Document Classification using Naïve Bayes Algorithm

Minal Zope (M.E.Computer)[1], Kiran Khade[2], Pratik More[3], Pratik Kamthe[4], Anuj Manekar[5]
Department of Computer Science
All India Shree Shivaji Memorial Society's Institute of Information Technology

Abstract **-** Every day the mass of information available, merely finding the relevant information is not the only task of automatic text classification systems. Instead the automatic text classification systems are supposed to retrieve the relevant information as well as organize according to its degree of relevancy with the given query. The main problem in organizing is to classify which documents are relevant and which are irrelevant. The Automated text classification consists of automatically organizing clustered data. We propose an automatic method of text classification using machine learning based on the disambiguation of the meaning of the word we use the word net to eliminate the ambiguity of words so that each word is replaced by its meaning in context. The closest ancestors of the senses of all the undamaged words in a given document are selected as classes for the specified document.

Keywords **-** Document clustering, feature selection, model selection, machine learning

I. INTRODUCTION

Every day the mass of information available to us increases. This information would be irrelevant if our ability to productively get to did not increment too. For most extreme advantage, there is need of devices that permit look, sort, list, store and investigate the accessible information. One of the promising region is the automatic text categorization. Envision ourselves within the sight of impressive number of texts, which are all the more effectively available on the off chance that they are composed into classes as per their topic. Obviously one could request that human read the text and arrange them physically. This assignment is hard if done on hundreds, even a huge number of texts. Thus, it appears to be important to have a computerized application, so here automatic text categorization is presented. An increasing number of data mining applications involve the analysis of complex and structured types of data and require the use of expressive pattern languages. Many of these applications cannot be solved using traditional data mining algorithms. This observation forms the main motivation for the Linear Regression.

Unfortunately, existing "upgrading" approaches, especially those using Logic Programming techniques, often suffer not only from poor scalability when dealing with complex database schemas but also from unsatisfactory predictive performance while handling noisy or numeric values in real-world applications. However, "flattening" strategies tend to require considerable time and effort for the data transformation, result in losing the compact representations of the normalized databases, and produce an extremely large table with huge number of additional attributes and numerous NULL values (missing values). As a result, these difficulties have prevented a wider application of multi relational mining, and post an urgent challenge to the data mining community. To address the above mentioned problems, this article introduces a Descriptive clustering approach where neither "upgrading" nor "flattening" is required to bridge the gap between propositional learning algorithms and relational.

In Proposed approach, Data analysis techniques, such as clustering it can be used to identify subsets of data instances with common characteristics. Users can explore the data by examining some instances in each group instead of rather than examining the instances of the complete data set. This allows users to focus efficiently on large relevant subsets Data sets, in particular for document collections. In particular, the descriptive grouping consists of automatic grouping sets of similar instances in clusters and automatically generates a description or a synthesis that can be interpreted by man for each group. The description of each cluster allows a user determine the relevance of the group without having to examine its content For text documents, a description suitable for each group can be a multi-word tag, an extracted title or a list of characteristic words . The quality of the grouping it is important, so that it is aligned with the idea of likeness of the user, but it is equally important to provide a user with a brief and informative summary that accurately reflects the contents of the cluster.

II. RELATED WORK

Literature survey is the most important step in any kind of research. Before start developing we need to study the previous papers of our domain which we are working and on the basis of study we can predict or generate the drawback and start working with the reference of previous papers.

In this section, we briefly review the related work on Text classification and their different techniques.

J.-T. Chien, describe the "Hierarchical theme and topic modeling," in that Taking into account hierarchical data sets in the body of text, such as words, phrases and documents, we perform structural learning and we deduce latent themes and themes for sentences and words from a collection of documents, respectively. The relationship between arguments and arguments in different data groupings is explored through an unsupervised procedure without limiting the number of clusters. A tree branching process is presented to draw the proportions of the topic for different phrases. They build a hierarchical theme and a thematic model, which flexibly represents heterogeneous documents

using non-parametric Bayesian parameters. The thematic phrases and the thematic words are extracted. In the experiments, the proposed method is evaluated as effective for the construction of a semantic tree structure for the corresponding sentences and words. The superiority of the use of the tree model for the selection of expressive phrases for the summary of documents is illustrated [1].

Bernardini, C. Carpineto, and M. D'Amico, describe the "Full-subtopic retrieval with keyphrase-based search results clustering," in that Consider the problem of restoring multiple documents that are relevant to the individual sub-topics of a given Web query, called "full child retrieval". To solve this problem, they present a new algorithm for grouping search results that generates clusters labelled with key phrases. The key phrases are extracted generalized suffix tree created by the search results and merge through a hierarchical agglomeration procedure improved grouping. They also introduce a new measure to evaluate the performance of full recovery sub-themes, namely "look for secondary arguments length under the sufficiency of k documents". they have used a test collection specifically designed to evaluate the recovery of the sub-themes, they have found that our algorithm has passed both other clustering algorithms of existing research results as a method of redirecting search results underline the diversity of results (at least for k> 1, that is when they are interested in recovering more than one relevant document by sub-theme) [2].

T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela, describe the "Self-organization of a massive document collection," this paper describes the implementation of a system that can organize large collections of documents based on textual similarities. It is based on the self-organized map (SOM) algorithm. