THINK INDIA JOURNAL

ISSN:0971-1260

Vol-22-Issue-41-December-2019

Performance Evaluation of Text Classifier Using SVM Algorithm

Rajesh Banala, K Sindhuja, D. Sindhuja and V. Keerthana

Sreyas institute of engineering and technology

ABSTRACT

Text categorization also known as text classification is the assignment of routinely sorting a hard and fast of files into categories from a predefined set. This assignment has several programs, such as computerized indexing of medical articles in line with predefined technical phrases, filing patents into patent directories, selective dissemination of facts to data purchasers, computerized population of hierarchical catalogues of internet sources, unsolicited mail filtering, identification of document style, authorship attribution, survey coding and even automated essay grading. Computerized textual content class is appealing as it frees organisation from the want of manually organizing document bases, which may be too expensive, or without a doubt no longer possible given the time constraints of the application or the number of files worried. The accuracy of modern-day text class competitors that of trained human specialists, thanks to a aggregate of statistics retrieval (IR) era and system studying (ML) generation. This project will outline the fundamental tendencies of the technology worried, of the applications that could feasibly be tackled thru textual content classification, and of the gear and assets that vicinity to be had to the researcher and developer wishing to take up those technologies for growing actual-international applications.

Document categorization won quite a few significance within the final years due to the growth in the range of virtual documents. This task analyzes the performance of various classification algorithms on text categorization troubles. Significance of parameter optimization on the overall performance of the algorithms is likewise mentioned. This venture particularly specializes in the aid Vector machine (SVM) algorithms. Considering the fact that in line with the literature, its overall performance is superior to different system learning algorithms in textual content categorization trouble.

Keywords: Support Vector Machine (SVM), Natural Language Processing (NLP), Neural Networks, Naive Bayes classification.

1. INTRODUCTION

One of the subfields of predictive modelling is supervised pattern type. Supervised pattern class is the task of education a version primarily based labeled schooling records which then may be used to assign a pre-defined class label to new gadgets. One example that we can explore is unsolicited mail filtering via the use of one of a kind system mastering algorithms like Naive Bayes, Neural Networks, guide Vector system (SVM) which will are expecting whether a brand new text message may be categorized as spam or not junk mail.

1.1 NATURAL LANGUAGE PROCESSING:

One of the extensively used natural language processing undertaking in unique commercial enterprise troubles is "textual content type". The goal of text type is to routinely classify the text documents into one or more defined categories. We have a few examples of textual content type are:

- understanding target audience sentiment from social media,
- Detection of spam and non-junk mail emails,
- automobile tagging of customer queries and
- Categorization of news articles into described topics.

ISSN:0971-1260

Vol-22-Issue-41-December-2019

1.2 NAIVE BAYES:

Naive Bayes is a own family of statistical algorithms we are able to employ while doing textual content type. One of the contributors of that circle of relatives is Multinomial Naive Bayes (MNB).Its main benefits is that you can get truly excellent consequences while facts to be had is not an awful lot and computational assets are scarce. All you want to recognize is that Naive Bayes is based totally on Bayes's Theorem, which enables us to compute the conditional possibilities of incidence of events based on the possibilities of occurrence of every character occasion. This means that any vector that represents a textual content will have to comprise information about the chances of appearance of the phrases of the text inside the texts of a given category in order that the algorithm can compute the likelihood of that textual content's belonging to the category.

1.3 NEURAL NETWORKS:

The neural networks is likewise a one of the supervised studying algorithm which we will select to apply in text category. The neural community may have issue converging earlier than the most range of iterations allowed if the records is not categorised. Multi-Layer perceptron (MLP) is touchy to characteristic scaling, so it is exceptionally encouraged to scale your information. Be aware that you have to observe the same scaling to the test set for meaningful outcomes. There are a number of one of a kind methods for type of information.

To create an example of the model, there are lots of parameters you can pick out to define and customize. On this we will define handiest the hidden layer sizes, to this parameters you pass a tuple together with the quantity of neurons you want at each layer, in which the last access inside the tuple represents the range of neurons within the closing layer of the MLP version, there are many ways to pick these numbers, but for simplicity we are able to choose three layers with the equal wide variety of neurons with the identical number of neurons as there are capabilities in our dataset.

1.4 SUPPORT VECTOR MACHINE:

Aid Vector system (SVM) is simply one out of many algorithms we are able to pick from when doing text category. Like naive Bayes, neural networks, SVM does not need much education information to begin imparting accurate outcomes. Although it wishes more computational sources than naive Bayes and neural networks, SVM acquire extra correct effects.

In brief SVM takes care of drawing a "line" or hyper plane that divides a space into subspaces, one subspace that incorporates vector that belong to a group and another subspace that contains vectors that don't belong to that organization. Those vectors are representations of your education texts and a group is a tag you have got tagged your texts.

2. PROBLEM STATEMENT

Report category or document categorization is a hassle in library technological know-how, statistics technological know-how and pc technology. The assignment is to assign a document to one or more instructions or classes. This may be performed "manually" (or "intellectually") or algorithmically. The intellectual classification of files has ordinarily been the province of library science, whilst the algorithmic category of documents is especially in facts technological know-how and laptop technological know-how. The issues are overlapping, but, and there is consequently interdisciplinary studies on document type.

The documents to be labeled may be texts, pix, music, and many others. Each form of document possesses its unique category issues. Whilst now not otherwise distinctive, text type is implied.

Files may be classified in step with their subjects or consistent with different attributes (consisting of record kind, author, printing 12 months and many others.). In the relaxation of this newsletter most effective situation category is taken into consideration. There are two main philosophies of challenge category of documents: the content material-based approach and the request-based totally method.

3. PROPOSED SYSTEM

Many applications in textual content processing require large human effort for both labeling large record collections or extrapolating policies from them. On this work, we describe away to lessen this attempt, while

ISSN:0971-1260

Vol-22-Issue-41-December-2019

maintaining the strategies accuracy, by constructing a hybrid classifier that makes use of the human reasoning over mechanically found textual content styles to supplement gadget studying. The use of a standard type dataset, we show that the ensuing method results in extensive reduction of the human attempt required to attain a given classification accuracy. Furthermore the hybrid textual content classifier also effects in a substantial boost in accuracy over device gaining knowledge of primarily based classifiers whilst a similar amount of categorised data is used.

That allows you to improve an automated system of organizing and classifying fitness care information we suggest an method this is textual content mining. It's miles new and exciting research place that rises to solve the studies overload problem by means of the usage of approach from system getting to know, NLP (herbal language processing), information mining, records extraction, and many others.And this is applicable the strategies of supervised device mastering and particular algorithms for text categorizations SVM (aid vector gadget) . The proposed statistics classifies the given information consistent with its precis. It classifies very massive amounts of statistics more correctly and efficiently.

NAIVE BAYES ALGORITHM INPUT:
Training dataset T,

F= (f1, f2, f3,...fn) //value of the predictor variable in testing dataset.

OUTPUT:
A class of testing dataset.

STEP:

1. Read the training dataset T;
2. Calculate the mean and standard deviation of the predictor variables in each class;
3. Repeat

   Calculate the probability of fi using the equation of each class;

   Until the probability of all predictor variables (f1, f2, f3,…fn) has been calculated.
4. Calculate the likelihood for each class;
5. Get the greatest likelihood;

3.1 NEURAL NETWORKS
Neural Networks are a system mastering framework that tries to mimic the gaining knowledge of sample of natural organic neural networks. Organic neural networks have interconnected neurons with dendrites that get hold of inputs, then based on those inputs they produce an output sign via an axon to some other neuron. We will attempt to mimic this technique thru the usage of synthetic Neural Networks (ANN), which we will just talk to as neural networks any further. The system of making a neural network starts with the maximum fundamental form, a single notion.

3.2 NEURAL NETWORKS ALGORITHM:
repeat

for each training vector pair(x,t)

evaluate the output y when x is the input if y!=t then

from a new weight vector w’ according to w’=w+alpha(t-y)x

else

do nothing end if

end for

ISSN:0971-1260

Vol-22-Issue-41-December-2019

until y=t for all training vector pairs

SVM ALGORITHM:
Input: Training set S; Number of base classifiers N; Test Instance x:

TRAINING PROCESS:
for i=1: N

Use Bootstrap technology on S to generate training subset S_i; Train the i^{th} SVM Classifier SVM_i ;

call SVM-OTHER algorithm to adjust the classification hyperplane of SVM_i to form $SVM\text{-}OTHER_i$;

end

TESTING PROCESS:
for i=1: N

put x into the classifier $SVM\text{-}OTHER_i$ to predict its class label y_i;

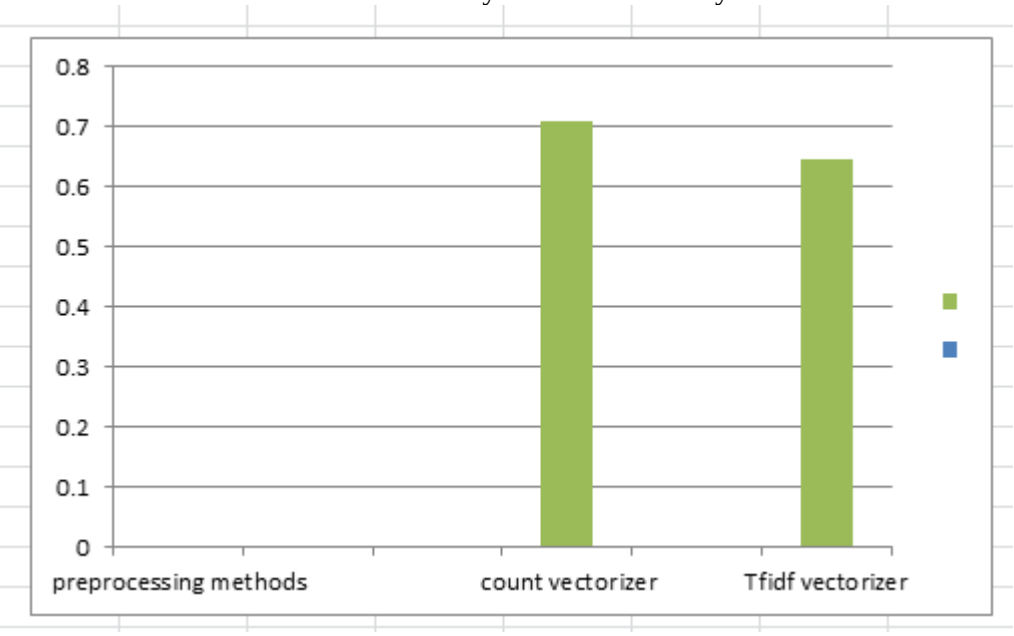
end

Output: y which is the predicted class label for test instance x

NAIVE BAYES GRAPH
Comparison between preprocessing methods

preprocessing methods	classification techniques
	Naive Bayes
count vectorizer	0.708
Tfidf vectorizer	0.645

Table 9.1: Accuracy values in naive bayes



ISSN:0971-1260

Vol-22-Issue-41-December-2019

NEURAL NETWORKS

preprocessing methods	classification techniques
	neural networks
count vectorizer	0.736
Tfidf vectorizer	0.743

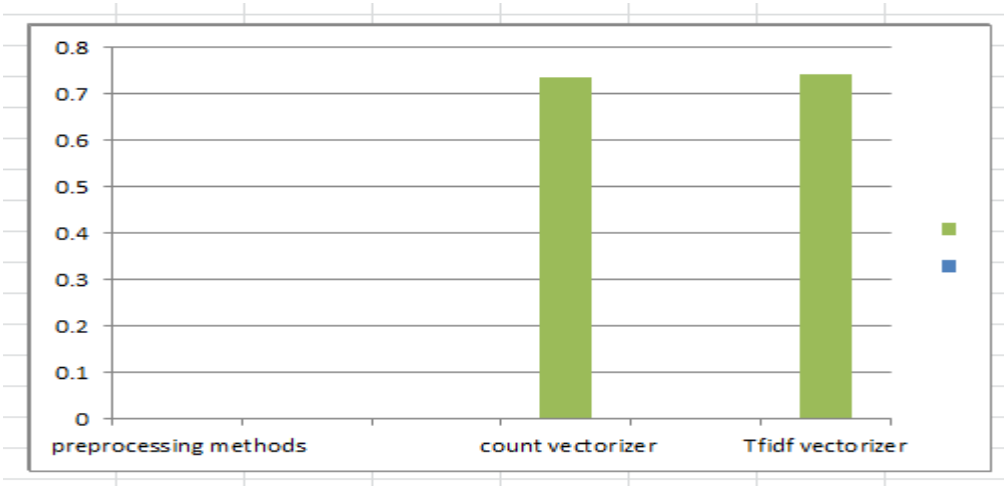
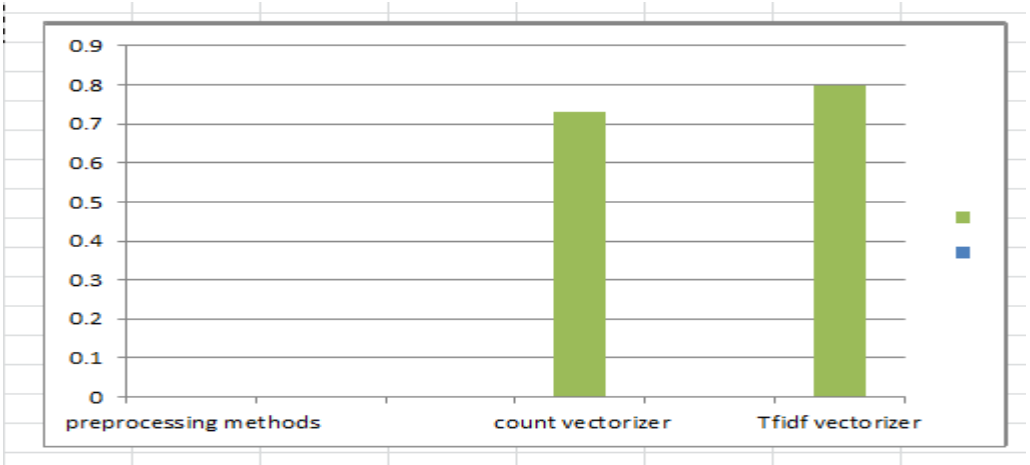


Table 9.2: Accuracy values in neural networks

Support vector machine

preprocessing methods	classification techniques
	SVM
count vectorizer	0.729
Tfidf vectorizer	0.798



ISSN:0971-1260

Vol-22-Issue-41-December-2019

Table 9.3 : Accuracy value in SVM

Comparison between Naive Bayes, Neural networks and Support vector machine

Preprocessing methods	classifaction techniques		
	naive bayes	neural networks	svm
count vectorizer	0.708	0.736	0.729
Tfidf vectorizer	0.645	0.743	0.798

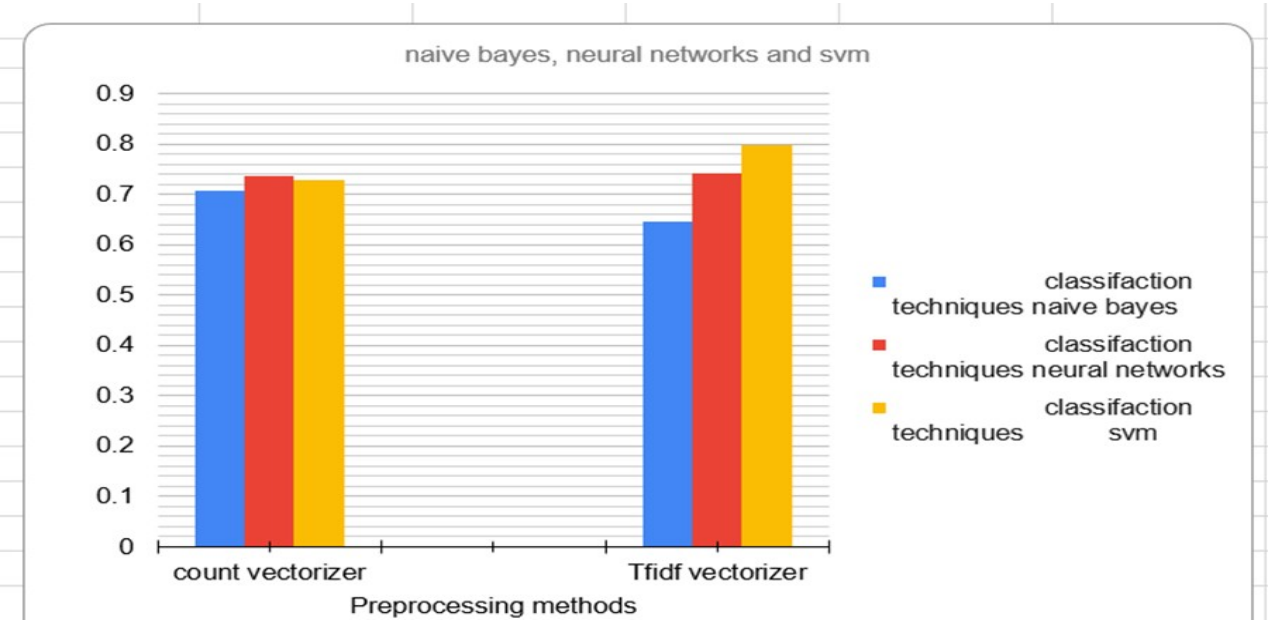


Table 9.4: To compare values of three algorithms

8. CONCLUSION

The textual content category problem is an synthetic intelligence research subject matter, specially given the big wide variety of documents available inside the shape net pages and other digital texts like emails, dialogue forum postings and other digital documents.

It has located that even for a specific class method, classification performances of the classifiers based totally on specific education text corpuses are specific; and in a few cases such variations are quite sizeable. This commentary implies that classifier overall performance is applicable to its schooling corpus in a few diploma and appropriate or high quality corpuses can also derive classifiers of properly performance. Unluckily, to date little studies paintings in the literature has been visible on the way to take advantage of schooling corpuses to enhance classifier overall performance.

A few important conclusions have no longer been reached yet, together with:

- Which characteristic choice techniques are each computationally scalable and excessive acting across classifiers and collections? Given the high variability of textual content collections, do such methods even exist?
- could combining uncorrelated, however properly acting strategies yield a performance increase?
- change the thinking from word frequency based vector space to concepts based totally vector area. Look at the technique of function choice under concepts, to look if those will assist in text categorization.
- Make the dimensionality discount greater efficient over huge corpus.

ISSN:0971-1260

Vol-22-Issue-41-December-2019

Moreover, there are other two problems in text categorization: polysemy, synonymy. Polysemy refers to the fact that a word will have multiple meanings. Distinguishing among extraordinary meanings of a phrase isn't clean. Synonymy way that distinctive words will have the same or similar that means.

9. REFERENCES

1. Crammer, K. & Singer, Y., On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research, 2, pp. 265–292, 2001.
2. Sebastiani, F., Machine learning in automated text categorization. ACM Com- puting Surveys, 34(1), pp. 1–47, 2002.
3. Debole, F. & Sebastiani, F., Supervised term weighting for automated text categorization. Proceedings of SAC-03, 18th ACM Symposium on Applied Computing, ACM Press, New York, US: Melbourne, US, pp. 784–788, 2003. An extended version appears as [64].
4. Yang, Y. & Pedersen, J.O., A comparative study on feature selection in text categorization. Proceedings of ICML-97, 14th International Conference on Machine Learning, ed. D.H. Fisher, Morgan Kaufmann Publishers, San Fran-cisco, US: Nashville, US, pp. 412–420, 1997.
5. Wiener, E.D., Pedersen, J.O. & Weigend, A.S., A neural network approach to topic spotting. Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, US, pp. 317–332, 1995.
6. Joachims, T., Text categorization with support vector machines: learning with many relevant features. Proceedings of ECML-98, 10th European Conference on Machine Learning, eds. C. Nedellec C. ouveirol, Springer Verlag, Heidelberg, DE: Chemnitz, DE, pp. 137–142, 1998. Published in the “Lecture Notes in Computer Science” series, number 1398.
7. Giorgetti, D. & Sebastiani, F., Automating survey coding by multiclass text categorization techniques. Journal of the American Society for Information Science and Technology, 54(12), pp. 1269–1277, 2003.
8. Rish, Irina (2001). An empirical study of the naive Bayes classifier (PDF). IJCAI Workshop on Empirical Methods in AI.
9. Rosenblatt, F. (1958). "The Perceptron: A Probabilistic Model for Information Storage And Organization In The Brain". Psychological Review. 65 (6): 386–40

8. CiteSeerX 10.1.1.588.3775. doi:10.1037/h0042519. PMID 13602029