Tshe Speculative Study On Machine Learning Algorithm For An Efficient Classification Of Text Document.

[1]k.Prashant Gokul ,[2] Dr.M.Sundararajan

[1]research Scholar, Bharath Institute Of Higher Education And Research,Chennai,India

[2]professor-Ece, Dean Research, Bharath Institute Of Higher Education And Research,Chennai,India

Abstract

Document Classification Is A Problem In Information Science. The Text Documents Are Assigned To One Or More Classes Or Categories For Classification. Classifying Text Documents Using Various Classifier Algorithms And Are Compared With Different Metrics. Based On The Experiment It Is Demonstrated That Svm Classifier Model Proves To Be Better Than Other Models, Whereas While Comparing The Accuracy Random Forests Model Proves To Have Good Accuracy Score But The Time Taken By The Training Phase Is Not Considerable. For This Experiment Three Classes Have Been Used Among That The "Administrative" Category Is The Most Frequent Category As Its Count Value Is Greater And Also It Confirms That There Are No Missing Values. The Accuracy Score Of Random Forest And Non Linear Svm Classifiers' Is Quite Appealing The Magnitude Is High When Considering The Training Time. The Chi-Square Test Is Applied To Select Less Number Of Features And The Algorithms Are Applied, And It Demonstrated That The Accuracy Rate Remains The Same.

Keywords: *Random Forest Classifier, Multinomial Naive Bayes, Naive Bayes, Stochastic Gradient Descent*

1. Introduction

The Basic Reason For Data And Information Fusion Is To Reduce The Uncertainty In Our Understanding And Prediction Of The State Of Certain Aspect Of The World. The Text Document Classification System Contains Four Different Levels Of Scope That Can Be Applied. The Document Level Is The First Level Where The Algorithms Are Used To Obtain The Relevance For Categorizing The Document. Where The Second Level Is The Paragraph Level, Where The Algorithms Are Used To Acquires The Relevance In Single Paragraph Are Categorized And Therefore Is Also Said That A Portion Of A Document Categorization. The Third Level Is The Sentence Level, Which Are Used To Find The Relevance In Single Sentence And Categorizing The Portion Of The Paragraph. The Fourth Level Is The Sub-Sentence Level, Where The Algorithms Are Used To Find The Relevance In Sub-Expressions In A Sentence.

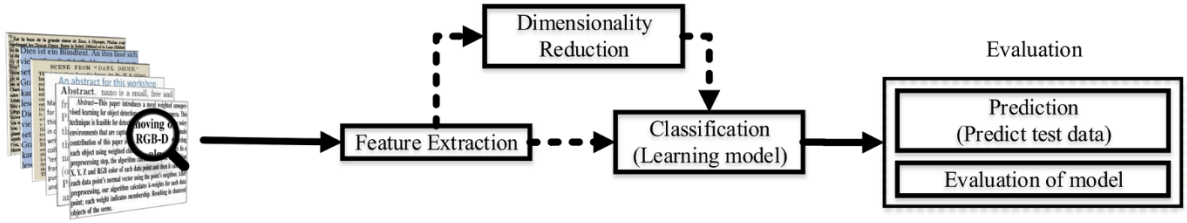

Fig1 System Architecture For Document Classifications.

1.1 The Category Variable To Find The Value Distribution Profile:

It Is More Complicated To Predict The Basic Statistics Of The Target Variable, Therefore To Ensure No Imbalance In Class Count The Instances Of Class Labels Are Counted.

ISSN: 2233-7857 IJFGCN

1.2 Text Data Analysed Using Lexical Analyzer:

In Pre-Processing Step, The Text Data Are Transformed To Columns And Therefore To Analyse The Richness Of The Vocabulary. By Using Lexical Analyser, The Uniqueness Of The Vocabulary Along With Data Rich Can Be Analysed.

2. Materials And Methods

3. 2.1 Boosting

An Ensemble Learning Is A Machine Learning Technique Which Is Used For Reducing The Variance In Supervised Learning. As Boosting Is A Strong Learning Algorithm It Is Used To Perform Well-Correlated And True Classification.

2.2 Naive Bayes Classifier

Naïve Bayes Technique Used For Text And Document Categorization, Where Naive Bayes Classifier (Nbc) Is A Generative Model That Can Widely Use For Information Retrieval. Which Developed By Using Term-Frequency (Bag Of Word) Feature Extraction Technique By Counting Number Of Words In Documents

2.3 Support Vector Machine (Svm)

The Svm Was Designed To Find Solution For Binary Classification Issues, It Have Been Used In Multi-Class Problem Based On Authoritative Technique.

2.4 Decision Tree

Decision Tree Algorithm For Successful Classification Of Text Document Are Structured In The Way Of Decomposition Of Data In Hierarchical. The Main Issues Is To Determine The Attributes Of Data Points For Tree Creation Which Leads Parent Level And That Will Be In Child Level.

2.5 Dataset

The Dataset Is Created By Using Jobs Requirement Posted On A Online Job Portal Which Comprised With The Title Of Jobs. Later The Data Are Collected And Labelled With The Corresponding Categories As Supervise Machine Learning Algorithm Is Going To Be Used. The Main Aim Is To Develop A Model By Training The Labelled Data That Can Perform Classification With More Accuracy. The Initial Steps Involved In Developing The Model Is Of Creating The Dataset, Cleaning The Data And Later Various Exploratory Analysis Are Applied For Preprocessing. As The Data Set Used To Have Many Features Which Leads To Multiple Dimensions. In Preprocessing The Dimensions Are Reduced By Applying Algorithm For Selecting The Required Features.

2.6 Naive Bayes:

Navie Bayes Algorithm Is One Of Classification Algorithm Widely In Text Mining Applications. It Helps In Prediction As It Uses The Bayes Theorem Therefore It Can Able To Make Assumption To Find The Class Based On The Probabilities Of Each Attribute. Navie Bayes Algorithm Comes In Two Form, Bernoulli And Multinominal. The Bernoulli Algorithm Is Designed For Classifying Boolean Or Binary Features, While Multinomial Algorithm Is Designed To Count The Occurrence Of Features.

2.7 Linear Svm

Support Vector Machine In A Linear Classification Algorithm Used For Classification Applications. For This Experiment The Svm Classifier Is Used Along With Python Libraries Such As Pandas, Numpy Sklearn Nltk Matplotlib[9] To Get Better Performance.

4. Experiment And Results

a. Experimental Setup

ISSN: 2233-7857 IJFGCN

Initially A Data-Frame Is Created To Load Data Into It. The Dataset With Header Row Is Also Created And Loaded Into The Data-Frame. As It Have Only Four Categories It Used To Have Four Colums And More Than 5000 Rows As A Dataset. Then In Pre-Processing Step The Missing Values Are Identified They Are Replaced With Nan Value. Therefore There Are Three Unique Approaches.
The Dataset Used To Have Three Unique Categories Such As Administrative, Which Is The Most Frequently Occurred Category. The Count Value Of This Category Determines That There Are No Missing Value Present In The Dataframe. If Missing Value Are Present Either It Can Be Removed Or It Can Be Replaced With Nan Value As Simpler Approach[15]. For This Experiment The Nan Value Is Used To Replace The Missing Value And It Ie Reloaded With The Added Parameters.

3.2 Data Pre-Processing:

The Raw Data Are Not Suitable For Developing The Analytical Model; Therefore, The Raw Data Are To Processed For Applying It For A Model. A Pre-Processing Step Is Required To Process The Raw Data. The Step Includes Regression, Classification Or Clustering, But Here Multiclass Classification Is Applied To Develop The Analytical Model [12]. Though Pre-Processing Depends On The Analytical Model. It Performs The Following Functions; First All The Text Data Are Converted To Lower Case. Second The Words Are Stemmed. So, It Reduces The Feature Size Automatically. In This Dataset Both The Attributes Title Of The Job And Job Description Are Used For The Classification [17]. Though It Can Able To Obtain Better Accuracy Rate.

3.3 Steps Involved In Automatic Document Classification

Step 1: Data Extraction

Initially The Data Are Generated Locally To Train The Proposed Model. The Function Namely Local-Data.Py Are Specified The Dates And The Key. This Code Will Create The Specified Dataset (Headers: Id,Text,Label) And Fix The Category And Labels File[15]. The Code Will Automatically Increment The Dates And Keep On Collecting The Data Specified By The Iterations. There Are Multiple Categories Such As News, Sport, Television & Radio, Environment, Science And Design, Global, Food, Culture, Community, Money And Technology.

Step 2: Create A Model Using The Extracted Data

Later The Dataset Is Uploaded And Created The Pipeline Model Using The Pipeline.Py. This File Will Save The Pipeline In A Folder And Will Show The Accuracy Of The Model By Using Dataset For 80% Training And 20% Testing.

Step 3: Run Stream_Producer.Py To Use The Extracted Model To Classify Streaming Data

Now The Stream_Producer.Py To Create For Stream Data And The Spark_Stream.Py File To Read The Stream And Load The Created Model And Generate The Prediction On The Fly. Please, Add The Saved Model Full Path In The Spark_Stream.Py File.

3.4Result Analysis.

ISSN: 2233-7857 IJFGCN

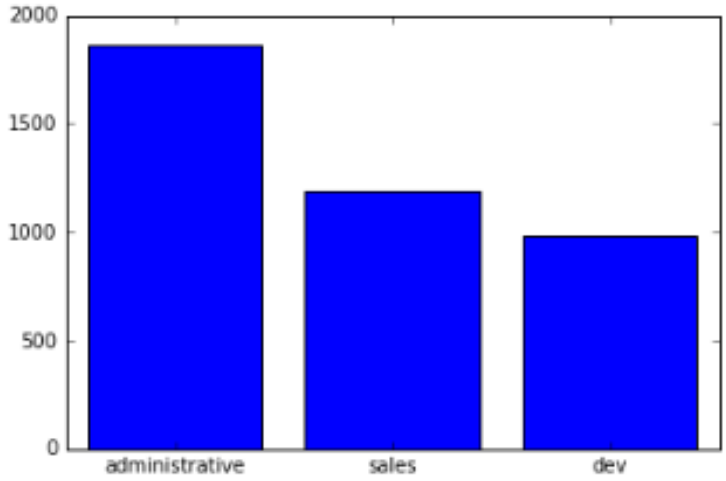

Fig 2: Relative Count Of Each Classes

The Graph Shows The Relative Count Value Of The Three Categories That Is Administrative, Sales And Development. The Difference In Count Of The Categories Are Relatively Equal It Is Considered To Be Balanced Dataset And It Can Apply As A Training Set For The Model Development. Imbalance Dataset Leads To Under-Sampling Or Over-Sampling During Training The Model. Therefore, To Minimize The Bias Rate Balanced Dataset Is To Be Used For Model Development.

A Navie Base Algorithm Approach Is Use For Classification. Bag Of Words Approach Is Relatively Better For Computing N Top Frequent Words. Eventhough It Has The Drawback Of Favouring These Terms While Repeated In Corpus But Any Way It Helps In Overall Classification Objective. Where Us The By Applying The Proposed Algorithm It Is Used Infact That Somehow The Occurance Of The Weight Or The Features Are Better In Computing The Weight Of The Occurrence Or The Feature Count. Therefore The Hypothesis That The Classifier Performance Is Improved And It Demonstrated The Better Accuracy Rate As A Result . That Is The Multi-Nominal Naive Bayes Obtains The Accuracy Rate 95 % Using Training Set And The Time Taken For Training Is 33 Sec Whereas The Test Set Obtaines The Accuracy Rate Of 93% And The Time Taken For Prediction Is 0.09 Sec, When Random Forest Classifier Is Applied With The Ensemble Learning Methods To Perform Classification And Regression Tasks. Though The Random Forests Have The Characteristic To Minimize Variance If Its There In The Data-Set. There Training Time Is Generally Quite Higher Which Is One Of Their Drawbacks To Be Used In Production Environment. However, We Can Apply Routine To Verify The Accuracy Of Random Forests Algorithm Using Different Number Of Trees A 50 Have Training Time As 6.059000s Where Accuracy Score As 0.9842.

ISSN: 2233-7857 IJFGCN

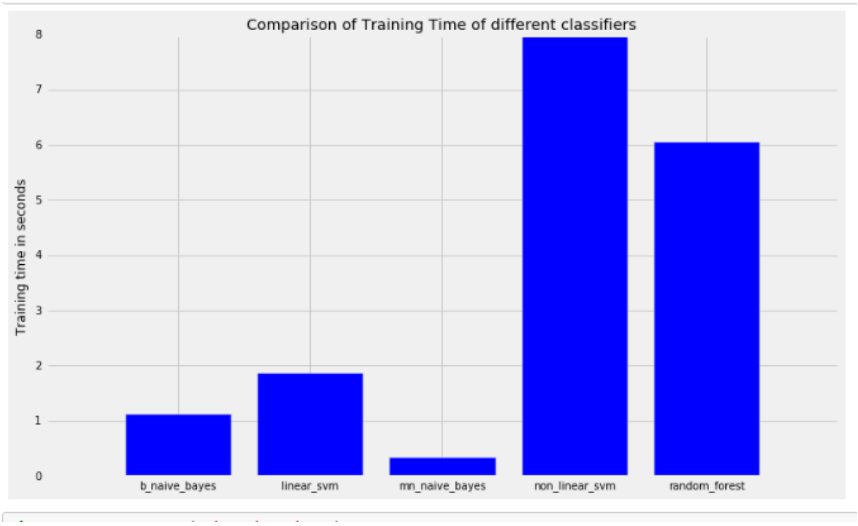

Fig 3: Comparison Of Training Time For Various Classifier

The Accuracy Rate Obtained By The Proposed Classifier Is Better But, The Time Taken For Training And Prediction Are Higher While Comparing The Time Taken By The Other Classifier.

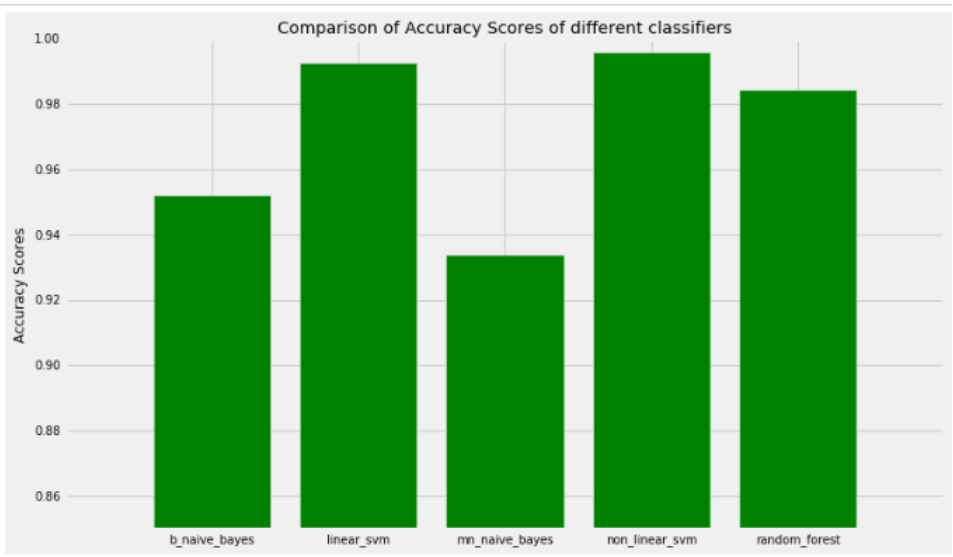

Fig 4: Comparison Of Accuracy For Various Classifier

The Ultimate Goal Is To Estimate The Performance Of The Predictive Model Accuracy Rate By Using Un-Seen Data To Reduce The Risk Of Overfitting.

5. Conclusion:

It Is Clearly Analysed From The Accuracy Rate That Linear Svm Classifier Demonstrated The Better Performance In Terms Of Obtaining Highest Accuracy Rate By Using Both Training Set And Testing Set. Whereas The Random Forest And Non Linear Svm Classifiers Has Obtained Better Accuracy Rate But The Time Taken For Training The Model And For Prediction Is Quite High While Comparing To Linear Svm. For This Experiment Three Classes Have Been Used Among That The "Administrative" Category Is The Most Frequent Category As Its Count Value Is Greater And Also It Confirms That There Are No Missing Values. The Accuracy Score Of Random Forest And Non

ISSN: 2233-7857 IJFGCN

Linear Svm Classifiers' Is Quite Appealing The Magnitude Is High When Considering The Training Time. The Chi-Square Test Is Applied To Select Less Number Of Features And The Algorithms Are Applied, And It Demonstrated That The Accuracy Rate Remains The Same.

