INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

REVIEW ON TEXT CLASSIFICATION USING DIFFERENT CLASSIFICATION TECHNIQUES

PRADIP INGLE[1], SNEHAL INGOLE[2], HEMANT DESHMUKH[2], IRSHAD ANSARI[2]

1. Department Of Information Technology, Anuradha Engg College Chikhli, Sant Gadge Baba Amravati University Amravati (Maharashtra) India.
2. Department Of Information Technology, Anuradha Engg College Chikhli, Sant Gadge Baba Amravati University Amravati (Maharashtra) India.

Accepted Date: 05/03/2015; Published Date: 01/05/2015

Abstract: Now a day's proper classification of large amount of information present over the internet is very critical step towards the business due to the explosive growth of the textual information day by day from the electronic documents and World Wide Web (www), so the proper classification of such enormous amount of information is became basic need towards the business success. Recently, numerous research activities have been conducted in the field of document classification, applying particularly for the formation of knowledge repositories, website classification, ontology mapping, spam filtering, and emails categorization. The most growing interest in the area text mining research is document classification. The identification of documents into particular category in correct manner is still big challenge because large and vast amount of features are present in the dataset. Regarding to the existing classification approaches such as Decision Tree, Neural Network etc. Due to its simplicity Naïve Bayes is potentially better than other approaches at serving as a document classification model. The main intention of this paper is to highlight the performance of Naïve Bayes classification modal by employing it in the area of document classification.

Keywords: Text classification, unstructured data, Naïve Bayes, Decision Tree, Neural Network

Corresponding Author: MR. PRADIP INGLE

Access Online On:

www.ijpret.com

How to Cite This Article:

Pradip Ingle, IJPRET, 2015; Volume 3 (9): 1158-1166

PAPER-QR CODE

INTRODUCTION

A “Text document” refers to printed, written, or online document that presents or communicates narrative or tabulated data in the form of an article, letter, memorandum, report, etc. The Text not only expresses a vast range of information, but also automatically encodes the information in the form that is difficult to decipher. The information is basically available in two form structured data and unstructured data, the term “structured data” is referred to data that resides in fixed fields either in a record or in a file, Relational databases and spreadsheets are examples of structured data, and the term “unstructured data” is referred to data that does not reside in fixed locations, it is also called as the “free-form text” which is ubiquitous. It is mostly related to computerized information that either does not have a data model or has one that is not easily usable by a computer program. The term distinguishes such information from the data stored in fields of databases or annotated in documents.

However, data mining deals with structured data, whereas text presents special characteristics and it is unstructured. The important things are Extraction, Integration and classification of electronic documents from different sources and knowledge discovery from these documents. Machine learning is often used in data mining, for Prediction or Classification. Prediction means extracting information from data and using it to predict future trends and behavior patterns.

Structured Data –

Structured data first depends on creating a data model – a model of the types of business data that will be recorded and how they will be stored, processed and accessed. This includes defining what fields of data will be stored and how that data will be stored: data type and any restrictions on the data input Structured data is often managed using Structured Query Language (SQL) – a programming language created for managing and querying data in relational database management systems. Originally developed by IBM in the early 1970s and later developed commercially by Relational Software, Inc.

Unstructured Data –

Unstructured data is all those things that can't be so readily classified and fit into a neat box: photos and graphic images, videos, streaming instrument data, webpage’s, pdf files, PowerPoint presentations, emails, blog entries, wikis and word processing documents. Examples of "unstructured data" may include journals, books, health records, audio, video,

documents, metadata, analog data, images, files, and unstructured text such as the body of an e-mail message, Web page, or word-processor document.

Fig.1 Unstructured data

Text mining algorithms:

Text classification by SVM:

The dimension of the text data is huge for the text documents are usually represented with the vector space model. To improve the executing efficiency of classification methods, they present a classification algorithm based on nonlinear dimensionality reduction techniques and support vector machines.

Experimental results demonstrate that the executing efficiency of categorization methods is greatly improved after decreasing the dimension of the text data without loss of the classification accuracy. After pre-processing and transformations, a machine learning algorithm is used for learning how to classify documents, i.e. creating a model for input-output mappings [2].

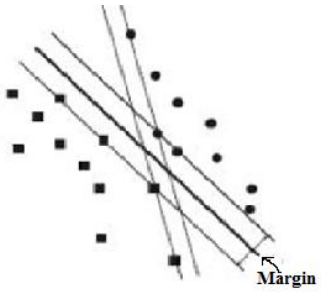

Fig.2. SVM decision margin

Advantages:-

1. It decreases the dimension without a loss of classification accuracy.

2. Decreases the memory requirements.

3. It can take high-dimensional input space.

Disadvantages:-

1. To lower the dimensionality need the dimensionality reduction techniques, this is also time consuming technique.

2. This is fully depending on vector space that we decides, if parameter get wrong then accuracy of the system also get affected.

