ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **9** (2020), no.6, 3477–3485
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.9.6.27 Spec. Issue on RDESTM-2020

# ANALYSIS OF DIFFERENT CLASSIFICATION ALGORITHMS FOR TEXT DATA MINING

CHARAN SINGH TEJAVATH[1] AND TRYAMBAK HIRWARKAR

ABSTRACT. Text Classification is a quickly developing area of Data Mining while Requirements Engineering is a less-investigated territory of Software Engineering that manages the way toward characterizing, archiving, and keeping up a software framework's prerequisites. At the point when researchers chose to mix these two streams in, there was look into on mechanizing the procedure of classification of software necessities articulations into classifications effectively conceivable for designers for quicker turn of events and conveyance, which till now was generally done physically by software engineers - without a doubt a monotonous activity. This research work examinations the utilization of classification algorithms and their uses to anticipate the uses of text mining. The purpose of this work is to introduce an investigation of ongoing distributions concerning text mining utilizing classification calculation specifically.

## 1. INTRODUCTION

Today web is the primary source of data. Google is the profoundly accessible Search Engine. About 57,000 searches for every second on a day as per Internet Live Stats, 2016. Web Content Mining is finding helpful data from web pages. Dissecting the web page by removing its unstructured data causes us to comprehend its convenience, foreseeing the future prerequisite of the client and some more. Web content Mining utilizes Text, Images, Audio and videos for extraction

[1]*corresponding author*

2010 *Mathematics Subject Classification.* ???, ???

*Key words and phrases.* Text mining, data mining, classification.

of data from Web. Among all text mining is extremely well known since a large portion of the pursuit absolutely utilizes just text archives. Text Mining assists with looking through related examples from web Repository. The task which is difficult in text mining is separating valuable data from unstructured text as there is no appropriate configuration of text in web. Measurable and Machine Learning algorithms utilized for Web Content Mining is to discover importance of web page substance. [1]

1.1. **Machine Learning.** Machine Learning is automatically figure out how to make predictions on current data dependent on previous history. It is partitioned into Supervised and Unsupervised Learning. Regulated Learning is when for each perception I = 1, 2, 3, 4âĂę n and a vector of estimation xi yet not related reaction yi. Unaided learning has inputs however no directing Outputs to learn Relationships and structure of data. Anticipating a consistent quantitative Output esteem is alluded as Regression Problem. Anticipating a non-numerical, Qualitative worth or absolute Output esteem is Classification. Watching just Input Variables and No Output factors and gathering those information factors relying upon their attributes called Clustering. Information factors are alluded as Predictors, Independent, Features or factors X. Yield factors are alluded as Response or Dependent variable Y. The connection between Y (Response) and X (Predictors) and it is composed as:

$$Y = f(X) + E,$$

where $f$ is fixed or unknown function and E is a random error term independent of X and mean zero. To estimate the function f apply a statistical learning method to the training data. Accuracy of f depends on two quantities Reducible error or Irreducible error. If the error can be reduced by increasing the accuracy then it is reducible error. If it cannot be reduced in any case then it is Irreducible error. When a given method yields a small training MSE but a large test MSE it is over fitting of data. Always training MSE smaller than the testing MSE. Variance is the amount by which 'f' would change if estimated it using a different training data set.

Function $f$ is fixed or obscure capacity and E is an irregular blunder term free of X and mean zero. To appraise the capacity f apply a factual learning strategy to the preparation data. Exactness of f relies upon two amounts Reducible blunder or Irreducible mistake. In the event that the mistake can be decreased by

expanding the exactness, at that point it is reducible blunder. On the off chance that it can't be decreased regardless, at that point it is Irreducible blunder. At the point when a given strategy yields a little preparing MSE however an enormous test MSE it is over fitting of data. Continually preparing MSE littler than the testing MSE. Fluctuation is the sum by which $f$ would change whenever evaluated it utilizing an alternate preparing data set.

The primary expanding volume of promptly accessible advanced texts makes characteristic language into a fruitful zone and turns out to be most significant data arrangement of Machine language application. They incorporate crucial language handling and semantic issues, distinguish a word's grammatical feature on its importance, discovering relations between words. In Machine Learning framework the exhibition ought to definitely improve the experience. The Knowledge in Machine Learning System is spoken to as Symbolic Declarative as choice trees, numeric configuration as Support Vector Machine and Naive Bayes or in Model–based ways, for example, Neural Network and Hidden Markov Model. An all around expressed Machine Learning issue needs its info and yield to be Quantified or Categorized. The Categorization of yield is simpler on the off chance that it depends on target factors, for what it's worth in record class by point.

