Efficient Text Classification of 20 Newsgroup Dataset using Classification Algorithm

Karishma Borkar1, Prof. Nutan Dhande2
1 Department of CSE,ACE Nagthana,Maharastra,India
2 Department of CSE,ACE Nagthana,Maharastra,India

Abstract: Text classification is the undertaking of naturally sorting an arrangement of archives into classifications from a predefined set. Content Classification is an information mining procedure used to anticipate bunch enrollment for information occurrences inside a given dataset. It is utilized for ordering information into various classes by thinking of some as compels. Rather than conventional component determination systems utilized for content archive grouping. We present another model in view of likelihood and over all class recurrence of term. The Naive Bayesian classifier depends on Bayes hypothesis with autonomy presumptions between indicators. A Naive Bayesian model is anything but difficult to work, with no confounded iterative parameter estimation which makes it especially valuable for substantial datasets. The paper demonstrates that the new probabilistic translation of tf×idf term weighting may prompt better comprehension of measurable positioning instruments.
.

Keywords: *Text classification, Documents classification, Modified Naïve Bayes.*

*****

I. INTRODUCTION

Information mining, the extraction of concealed prescient data from extensive databases, is a capable new innovation with awesome potential to help organizations concentrate on the most essential data in their information stockrooms. Information mining devices anticipate future patterns and practices, permitting organizations to make proactive, learning driven choices. Information mining devices can answer business addresses that generally were excessively tedious, making it impossible to determine. They secure databases for concealed examples, finding prescient data that specialists may miss since it lies outside their desires.

Content mining, now and again on the other hand alluded to as content information mining, generally proportionate to content examination, alludes to the way toward getting top notch data from text.Text mining normally includes the way toward organizing the information content (typically parsing, alongside the expansion of some inferred phonetic components and the expulsion of others, and consequent addition into a database), determining designs inside the organized information, lastly assessment and translation of the yield. Run of the mill content mining assignments incorporate content order, content bunching, idea/substance extraction, archive outline, and element connection demonstrating (i.e., learning relations between named elements).

Innocent Bayes has been one of the well known machine learning strategies for a long time. Its effortlessness makes the structure appealing. Consequently, there likewise have been many intriguing works of exploring guileless Bayes. Particularly, that credulous Bayes can perform shockingly well in the arrangement undertakings where the likelihood itself figured by the guileless Bayes is not essential. With this foundation, content classifiers in light of gullible Bayes have been contemplated widely by a few specialists. In their credulous Bayes classifiers, a report is considered as a parallel element vector speaking to whether each word is available or absent.

II. RELATED WORK

Bo Tang, Student Member, IEEE, Steven Kay, Fellow, IEEE, and Haibo He, Senior Member, IEEE: Feb 2016 , Proposed [1] Automated element determination is imperative for text arrangement to lessen highlight survey and to speed learning procedure of classifiers. In this paper, we display a novel and productive element determination system in view of the Information Theory, which plans to rank the components with their discriminative limit with respect to grouping. We first return to two data measures: Kullback-Leibler difference and Jeffreys disparity for paired theory testing, and break down their asymptotic properties identifying with sort I and sort II mistakes of a Bayesian classifier. We then present another disparity measure, called Jeffreys-Multi-Hypothesis (JMH) dissimilarity, to quantify multi-dissemination difference for multi-class arrangement. In light of the JMH-uniqueness, we create two productive element determination techniques, named most extreme separation (MD) and MD_ x2 strategies, for text arrangement. The promising consequences of broad examinations show the adequacy of the proposed approaches.

OmkarArdhapure, GayatriPatil, DishaUdani, KamleshJetha, Proposed [2] Text order is a procedure in information mining which appoints predefined classes to free-text archives utilizing machine learning methods. Any report as text, picture, music, and so on can be grouped utilizing some order procedures. It gives calculated perspectives of the gathered records and has vital applications in this present reality. Text based arrangement is made utilization of for archive order with example acknowledgment and machine learning. Points of interest of various grouping calculations have been considered in this paper to arrange records. A case of these calculations is: Naive Bayes' calculation, K-Nearest Neighbor, Decision Tree and so on. This paper exhibits a similar investigation of favorable circumstances and inconveniences of the previously mentioned arrangement calculation.

