative mark in decision tree methodology is set to as 'over-fitting'. It is basic and advantageous to utilize. Bayesian methodologies are classified

into Naïve and Non-Naïve Bayesian methodologies. The naïve methodologies work in two classes in particular, multivariate and multinomial classes. Both the classes execute on terms dispersion in the data documents [10]. An N-gram is defined as an endless sequence [11] which comprises of n characters of lengthy part of a substance. The most progressive N-grams are preserved. This method creates a basic number of parts appeared differently in relation to the text analysis by considering the word separators and punctuation marks, besides, it is to a great degree tolerant to the spelling errors and it stays faultless to all progressions models on a discrete letters (Chinese language, DNA orders…). Its accomplishment in dialect recognition is gone to its operation in content categorization [12].

Essentially, there are two sorts of vector oriented strategies in particular, Centroid technique and Support Vector Machines (SVMs) [13]. Centroid technique is the easiest technique. In the learning framework, a vector capacity for centroid is evaluated. The centroid technique is effectively utilized as a part of short databases [14]. Other framework, Support Vector Machine (SVM) is ascertained for characterizing the novel data documents. Chaitrali S. Dangare [16] have examined in coronary diseases prediction frameworks which consolidate a broad utilization of features to recognize the similitude scores of patient possessing coronary diseases or not.

D. Lavanya et al [17] proposed a hybrid technique for anticipating breast cancer diseases. They used the attributes of CART, which incorporates determination and packing strategies for measuring the execution performance. They identified the diseases at an inconvenient stage with exact outputs. Lior Rokach [18] was reviewed a top-down technique in the bits of knowledge of decision tree technique. They portrayed a new technique for partitioning criteria and pruning methods.

An improved anticipation technique was examined at [19] educational databases for classifying the vocation choices for end users. As per the behavior of students, the execution graph was assessed and proposed them an enhancement path for scholastic with utilization of Rapid Miner which is a data mining tool. The characterization framework was additionally tried in blood donar frameworks [20]. They utilized the approach of CART decision tree technique and established by utilizing WEKA tool.

3. CLASSIFIERS

There are three types of classifiers are taken to analyze the text dataset are,

- Naïve Bayes
- Decision Tree J48
- Multi-Layer Perceptron (MLP)

3.1 Naive Bayes

Naive Bayes is one of the basic approaches to build classifiers: frames that allocate labels for class to issue cases, demonstrated as vectors of attribute measures, where the labels are extracted from some limited set. It is not an individual technique to train these classifiers, however, a group of techniques as per a typical rile: entire naïve Bayes classifiers suppose that the measure of a specific attribute is autonomous of the measure of some other attributes, provided the class variable. For instance, a fruit might be thought to be an apple on the off chance that it is red, round, and around 10 cm in width. The naïve Bayes classifier takes every one of these attributes for contributing autonomously to the possibility that this organic product is an apple, without regarding any probable correlations among the shading, roundness, and diameter attributes. For a few sorts of probability frameworks, naive Bayes classifiers can be effectively trained in a context of supervised learning.

In numerous real time applications, parameter approximation for naive Bayes frameworks utilizes the strategy for maximum likelihood. Alternatively, one can operate with the naive Bayes framework without taking Bayesian likelihood or utilizing any Bayesian techniques. Regardless of their naive model and evidently distorted presumptions, naive Bayes classifiers have operated well in numerous unpredictable real time circumstances. In 2004, an investigation of the Bayesian characterization issue demonstrated that there are sound hypothetical purposes behind the evidently improbable viability of naive Bayes classifiers [5]. In 2006, a comprehension based comparison with other classification techniques demonstrated that Bayes classification is exceeded by different methodologies, for example, random forest [6].

3.2 Decision Tree J48

As the name itself implies J48 is the best known decision tree based classification technique. Initially it classifies the images as per the attributes and forms tree structure respectively. The tree hierarchy is explained in an understandable way. The Decision Tree J48 is extended from ID3 and it is performed mainly for its simple methodology in identifying the hidden pixels in the images. Under classification the images were arranged in a leaf structure and get pruned. By labeling these pixels were grouped and on each pixel the information's were extracted then tested. From resultant pixel the perfect one is selected and these classifiers are appreciated for handling both discrete and continuous values.

Pesodocode for j48 Check for the above base cases.