1.2. **Classification.** Classification and Prediction are the two significant strategies for data examination [3]. The initial step of classification process is gathering the records in various expansions. The gathered archives to be changed over into a pre-handled report with techniques like tokenization, Stop Word Removal, Stemming. This makes the report to get decreased from its unique size and simple to handle. At that point Indexing is finished utilizing Vector space model, Semantic portrayal, and Ontological portrayal, N-Grams, Boolean weighting and some more. To improve the Efficiency, Scalability and Accuracy of text highlight Selection is made utilizing Term Frequency, Chi-Square, Information Gain and Genetic Algorithm Optimization. Classification is done subsequent to choosing highlight utilizing some machine learning algorithms Bayesian Classifier, Decision Tree, K-Nearest Neighbor, Support Vector Machines and Neural Networks. At the point when classification is done it ought to be assessed tentatively through assessment techniques, for example, Precision, Recall, Accuracy and a lot more [4]. True Positive is Number of effectively Classified positive

models. Bogus Positive is erroneously ordered Positive models. Bogus Negative is inaccurately grouped Negative models. Accuracy gauges what number of the classifier is right. Review gauges what a small number of right decisions classifier has missed. In application where positive and negative models similarly treated Standard Accuracy is determined. Choice of text portrayal highlights can have any kind of effect among fruitful and Unsuccessful applications [12].

## 2. Classification Algorithms

2.1. **Random Forest.** Ensemble Learning algorithms are precise and strong to noise since it is a combination of more than one classifier. It performs well than single Classifier. Breiman in 2001 proposed this classifier with numerous points of interest, for example, proficient, more information factors took care of, significance of factors, strong to commotion and furthermore exceptions and it is lighter than other outfit algorithms [6]. Random forests helps in ranking the factors in regression or classification.

2.2. **Support Vector Machine.** Support vector machines (SVMs) are one of the discriminative classification techniques which are usually perceived to be progressively precise. The SVM classification strategy depends on the Structural Risk Minimization standard from computational learning hypothesis [9]. The possibility of this rule is to discover a theory to ensure the most reduced genuine blunder. Furthermore, the SVM is all around established that exceptionally open to hypothetical comprehension and investigation. The SVM need both positive and negative preparing set which are extraordinary for other classification strategies. These positive and negative preparing sets are required for the SVM to look for the choice surface that best isolates the positive from the negative data in the dimensional space, supposed hyperplane. The record agents which are nearest to the choice surface are known as the support vector. The presentation of the SVM classification stays unaltered if records that don't have a place with the support vectors are expelled from the arrangement of preparing data.

The SVM classification technique is extraordinary from the others with its exceptional classification viability. Moreover, it can deal with archives with high-dimensional info space and winnows out the vast majority of the superfluous highlights.

Support vector machines (SVM) as illustrative of directed techniques just as inactive semantic ordering (LSI) and self-sorting out maps (SOM) techniques for unaided strategies for framework execution. In [10] the creator examinations and contrasts SVM troupes and four diverse group developing techniques, specifically sacking, AdaBoost, Arc-X4, and an altered AdaBoost. Twenty true data sets from the UCI storehouse are utilized as benchmarks to assess and look at the presentation of these SVM troupe classifiers by their classification exactness. An ideal SVM calculation through different ideal methodologies is created in [9], for example, a novel significance weight definition, the component determination utilizing the entropy weighting plan, the ideal parameter settings. The SVM is the best procedure for the records classification.

TABLE 1. SVM Matrix with Precision and Recall Values

| | **True Negative** | **True Positive** | **Class Precision** |
|---|---|---|---|
| **Pred. Negative** | 233 | 8 | 96.68 % |
| **Pred. Negative** | 5 | 254 | 98.07 % |
| **Class Recall** | 97.90 % | 96.95 % | |