Aaditya Jain, JyotiMandowara, Proposed [3] Text order or report grouping is one of the significant undertakings in text information mining and data recovery. Numerous proficient classifiers for text characterization have been proposed till date. Be that as it may, the individual classifiers demonstrate constrained appropriateness as indicated by their separate spaces and extensions. Late research works assessed that the blend of classifiers when utilized for grouping indicated preferred execution over the individual ones. Our work gives portrayal about text characterization prepare and related mainstream classifiers. In this paper, the quantities of methodologies managing joining text classifiers for enhancing the effectiveness in the field of text order are additionally overviewed.

Adel Hamdan Mohammad, Omar Al-Momani and Tariq Alwada'n, Proposed [4] No uncertainty that text order is a critical research zone in data recovery. Actually there are many examines about text characterization in English dialect. A couple of analysts by and large discuss text order utilizing Arabic informational collection. This exploration applies three understood arrangement calculation. Calculation connected are Key Nearest neighbor (K-NN), C4.5 and Rocchio calculation. These notable calculations are connected on in-house gathered Arabic informational collection. Informational collection utilized comprises from 1400 records has a place with 8 classes. Comes about demonstrate that exactness and review values utilizing Rocchio classifier and K-NN are superior to C4.5. This examination makes a relative review between said calculations. Additionally this review utilized a settled number of records for all classes of reports in preparing and testing stage.

Kapila Rani, Satvika, Proposed [5] Text classification can be quickly depicted as the automatization of the report association procedure to an arrangement of pre-characterized classifications. Programmed Text Classification is an essential application and research subject for the distinguishing proof of computerized reports. A text order framework is utilized to list the records for the data recovery assignments and to the grouping of reminders, messages or website pages. Text Classification speaks to the high dimensionality of the element space. The Text Classification is utilized to allocate the classification marks to the new reports at the preparation arrange which depend on the learning picked up in a grouping framework. In the preparation stage, an arrangement framework is fabricated utilizing a learning strategy and an arrangement of archives which are given, appended with class marks, machine learning groups.

III. PROPOSED SYSTEM

Naïve Bayes Classifiers are simple probabilistic classifiers based on the Bayes Theorem [5]. These are highly scalable classifiers involves a family of algorithms based on a common principle assuming that the value of a particular feature is independent of the value of any other feature, given the class variable. In practice, the independence assumption is often violated, but Naive Bayes classifiers still tend to perform very well under this unrealistic assumption and very popular till date.

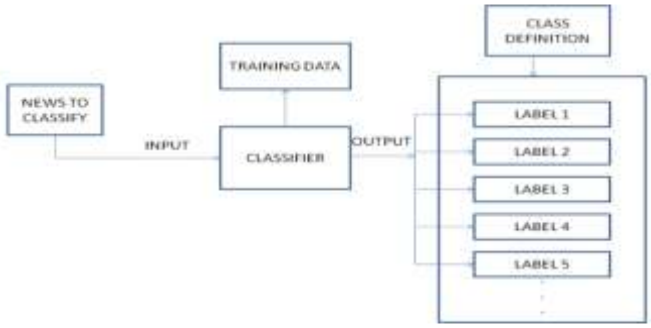


Fig- System Architecture

From the above figure, you can see the framework design plainly. Framework comprise of new to be named its information and the yield will be the name to which the news likely has a place with. Classifier is the fundamental module of the framework which is the execution of the guileless bayes calculation. It utilizes the preparation information as its information and characterizes the information records. Preparing information comprise of substantial number of records preprocessed i.e. term recurrence and record recurrence is figured. Utilizing this information the information record is arranged.

Order methods can deal with preparing of largevolume of information. It can foresee straight out class names and characterizes information in light of model worked by utilizing preparing set and related class names and after that can be utilized for arranging recently accessible test information. In this way, it is delineated as a vital piece of information investigation and is increasing greater ubiquity. Characterization utilizes directed learning approach. In regulated taking in, a preparation dataset of records is accessible with related class marks. Characterization process is partitioned into two fundamental strides. The first is the preparation step where the grouping model is constructed. The second is simply the arrangement, in which the prepared model is connected to allot obscure information protest one out of a given arrangement of class mark. This paper concentrates on an overview of different arrangement systems that are most generally utilized as a part of information mining. The near review between various calculations (Bayesian system) is utilized to demonstrate the quality and exactness of every grouping calculation in term of execution effectiveness and time multifaceted nature. A similar review would draw out the points of interest and impediments of one technique over the other. This would give the rule to fascinating exploration issues which thusly help different analysts in creating imaginative calculations for applications or necessities which are not accessible.

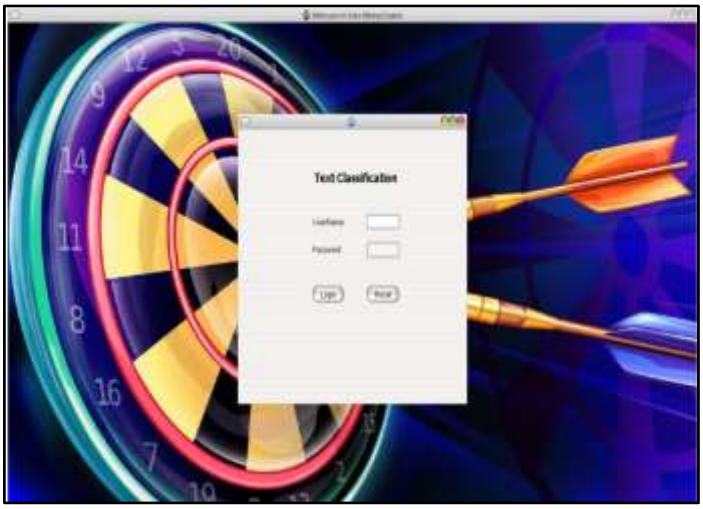
Algorithm MNB(Buffer)

```
Step 1:
Read data from buffer into temp array
Step 2:
Preprocess array and remove stop words and unwanted
special symbols and spaces

Step 3:For I =0 to N
    For each word in temp as j do
        Fetch avgtf*idf from database and place in
        decisionmatrix[i][j]
    End for
End for
Step 4:
For I =0 to N
    For j=0 to words in temp
        Sum[i] = sum[i] * decisionmatrix[i] [j]
    End forEnd for
Step 5:
Calculate index of the max value in sum[] as index
Step 6:
Return index
```

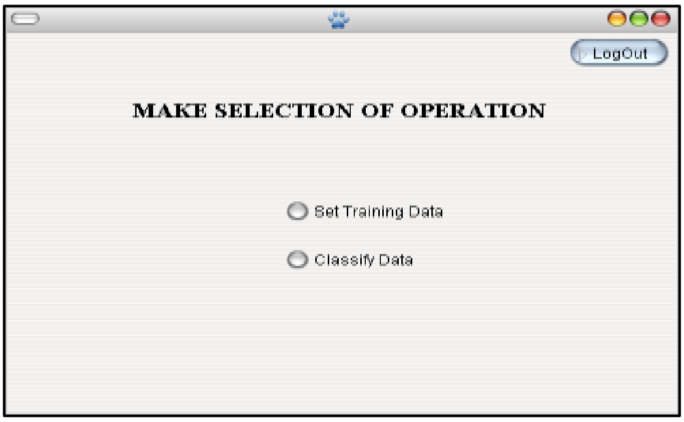
The generic strategy for text classification is depicted in Fig 1.The main steps involved are i) document pre-processing, ii)feature extraction / selection, iii) model selection, iv) training and testing the classifier. Data pre-processing reduces the size of the input text documents significantly. It involves activities like sentence boundary determination [2], natural language specific stop-word elimination [1] [2] [3] and stemming [2] [4]. Stop-words are functional words which occur frequently in the language of the text (for example, „a‟, ‟the‟, ‟an‟, ‟of‟ etc. in English language), so that they are not useful for classification. Stemming is the action of reducing words to their root or base form. For English language, the Porter‟s stemmer is a popular algorithm [4] [12], which is a suffix stripping sequence of systematic steps for stemming an English word, reducing the vocabulary of the training text by approximately one-third of its original size [4]. For example, using the Porters stemmer, the English word "generalizations" would subsequently be stemmed as "generalizations → generalization → generalize → general → gener". In cases where the source documents are web pages, additional pre-processing is required to remove / modify HTML and other script tags [13]. Feature extraction / selection helps identify important words in a text document. This is done using methods like TF-IDF (term frequency-inverse document frequency) [14], LSI (latent semantic indexing) [15], multi-word [2][16] etc. In the context of text classification, features or attributes usually mean significant words, multi-words or frequently occurring phrases indicative of the text category. After feature selection, the text document is represented as a document vector, and an appropriate machine learning algorithm is used to train the text classifier. The trained classifier is tested using a test set of text documents. If the classification accuracy of the trained classifier is found to be acceptable for the test set, then this model is used to classify new instances of text documents.

Simulation Screenshots



Snapshot 1: login page

This is the authentication page of our project. In this stage we provide the username and password to the system if the username and password is correct then system permits the log in to the user.



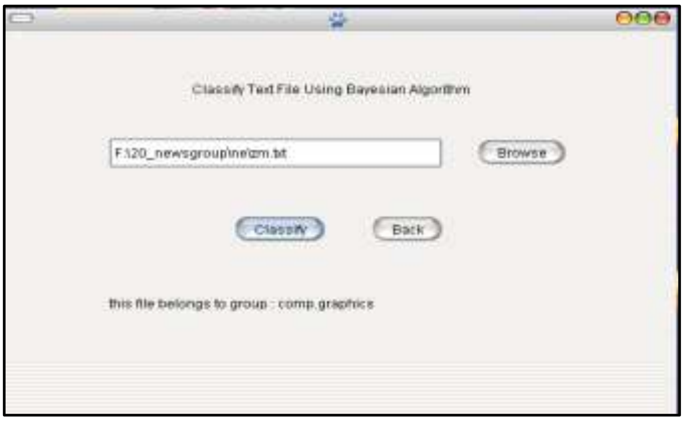
MAKE SELECTION OF OPERATION

○ Set Training Data

○ Classify Data

Snapshot 2: Training andTesting module

This training and the testing module gives the option to user to make selection between the training data and classifying it. Set training data can be used to set any one file into the dataset after preprocessing TF iDf calculation and other means of processing and frequency generation. Classify data can be used to test whether the algorithm is working properly or not.



Classify Text File Using Bayesian Algorithm

F:\20_newsgroup\nelzm.txt

this file belongs to group : comp.graphics

Snap shot 3: classifying text

Above snapshot shown the working of the simulation where one file can be randomly selected from the directories and classify it using Naive Bayes Algorithm.

5.CONCLUSION

The Text Classification using analytical approach project proposed a design of the application that can effectively classify text files into appropriate folder depending upon the theme of the file, using the training data to model the classifier. This application automates the text classification process otherwise would take long time doing manually the same task. Text file are appropriately classified using this application. This application allows you to select the test data, training data. In the future, a similar concept can be used for different purposes like arrange your computer, classify various documents with various applications and analyze them.

6.FUTURE SCOPE

Accuracy analysis is the how perfectly does the text gets classified. Currently we are using 18,000 files as our training data due to this sometimes the file may not be classified correctly. Hence in future by increasing the amount of training data accuracy can be increased or else by using some other technique accuracy can be increased.

References

[1] Jiawei Han and MichelineKamber “Data Mining Concepts And Techniques” ,Morgan kaufman publishers, San Francisco, Elsevier, 2011, pp. 285-351

[2] M.Sukanyal, S.Biruntha2 "Techniques on Text Mining" International Conference on Advanced Communication Control and Computing Technologies, IEEE-2012

[3] Sonali Vijay Gaikwad, ArchanaChaugule, PramodPatil "Text Mining Methods and Techniques"International Journal of Computer Applications (0975 – 8887) Volume 85 – No 17, January 2014

[4] Nidhi, Vishal Gupta "Recent Trends in Text Classification Techniques" International Journal of Computer Applications (0975 – 8887) Volume 35– No.6, December 2011

[5] S. Subbaiah "Extracting Knowledge using Probabilistic Classifier for Text Mining" Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22, IEEE-2013

[6] M. JanakiMeena , K. R. Chandran "Naive Bayes Text Classification with Positive Features Selected by Statistical Method" ©2009 IEEE vaishaliBhujade, N.J.Janwe "knowledge discovery in text mining techniques using association rule extraction" International Conference on Computational Intelligence and Communication Systems, IEEE- 2011

[7] Zhou Faguo, Zhang Fan "Research on Short Text Classification Algorithm Based on Statistics and Rules" 2010 Third International Symposium on Electronic Commerce and Security © 2010 IEEE

[8] Shuzlina Abdul-Rahman, SofianitaMutalib, Nur Amira Khanafi, AzlizaMohd Ali "Exploring Feature Selection and Support Vector Machine in Text Categorization" 16th International Conference on Computational Science and Engineering, IEEE-2013

[9] Xianfei Zhang, Bicheng Li, Xianzhu Sun "A k-Nearest Neighbor Text Classification algorithm Based on Fuzzy Integral" Sixth International Conference on Natural Computation, IEEE-2010

[10] Liu T., Chen Z., Zhang B., Ma W., and Wu G. 2004.” Improving text classification using local latent semantic indexing”. In proceedings of the 4th IEEE international conference on Data Mining , pp. 162-169.

[11] M. M. SaadMissen, and M. Boughanem. 2009. Using WordNet ‟semantic relations for opinion detection in

blogs". ECIR 2009, LNCS 5478, pp. 729-733, Springer Verlag Berlin Heidelberg.

[12] Balahur A., and MontoyoA.. 2008. "A feature dependent method for opinion mining and classification". In proceedings of the IEEE international conference on Natural Language Processing and Knowledge Engineering, pp. 1-7.

[13] Zhao L., and Li C.. 2009. "Ontology based opinion mining for movie reviews". KSEM 2009, LNAI 5914, pp. 204-214, Springer-Verlag Berlin Heidelberg.

[14] Durant K. T., Smith M. D. 2006. "Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection".

[15] WebKDD 2006, LNAI 4811, pp. 187-206, Springer-Verlag Berlin Heidelberg.

[16] Polpinij J., and Ghose A. K. 2008. "An ontology-based sentiment classification methodology for online consumer reviews". In proceedings of the IEEE international conference on Web Intelligence and Intelligent Agent Technology, pp. 518-524.

[17] HeideBrücher, Gerhard Knolmayer, Marc-André Mittermayer; "Document Classification Methods for Organizing Explicit Knowledge", Research Group Information Engineering, Institute of Information Systems, University of Bern, Engehaldenstrasse 8, CH - 3012 Bern, Switzerland. 2002.

[18] Andrew McCallum, Kamal Nigam; "A Comparison of Event Models for Naïve Bayes Text Classification", Journal of Machine Learning Research 3, pp. 1265-1287. 2003.

[19] Irina Rish; "An Empirical Study of the Naïve Bayes Classifier", In Proceedings of the IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence. 2001.

[20] Irina Rish, Joseph Hellerstein, JayramThathachar; "An Analysis of Data Characteristics that affect Naïve Bayes Performance", IBM T.J. Watson Research Center 30 Saw Mill River Road, Hawthorne, NY 10532, USA. 2001.

[21] Pedro Domingos, Michael Pazzani; "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, Machine Learning", Vol. 29, No. 2-3, pp.103-130. 1997.

[22] Sang-Bum Kim, Hue-Chang Rim, Dong-Suk Yook, Huei-Seok Lim; "Effective Methods for Improving Naïve Bayes Text Classification", 7th Pacific Rim International Conference on Artificial Intelligence, Vol. 2417. 2002.

[23] Susana Eyheramendy, Alexander Genkin, Wen-Hua Ju, David D. Lewis, and David Madigan; "Sparce Bayesian Classifiers for Text Categorization", Department of Statistics, RutgersUniversity.2003.

[24] Miguel E. Ruiz, Padmini Srinivasan; "Automatic Text Categorization Using Neural Network",In Proceedings of the 8th ASIS SIG/CR Workshop on Classification Research, pp. 59-72. 1998.

[25] Petri Myllymaki, Henry Tirri; "Bayesian Case-Based Reasoning with Neural Network", In Proceeding of the IEEE International Conference on Neural Network'93, Vol. 1, pp. 422-427. 1993.

[26] Y. Aphinyanaphongs, L. D. Fu, Z. Li, E. R. Peskin, E. Efstathiadis, C. F. Aliferis, and A. Statnikov, "A comprehensive empirical comparison of modern supervised classification and feature selection methods for text categorization," Journal of the Association for Information Science and Technology, vol. 65, no. 10, pp. 1964–1987, 2014.