6. References.

1. F.Kboubi, Etal., Table Recognition Evaluation And Combination Methods For Document Analysis And Recognition, 2005 Proceedings Eighth International Conference , Ieee (2005).
2. M.T. Luong,Etal, Logical Structure Recovery In Scholarly Articles With Rich Document Features, Multimedia Storage Retrieval Innovat. Digital Lib. Syst. (2012).
3. H.Chao,Eta, Fan Layout And Content Extraction For Pdf Documents, Document Analysis Systems Vi, Springer (2004).
4. F.Boudin,J.Y. Nie,Etal, Combining Classifiers For Robust Pico Element Detection, Bmc Med. Inform. Decis. Mak., 10 (2010).
5. A.Constantin, S.Pettifer, Etal, Pdfx Fully-Automated Pdf-To-Xml Conversion Of Scientific Literature, Proceedings Of The 2013 Acm Symposium On Document Engineering, Acm (2013).
6. K.G.Shojania,Etal, How Quickly Do Systematic Reviews Go Out Of Date? A Survival Analysis Ann. Intern. Med., 147 (4) (2007).
7. P.Bragge, O.Clavisi,Et.Al. Gruen The Global Evidence Mapping Initiative: Scoping Research In Broad Topic Areas, Bmc Med. Res. Methodol., 11 (1) (2011).
8. J.P.Higgins,Etal., Cochrane Handbook For Systematic Reviews Of Interventions Wiley Online Library (2008).
9. D.D.Bui, S. Jonnalagadda, G. Del Fiol, Automatically Finding Relevant Citations For Clinical Guideline Development, J. Biomed. Inform.(2015)
10. R.L.Summerscales, Automatic Summarization Of Clinical Abstracts For Evidence-Based Medicine, Illinois Institute Of Technology (2013)
11. K.-C.Huang, I.J. Chiang, F. Xiao, C.-C. Liao, C.C.-H. Liu, J.-M. Wong, Pico Element Detection In Medical Text Without Metadata: Are First Sentences Enough?J. Biomed. Inform., 46 (5) (2013)
12. H.Han, C.L. Giles, E. Manavoglu, H. Zha, Z. Zhang, E. Fox (Eds.), Automatic Document Metadata Extraction Using Support Vector Machines, Digital Libraries 2003 Proceedings 2003 Joint Conference On, Ieee (2003)
13. H.Zhu, Y.Ni, P.Cai, Z.Qiu, F.Cao, Automatic Extracting Of Patient-Related Attributes: Disease, Age, Gender And Race, Stud. Health Technol. Inform., 180 (2012).
14. D.P.A. Corney,Etal., "Biorat: Extracting Biological Information From Full-Length Papers", Bioinformatics (Oxford, England), 20 (17) (2004).
15. Verspoor K, Mackinlay A, Cohn Jd, Wall Me. Detection Of Protein Catalytic Sites In The Biomedical Literature. In: Pac Symp Biocomput.(2013).
16. J.Hakenberg,Et Al., Efficient Extraction Of Protein–Protein Interactions From Full-Text Articles, Ieee/Acm Trans. Comput. Biol. Bioinform., 7 (3) (2010).
17. W.Hsu, W. Speier, R.K. Taira, Automated Extraction Of Reported Statistical Analyses: Towards A Logical Representation Of Clinical Trial Literature, In: Amia Annual Symposium Proceedings/Amia Symposium Amia Symposium (2012).
18. B.De Bruijn,Etal., Automated Information Extraction Of Key Trial Design Elements From Clinical Trial Publications, In: Amia Annual Symposium Proceedings/Amia Symposium Amia Symposium.(2008).
19. S.Kiritchenko, Etal., I. Sim Exact: Automatic Extraction Of Clinical Trial Characteristics From Journal Publications, Bmc Med. Inform. Decis. Mak., 10 (2010).R. Kern, K. Jack, M. Hristakeva, M. Granitzer, Teambeam Meta-Data Extraction From Scientific Literature, D-Lib Magazine, 18 (7) (2012).

ISSN: 2233-7857 IJFGCN

20. M. Granitzer, M. Hristakeva, R. Knight, K. Jack, R. Kern (Eds.), A Comparison Of Layout Based Bibliographic Metadata Extraction Techniques, Proceedings Of The 2nd International Conference On Web Intelligence, Mining And Semantics, Acm (2012).

ISSN: 2233-7857 IJFGCN