1. For each attribute a, find the normalized information gain ratio from splitting on a.
2. Let a_best be the attribute with the highest normalized information gain.

3. Create a decision node that splits on a_best.
4. Recur on the sublists obtained by splitting on a_best, and add those nodes as children of node.

3.3 Multi-Layer Perceptron (MLP)

The Multi-Layer Perceptron (MLP) is a feed forward ANN technique, here by mapping the classification were done on the input images. The mapping is done on the features of the training and testing dataset. Here the mapping is done by applying back propagation algorithm. By means of that the MLP constructs nodes as a directed graph and then connected to each other. Each individual node in the graph is provided with non-linear activation function. Additionally the datasets of MLP were trained by supervised learning techniques which are also helpful in classifying non-linear data's. It operates fitness function in a stochastic manner for solving the complexities.

4. EXPERIMENTAL RESULTS

In this paper, 'Letter dataset' is taken for text analysis and text categorization. The classification is conducted in Weka tool where three classifiers are taken to perform classification such as Naive Bayes, Decision Tree J48, and Multi-Layer Perceptron (MLP). In this section, the results of each classifier are given in a detailed manner.

Time taken to build model: 1.49 seconds

4.1 Results of Naïve Bays classifier

=== Evaluation on training set ===

Correctly Classified Instances %	12881	64.405
Incorrectly Classified Instances %	7119	35.595
Kappa statistic	0.6298	
Mean absolute error	0.032	
Root mean squared error	0.1383	
Relative absolute error	43.3058 %	
Root relative squared error	71.9279 %	
Total Number of Instances	20000	

Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC	Area	Class
0.873	0.007	0.841	0.873	0.857	0.968		A
0.719	0.03	0.485	0.719	0.58	0.969		B
0.75	0.009	0.771	0.75	0.76	0.971		C
0.703	0.019	0.614	0.703	0.655	0.967		D
0.357	0.009	0.608	0.357	0.45	0.943		E
0.73	0.012	0.703	0.73	0.716	0.958		F
0.545	0.018	0.546	0.545	0.545	0.946		G
0.312	0.01	0.545	0.312	0.397	0.892		H
0.767	0.027	0.528	0.767	0.626	0.944		I
0.718	0.006	0.825	0.718	0.767	0.947		J
0.455	0.018	0.491	0.455	0.472	0.952		K
0.761	0.001	0.962	0.761	0.85	0.931		L
0.895	0.018	0.666	0.895	0.764	0.989		M
0.696	0.004	0.865	0.696	0.771	0.982		N
0.733	0.033	0.466	0.733	0.57	0.96		O

0.743	0.004	0.875	0.743	0.804	0.974	P
0.539	0.016	0.581	0.539	0.559	0.956	Q
0.67	0.017	0.61	0.67	0.639	0.979	R
0.303	0.025	0.322	0.303	0.313	0.914	S
0.702	0.013	0.699	0.702	0.701	0.956	T
0.723	0.005	0.856	0.723	0.784	0.962	U
0.815	0.016	0.672	0.815	0.737	0.971	V
0.799	0.014	0.692	0.799	0.742	0.99	W
0.456	0.026	0.415	0.456	0.434	0.95	X
0.337	0.006	0.703	0.337	0.456	0.964	Y
0.61	0.007	0.774	0.61	0.682	0.971	Z
Weighted Avg.	0.644	0.014	0.66	0.644	0.641	0.958

4.2 Results of Decision Tree J48 classifier

Correctly Classified Instances	19269	96.345 %
Incorrectly Classified Instances	731	3.655 %
Kappa statistic	0.962	
Mean absolute error	0.0044	
Root mean squared error	0.0471	
Relative absolute error	5.9985 %	
Root relative squared error	24.4918 %	
Total Number of Instances	20000	

Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.986	0.001	0.982	0.986	0.984	1	A
0.949	0.003	0.931	0.949	0.94	0.999	B
0.961	0.001	0.966	0.961	0.963	0.999	C
0.973	0.002	0.949	0.973	0.961	0.999	D
0.969	0.002	0.953	0.969	0.961	1	E
0.95	0.002	0.956	0.95	0.953	0.999	F
0.965	0.003	0.934	0.965	0.949	0.999	G
0.935	0.002	0.942	0.935	0.938	0.999	H

	0.98	0.001	0.978	0.98	0.979	1	I
	0.964	0.002	0.959	0.964	0.961	1	J
	0.955	0.002	0.95	0.955	0.953	0.999	K
	0.967	0.001	0.987	0.967	0.977	1	L
	0.975	0.001	0.978	0.975	0.977	1	M
	0.977	0.001	0.971	0.977	0.974	1	N
	0.954	0.002	0.948	0.954	0.951	0.999	O
	0.958	0.002	0.952	0.958	0.955	0.999	P
	0.963	0.002	0.954	0.963	0.959	1	Q
	0.949	0.002	0.959	0.949	0.954	0.999	R
	0.944	0.002	0.951	0.944	0.948	0.999	S
	0.969	0.001	0.971	0.969	0.97	1	T
	0.969	0.001	0.984	0.969	0.976	1	U
	0.962	0.001	0.981	0.962	0.972	1	V
	0.973	0.001	0.985	0.973	0.979	1	W
	0.967	0.001	0.973	0.967	0.97	1	X
	0.966	0.001	0.973	0.966	0.969	1	Y
	0.969	0.001	0.985	0.969	0.977	1	Z
Weighted Avg.	0.963	0.001	0.964	0.963	0.963	1	

4.3 Results of Multilayer perceptron

Time taken to build model: 344.96 seconds

=== Evaluation on training set ===

Correctly Classified Instances	16491	82.455 %
Incorrectly Classified Instances	3509	17.545 %
Kappa statistic	0.8175	
Mean absolute error	0.0153	
Root mean squared error	0.1093	
Relative absolute error	20.6997 %	
Root relative squared error	56.8155 %	
Total Number of Instances	20000	

Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.883	0.005	0.885	0.883	0.884	0.965	A
	0.856	0.018	0.656	0.856	0.743	0.956	B
	0.887	0.008	0.802	0.887	0.843	0.986	C
	0.887	0.014	0.725	0.887	0.798	0.972	D
	0.823	0.01	0.763	0.823	0.792	0.945	E
	0.827	0.008	0.801	0.827	0.814	0.976	F
	0.719	0.004	0.89	0.719	0.795	0.887	G
	0.26	0	0.995	0.26	0.413	0.681	H
	0.792	0	0.985	0.792	0.878	0.882	I
	0.839	0.005	0.874	0.839	0.857	0.955	J
	0.855	0.012	0.735	0.855	0.79	0.98	K
	0.846	0.003	0.916	0.846	0.88	0.937	L
	0.936	0.009	0.816	0.936	0.872	0.986	M
	0.831	0.003	0.926	0.831	0.876	0.968	N
	0.795	0.008	0.804	0.795	0.8	0.964	O
	0.857	0.006	0.861	0.857	0.859	0.987	P
	0.754	0.011	0.739	0.754	0.746	0.947	Q
	0.806	0.01	0.765	0.806	0.785	0.972	R
	0.689	0.007	0.801	0.689	0.74	0.92	S
	0.871	0.004	0.897	0.871	0.883	0.971	T
	0.891	0.005	0.88	0.891	0.885	0.963	U
	0.895	0.004	0.888	0.895	0.892	0.969	V
	0.923	0.007	0.835	0.923	0.877	0.989	W
	0.895	0.01	0.78	0.895	0.833	0.988	X
	0.902	0.007	0.841	0.902	0.87	0.995	Y
	0.881	0.005	0.871	0.881	0.876	0.974	Z
Weighted Avg.	0.825	0.007	0.836	0.825	0.82	0.951	

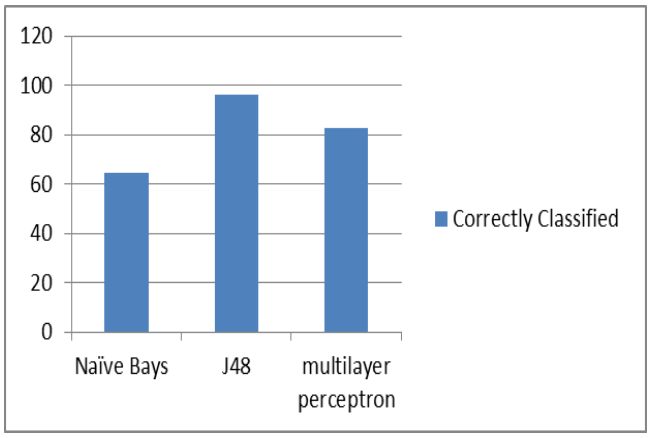

Figure 1: Comparison of classifiers

5. CONCLUSION

In this paper, the text analysis and categorization is done using 'Letter dataset'. Using this dataset, a performance analysis is conducted for three classifiers such as Naive Bayes, Decision Tree J48, and Multi-Layer Perceptron (MLP) for text datasets. These are simulated in weka and the obtained results are tabulated and compared. From the results, it is clear that decision Tree J48 classifier provides better performance in terms of increased accuracy and precision rate.

REFERENCES

[1] Bruno Trstenjaka, Sasa Mikacb, Dzenana Donkocm, "KNN with TF-IDF Based Framework for Text Categorization", 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 69: 1356 – 1364, 2014.

[2] Michal Hrala and Pavel Kral, "Evaluation of the Document Classification Approaches", doi: 10.1007/978-3-319-00969-8_86, 2013.

[3] Ashis Kumar Mandal1and Rikta Sen, "Supervised Learning Methods for Bangla Web Document Categorization", International Journal of Artificial Intelligence & Applications (IJAIA), 5(5), 2014.

[4] Erlin, Unang Rio, "Text Message Categorization of Collaborative Learning Skills in Online Discussion Using Support Vector Machine", 2013 International Conference on Computer, Control, Informatics and Its Applications , 2013.

[5] Joachims, T, "Transductive inference for text classification using support vector machines", Proceedings of ICML-99, 16th International Conference on Machine Learning, eds. Morgan Kaufmann Publishers, San Francisco, US: Bled, SL, 1999, pp. 200–209

[6] Addis, A., "Study and Development of Novel Techniques for Hierarchical Text Categorization", PhD Thesis Electrical and Electronic Engineering Dept., University of Cagliari, Italy, 2010.

[7] Feldman, R & Sanger, J, "The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data", Cambridge University Press New York, 2006.

[8] D. E. Johnson, F. J. Oles, T. Zhang, T. Goetz, "A decision-tree-based symbolic rule induction system for text categorization", IBM Systems Journal, 2002.

[9] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer, "KNN Model-Based Approach in Classification doi: 10.1007/978-3-540-39964-3_62: 986-996, 2003.

[10] C. C. Aggarwal, and C. Zhai, "Mining text data", doi: 10.1007/978-1-4614-3223-4, 2012

[11] Hamood Alshalabi, Sabrina Tiun, Nazlia Omar, Mohammed Albared, "Experiments on the Use of Feature Selection and Machine Learning Methods in Automatic Malay Text Categorization", 4th International Conference on Electrical Engineering and Informatics , pp. 734-739, 2013.

[12] Z. Wei, D. Miao, J.-H. Chauchat, "N-grams based feature selection and text representation for Chinese Text Classification", International Journal of Computational Intelligence Systems, 2 (4), 2009, pp. 365-374.

[13] T. Joachims, "A statistical learning model of text classification for support vector machines", in: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2001, pp. 128- 136.

[14] Forman, G. "An Experimental Study of Feature Selection Metrics for Text Categorization", Journal of Machine Learning Research, 2003, pp. 1289-1305.

[15] Chaitrali S. Dangare and Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications, 47(10), 2012.

[16] D.Lavanya and Dr.K.Usha Rani, "Ensemble Decision Tree Classifier for Breast Cancer Data", International Journal of Information Technology Convergence and Services (IJITCS), 2(1), 2012.

[17] Elakia, Gayathri, Aarthi, Naren J, "Application of Data Mining in Educational Database for Predicting Behavioral Patterns of the Students", International Journal of Computer Science and Information Technologies, 5(3), 2014, pp. 4649-4652.

[18] T. Santhanam and Shyam Sundaram, "Application of CART Algorithm in Blood Donors Classification ", Journal of Computer Science, 6 (5), 2010, pp. 548-552.

[19] Jaydeep Jalindar Patil, Nagaraju Bogiri, "Automatic Text Categorization Marathi Documents ", International Journal of Advance Research in Computer Science and Management Studies, 3(3), 2015.

[20] Ashis Kumar Mandal, Rikta Sen, "Supervised Learning Methods for Bangla Web Document Categorization", International Journal of Artificial Intelligence & Application (IJAIA), DOI:10.5121/ijaia.2014.5508, 2014.