3. Due dimension reduction there may be chance of data loss.

Text classification by Decision Tree (DT):

The Decision Tree is introduced in the text classification. An input string determines a unique path from the root to a leaf at each internal node the right edge to a child is taken if the input string contains the string labeled at that internal node as a substring [8].

Advantages:-

1. The algorithm does not need any natural language processing technique.

2. The algorithm is robust for classification of noise contained by the sample.

3. Due to tree structure decision is taken very fast.

Disadvantages:-

1. Decision trees are easy to use compared to other decision-making models, but preparing decision trees, especially large ones with many branches, are complex and time-consuming affairs.

2. Cost of decision tree implementation is very high for good decisions.

3. Analytical area of decision tree is imitated.

Text classification by Neural Network:-

The structure of web classification mining system based on wavelet neural network is given. With the ability of strong nonlinear function approach and pattern classification and fast convergence of wavelet neural network, the classification mining method can truly classify the web text information. The neural network is a high nonlinearity dynamics system, and the method of searching problem generally uses the gradient descent method and the random search method. Wavelet neural network is new kinds of network based on the wavelet transform theory and the artificial neural network[4]. Meanwhile the wavelet neural network has the simple implementation process and fast convergence rate.

Advantages:-

1. It does follow pattern classification technique so that time consumption and accuracy is more.

2. This classification method shows the results feasible and effective.

3. Speed of classification is fast.

Disadvantages:-

1. It works on the trained patterns so that it does not work for other requirement. It needs training to node to get that.

2. Neural network requires training to their nodes. So that it is time consuming process.

3. Due pre-requisite training cost of implementation also increases.

Text mining With Naïve bayes

The document representation is the pre-processing process that is used to reduce the complexity of the documents and make them easier to handle, which needs to be transformed from the full text version to a document vector. Dimensionality reduction (DR) is a very important step in text categorization, because irrelevant and redundant features often degrade the performance of classification algorithms both in speed and classification accuracy, the current literature shows that lot of work are in progress in the pre-processing and DR, and many models and techniques have been proposed. DR techniques can classify into Feature Extraction (FE) approaches and feature Selection (FS), as discussed below.

- Feature Extraction
- Feature selection

- Semantic and ontology based document representation
- Learning algorithm

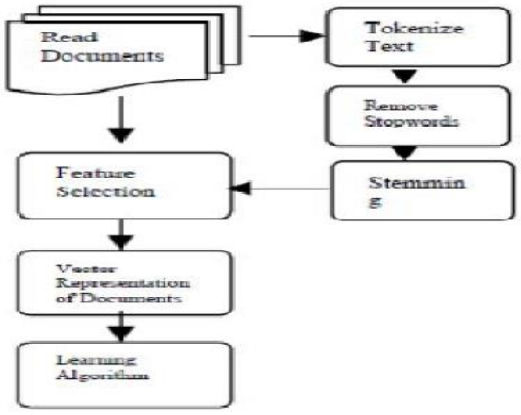

Fig. 3. Proposed system

Feature extraction:-

The process of feature extraction is to make clear the border of each language structure and to eliminate as much as possible the language dependent factors, tokenization, stop words removal, and stemming. Feature Extraction is fist step of pre processing which is used to presents the text documents into clear word format. The documents in text classification are represented by a great amount of feature and most of then could be irrelevant or noisy [7-8] Dimension reduction is the exclusion of a large number of keywords, base preferably on a statistical criterision, to create a low dimension vector.

Feature selection:-

After feature extraction the important step in pre-processing of text classification, is feature selection to construct vector space or bag of words, which improve the scalability, efficiency and accuracy of a text classifier [7-8]. In general, a good feature selection method should consider domain and algorithm characteristics. The main idea of FS is to select subset of feature from the original documents. FS is performed by keeping the words with highest score according to predetermined measure of the importance of the word. Hence feature selection is commonly used in text classification to reduce the dimensionality of feature space and improve the efficiency and accuracy of classifiers.