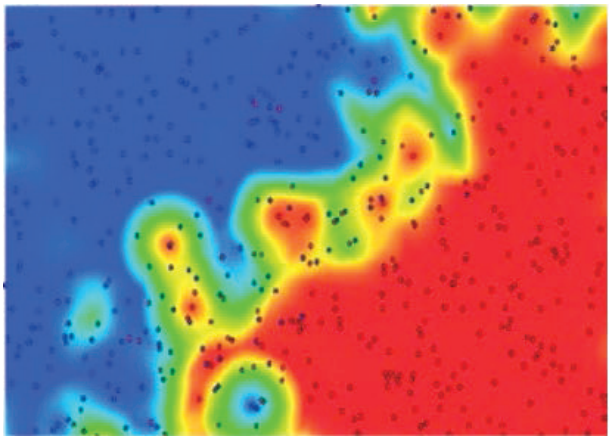

FIGURE 1. SVM Density Graph

2.3. **Naive Bayes.** Naive Bayes classifier is a straightforward probabilistic classifier dependent on applying Bayes Theorem with solid freedom suppositions. An increasingly distinct term for the fundamental likelihood model would be autonomous component model. These autonomy presumptions of highlights

make the highlights request is insignificant and therefore that the present of one element doesn't influence different highlights in classification assignments. These suppositions make the calculation of Bayesian classification approach increasingly productive, however this presumption seriously constrains its materialness. Contingent upon the exact idea of the likelihood model, the naive Bayes classifiers can be prepared productively by requiring a moderately limited quantity of preparing data to evaluate the parameters important for classification. Since free factors are accepted, just the changes of the factors for each class should be resolved and not the whole covariance framework.

Because of its obviously over-rearranged suspicions, the naive Bayes classifiers regularly work much better in numerous intricate genuine circumstances than one may anticipate. The naive Bayes classifiers has been accounted for to perform shockingly well for some true classification applications under some particular conditions. A preferred position of the naive Bayes classifier is that it requires a limited quantity of preparing data to evaluate the parameters important for classification. Bayesian classification approach shows up at the right classification as long as the right classification is more likely than the others. Class' probabilities don't need to be assessed quite well. As it were, the general classifier is sufficiently vigorous to overlook genuine lacks in its fundamental Naive likelihood model.

TABLE 2. Naive Bayes Matrix with Precision and Recall Values

| | **True Negative** | **True Positive** | **Class Precision** |
|---|---|---|---|
| **Pred. Negative** | 216 | 27 | 88.89 % |
| **Pred. Negative** | 22 | 235 | 91.44 % |
| **Class Recall** | 90.76 % | 89.69 % | |

2.4. **Decision Tree.** The decision tree rebuild the manual classification of preparing archives by developing very much characterized valid/bogus questions as a tree structure. In a decision tree structure, leaves speak to the comparing classification of records and branches speak to conjunctions of highlights that lead to those classes. The efficient decision tree can without much of a stretch group an archive by placing it in the root hub of the tree and let it go through the inquiry structure until it arrives at a specific leaf, which speaks to the objective for the classification of the report.

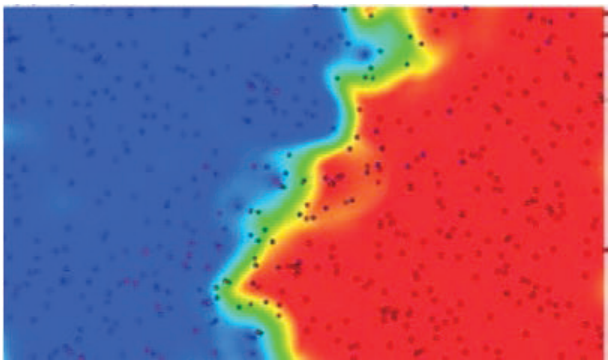

FIGURE 2. Naive Bayes Density Graph

The decision tree classification strategy is remarkable from other decision support apparatuses with a few points of interest. The principle favorable position of decision tree is its effortlessness in comprehension and deciphering, in any event, for non-master clients. Additionally, the clarification of a given outcome can be effortlessly imitated by utilizing basic science algorithms, and give a merged perspective on the classification rationale, which is a helpful data of classification. It tends to be demonstrated tentatively that text classification errands habitually include countless significant highlights. Consequently, a decision tree's inclination to put together classifications with respect to as barely any tests as conceivable can prompt terrible showing on text classification. In any case, when there are few organized qualities, the exhibition, straightforwardness and understandability of decision trees for content-based models are for the most part focal points. The [12] portray an utilization of decision trees for customizing promotions on web pages. The significant danger of executing a decision tree is it over fits the preparation data with the event of an elective tree that sorts the preparation data more regrettable however would arrange the archives to be classified better. This is because of the classification calculation of decision tree is made to sort preparing data viably, anyway disregard the presentation of grouping different records. In addition, tremendous and too much complex structure of tree is worked from a dataset with enormous number of sections.

