DOCUMENT CLASSIFICATION USING MACHINE LEARNING

Prajakta Pawale[1], Poonam Masal[2], Ankita Jadhav[3], Prof. Shivraj B Kone[4]

***Abstract*: The main objective is to classify the sector from IT analysis papers and compare the accuracy in 2 classifiers. Classification is the sort of information analysis that may be wont to extract models describing vital information category or to predict future information trends. The foremost vital options area unit designated and information area unit ready for learning and classification. Text classification is that the method of assigning a document to at least one or additional target classes supported its contents. Coaching and classification area unit performed exploitation Naïve Bayes (NB) classifiers.**

Experimental results show that the ways area unit favorable in terms of their effectiveness and potency. This technique summarizes text on 10 classes like "Big Data", "Image Processing"," Information Mining", "Artificial Intelligent", "Ontology", "Data Base Management System", "Management data System" and "Software Engineering" then on. Technique calculates the accuracy of testing information exploitation.

***Keywords*: Naïve Bayes (NB), Machine Learning, Natural Language processing, TF-IDF(Term-Frequency inverse document frequency)**

Introduction:

Document classification is that the work of uncertain documents into classes supported their content. There area unit several classification strategies for documents. Classification is outlined as categorizing document into one in every of a set variety of predefined categories with one document happiness to just one category. The document is entered as input during this system, then the system can do preprocessing steps. Main words area unit solely retrieved. That means summarize words and words area unit matched with all the information needed for every field hold on within the info.

Then count the necessary words exploitation term frequency (TF) in feature choice. Finally, calculate the accuracy by exploitation holdout technique. This paper is comparing the results of Naïve Bayes classifier techniques. Highlighted 5 algorithms that area unit applied to the text classification and expressed comparative study on differing types of classifiers with their enhancements.

This project uses machine learning techniques for classification and summarizers of documents. Machine Learning enables systems to recognize patterns based on existing algorithms and data sets and to develop adequate solution concepts. Machine Learning undoubtedly helps people to work more creatively and efficiently. Therefore, in Machine Learning, artificial knowledge is generated based on experience. In this project, by classifying and summarize a document, one or more categories are assigned to a document, making it easier to manage and sort. This is especially useful for publishers, news sites, blogs or anyone who deals with a lot of content.

Related work:

Nowadays finding the papers related to a particular domain is very difficult until we read the whole paper. Classification and summarize is a machine learning technique that assigns categories to a collection of data to aid in more accurate predictions and analysis. The document is entered as input in this system, and then the system will do preprocessing steps. The main words are only retrieved. The main words are matched with all the data required for each field stored in the database and then count the important words using Term Frequency in feature selection. The document is entered as input in this system, and then the system will do preprocessing steps. The main words are only retrieved. The main words are matched with all the data required for each field stored in the database. Summarize that document by using NLP algorithms and then count the important words using Term Frequency (TF) in feature selection. Finally, calculate the accuracy by using the holdout method. This section provides the characteristics and descriptions of the data sets used for performing our experiments. In this system, the user can know any field of document and calculate the probability of the words of the document. Only words including more and more in a document are to display what kind of field. But if the count of words is the same, it needs to make accuracy so that the program displays what kind of field. This classification and technique are based on Bayes theorem with an assumption of independence between predictors. NLP is used for summarization techniques. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

1. Classification of text documents based on naïve Bayes using n-gram features

Description: Text and document classification processes are often used in areas such as sentiment analysis, text summarization, etc. The author Mehmet Baygin has performed document classification using the Naive Bayes approach.

2. Document clustering: TF-IDF approach

Description: Clustering techniques can be applied only on structured data. So unstructured data need to be converted to structured data. But while converting unstructured data into structured data the algorithm efficiency decreases. So to increase the efficiency For this, we are going to use the TF-IDF approach.

3. An outcome-based comparative study of different text classification algorithms

Description: Text classification has become more important due to the growth of big data with which we could obtain huge data daily. It has many applications like information retrieval, spam detection, language identification, sentiment analysis and plays a major role in natural language processing as well

4. Text classification and classifiers: a survey

Description: In this paper, Namrata Mahender and Vandana Korde have tried to give the introduction of text classification its process and also the overview of classifiers. They also tried to compare some existing classifiers. The existing classification methods are compared and contrasted based on various parameters. They also found that the performance of the classification algorithm is greatly affected by the quality of the data source.

5. Improved c4.5 decision tree classifier algorithm for analysis of data mining application

Description: A decision tree is an important method for both induction research and data mining which is mainly used for model classification and prediction. ID3 and C4.5 algorithm is the most widely used algorithm in the decision tree. We aim to implement the algorithms in a very time and space-effective manner and throughput and response time for the application will be promoted as the performance measures. Our project aims to implement these algorithms and graphically compare the complexities and efficiencies of these algorithms.

Motivation:

Due to the Huge domains in the IT industry, it's very difficult to find classify the domains from a huge number of PDFs. The main motivation of this project to classify the field from IT research papers and compare the accuracy of two Algorithms Naïve Bayes. We are going to generate a short Description of that paper so that we can save the time of classification of Papers

System Architecture:

The system architecture describes the overall flow of the system. This system is useful for the early classification of the post. The user who will use this system needs to first register into the system. The details will be stored in the database. After registration, the user will log in to the system using the login.JSP page. Now the user will enter the details like age, gender, etc which is mentioned in the form which we are using. The algorithm used in the system is Core NLP for text mining. For classification, the Naïve Bayes algorithm is used. The prediction result is shown on the Notification page.

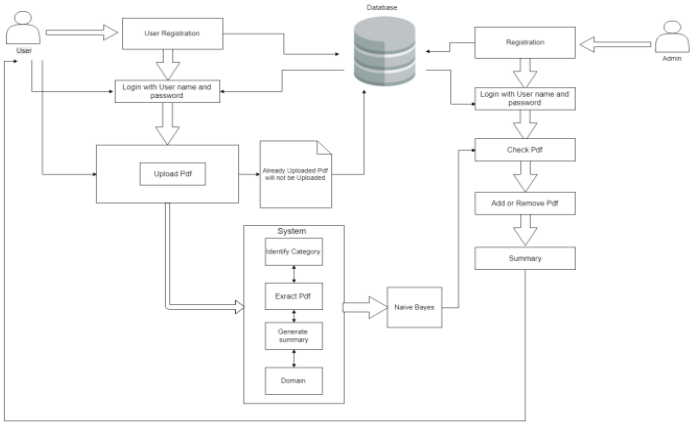

Fig. System Architecture

Conclusion:

In this system, we have developed an application by which we can identify the domain of our document. Due to the use of our system, we can easily be finding the domain of our document which is useful business or education purposes. This paper expressed the extraction of fields document related to IT research paper. It applied the Naive Bayes algorithms to classify documents automatically. This classifier gives a correct and accurate result.

References:

[1] Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In mining text data (pp. 163-222). Springer US.

[2] Mamoun, R., & Ahmed, M. A. (2014). A Comparative Study on Different Types of Approaches to the Arabic text classification. In Proceedings of the 1st International Conference of Recent Trends in Information and (Vol. 2, No. 3).

[3] A.Helen Victoria, M.Vijayalakshmi.An Outcome-based Comparative study of different Text Classification Algorithm. Volume 118 No. 22 2018, 1871-1877

[4] Ahmed, H. A ., & Esrra, H. A. A (2017). Comparative Study of Five Text Classification Algorithms with their Improvements International Journal of Applied Engineering Research, 12(14),4309-4319

[5] Korde, V., & Mahender, C. N. (2012). Text classification and classifiers: A survey. International Journal of Artificial Intelligence & Applications, 3(2), 85.

[6] Badgujar, M. G. V., & Sawant, K. (2016). Improved C4. 5 Decision Tree Classifier Algorithms for Analysis of Data Mining Application. International Journal, 1(8).

[7] Amna Rahman, Usman Qamar(2016). A Bayesian Classifiers based Combination Model for Automatic Text Classification. International Journal, 1(6).

[8] Petre, R. (2015). Enhancing Forecasting Performance of Naive-Bayes Classifiers with Discretization Techniques. Database Systems Journal, 6(2), 24-30.

[9] Mehdi Allahyari, Seyedamin Pouriyeh, A Brief of Text Mining: Classification, Clustering, and Extraction Techniques. KDD Bigdas, August 2017, Halifax, Canada.

[10] Mehdi Allahyari and Krys Kohut. 2016. Discovering Coherent Topics with Entity Topic Models. In Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on. IEEE, 26–33.