There are mainly two types of feature selection methods in machine learning; wrappers and filters. As opposed to wrappers, filters perform feature selection independently of the learning

algorithm that will use the selected features The term (word) frequency/inverse document frequency (TF-IDF) approach is commonly used to weight each word in the text document according to how unique it is. In other words, the TF-IDF approach captures the relevancy among words, text documents and particular categories. Some of the recent literature shows that works are in progress for the efficient selection of the feature selection to optimize the classification process. We in developed a new feature scaling method, called class–dependent–feature–weighting (CDFW) using naive Bayes (NB) classifier. Many feature evaluation metrics have been explored, notable among which are information gain (IG), term frequency, Chi-square, expected cross entropy, Odds Ratio, the weight of evidence of text, mutual information, Gini index. **Semantic and ontology**

Learning Algorithm: (Bayesian Theorem):-

The goal of using Bayes rules is to correctly predict the value of designated discrete class variable given a vector of predictors or attributes [8]. These pages will introduce the theorem and its use in the philosophy of science.

Following are the formulas used –

Baye's Formula –

Let $B_1, B_2, B_3, \ldots\ldots B_n$ be a partition of Ω (space) such that $P(B_n) \neq 0$ for any

n = 1, 2, 3... and let P (A) ≠ 0.Then,

$$P(A|B_n) = \frac{P(B_n|A)P(B_n)}{\sum P(B_n|A)P(B_n)}$$

Where, n =1, 2, 3, 4...

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.$$

Mathematically it is represented as –

n = 1, 2, 3... and let P (A) ≠ 0.Then,

$$\boldsymbol{P(A|B_n) = \frac{P(B_n|A)P(B_n)}{\sum P(B_n|A)P(B_n)}}$$

By using Bayesian network and feature variable Naïve Bayes classifies the data with following formula

$$\text{classify}(f_1, \ldots, f_n) = \underset{c}{\operatorname{argmax}}\, p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c).$$

Here f1…fn are the features set, we can calculate feature probability overclass for 1 to n scope.

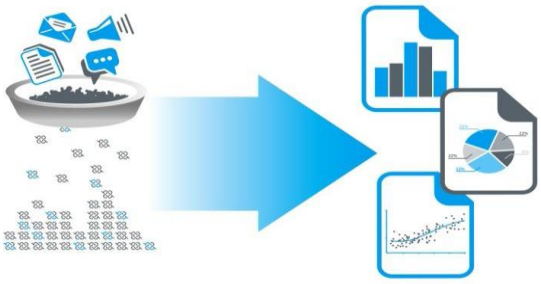

Fig. 3. Process

CONCLUSION

Naïve Bayes classifier has been discussed as the best document classifier, which satisfies the literature result, through the implementation of different feature selection and classifier. There are many words in the documents, therefore when we captured the terms from these documents, thousands of terms are found. However, there are some terms that are usefulness and uninteresting to the results, it is then important to discover and interpret which features are useful and critical.

REFERENCES

1. Text Categorization and Support Vector Machines, István Pilászy, Department of Measurement and Information Systems, Budapest University of Technology and Economics.

2. Fabrizio Sebastiani, “Machine learning in automated text categorization”, ACM Computing Surveys, vol. 34, no. 1, pp.1–47, 2002.

3. Thorsten Joachims, “Text categorization with support vector machines: learning with many relevant features”, Proc. of ECML-98, 10th European Conference on Machine Learning, pp. 137–142, Springer Verlag, Heidelberg, DE, 1998.

4. NTC (Neural Text Categorizer): Neural Network for Text Categorization, Taeho Jo School of Information Technology & Engineering, Ottawa University, Ontario, Canada, Vol 2, issue 2, April 2010.

5. Is Naïve Bayes a Good Classifier for Document Classification?, S.L. Ting, W.H. Ip, Albert H.C. Tsang, Vol. 5, No. 3, July, 2011

6. Naïve Bayes, http://www.wikipedia.com/Naive %20Bayes

7. Aurangzeb Khan, Baharum B. Bahurdin, Khairullah Khan, An Overview of E-Documents Classification, 2009 International Conference on Machine Learning and Computing.

8. Bayesian Theorem, http://www.wikipedia.com/bayesian_theorem

9. George Forman, Evan Kirshenbaum, Extremely Fast Text Feature Extraction for Classification and Indexing, HP Laboratories

10. Pingpeng Yuan, Yuqin Chen, Hai Jin, Li Huang, MSVM-kNN: Combining SVM and k-NN for Multi-Class Text Classification, IEEE International Workshop on Semantic Computing and Systems

11. Eibe Frank and Remco R. Bouckaert, Naive Bayes for Text Classification with Unbalanced Classes.

12. Zeeshan Ahmed and Saman Majeed Machine Learning and Data Optimization using BPNN and GA in DOC Int. J. Emerg. Sci., 1(2), 108-119, June 2011