## 3. Conclusion

Data classification is presently a typical errand applied in numerous application regions, for example, gathering comparative useful genomes, text that exhibit a similar example, apportioning web pages demonstrating a similar structure, etc. This examination places of business different strategies, techniques and execution of Classification Algorithms in Text mining. It is preposterous to expect to anticipate and propose the best algorithms for any sort of utilizations in data mining in light of the fact that, the outcomes are vary starting with one application then onto the next application.

Other references on the topic are [2, 5, 7, 8, 11, 13, 14, 15, 16, 17].

## References

[1] Dasgupta: *Feature selection methods for text classification*, Proceedings of the 13th ACM-SIGKDD international conference on Knowledge discovery and data mining, (2007), 230–239.

[2] G. James, D. Witten, T. Hastie: *An Introduction to Statistical Learning: With Applications in R*, Springer 2004.

[3] S. Vijayarani, M. Muthulakshmi: *Comparative Study on Classification Meta Algorithms*, International Journal of Innovative Research in Computer and Communication Engineering, **1**(8) (2013), 1768–1774.

[4] P. Raghavan, S. Amer-Yahia, L. Gravano: *Structure in Text: Extraction and Exploitation*, Proceeding of the 7th International Workshop on the Web and Databases(WebDB), ACM SIGMOD/PODS 2004, ACM Press **67**, 2004.

[5] M. Lynch: *e-Business Analytics: Depth Report*, Nov. 2000.

[6] V. Korde, C. N. Mahender: *Text classification and classifiers: A survey*, International Journal of Artificial Intelligence and Applications, **3**(2) (2012), 23-33.

[7] S. Das, A. Dey, A. Pal, N. Roy: *Applications of Artificial Intelligence in Machine Learning: Review and Prospect*, International Journal of Computer Applications, **115**(9) (2015), 67–83.

[8] V. F. Rodriguez-Galiano: *An assessment of the effectiveness of a Random forest classifier for land-cover classification*, ISPRS Journal of Photogrammetry and Remote Sensing, **67** (2012), 93–104.

[9] R. Aparicio, E. Acuna: *Using Ontologies To Improve Document Classification with Transductive Support Vector Machines*, International Journal of Data Mining and Knowledge Management Process, **3**(3) (2013), 111–125.

[10] R. P. RAJESWARI, K. JULIET, ARADHANA: *Text Classificationfor Student Data Set using Naive Bayes Classifier and KNNClassifier*, International Journal of Computer Trends and Technology, **43**(1) (2017), 8–12.

[11] J. MAROCO, D. SILVA, A. RODRIGUES, M. GUERREIRO, I. SANTANA, A. DE MENDONCA: *Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification treesand random forests*, BMC research notes, **4**(1) (2011), 55–65.

[12] V. MITRA, C. J. WANG, S. BANERJEE: *Text classification: A leastsquare support vector machine approach*, Applied SoftComputing, **7**(3) (2007), 908–914.

[13] J. FU, C. HUANG, S. LEE: *A multi-class svm classification systembased on methods of self-learning and error filtering*, Taiwan,Republic of China, 2008.

[14] E. SARAVANA KUMAR, K. VENGATESAN, R. P. SINGH, C. RAJAN: *Biclustering of Gene Expression data using Biclustering Iterative Signature Algorithm and Biclustering Coherent Column*, International Journal of Biomedical Engineering and Technology, **26**(3-4) (2018), 341–352.

[15] K. VENGATESAN, A. KUMAR, R. NAIK, D. K. VERMA: *Anomaly Based Novel Intrusion Detection System For Network Traffic Reduction*, 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on, Palladam, India, 2018, 688–690.

[16] M. SANTOSH, A. SHARMA: *A Proposed Framework for Emotion Recognition Using Canberra Distance Classifier*, J. Comput. Theor.Nanosci, **16**(9) (2019), 3778–3782.

[17] J. KAUR, A. SHARMA: *A Novel Method for Video Authenticity Based on the Fingerprint of Scene Frame and Black Frame Identification*, International Journal of Advanced Science and Technology, **27**(1) (2019), 97–103.

CHARAN SINGH TEJAVATH
DEPT. OF COMPUTER SCIENCE AND ENGINEERING
SRI SATYA SAI UNIVERSITY OF TECHNOLOGY AND MEDICAL SCIENCES, SEHORE
BHOPAL-INDORE ROAD, MADHYA PRADESH, INDIA

DEPT. OF COMPUTER SCIENCE AND ENGINEERING
SRI SATYA SAI UNIVERSITY OF TECHNOLOGY AND MEDICAL SCIENCES, SEHORE
BHOPAL-INDORE ROAD, MADHYA PRADESH, INDIA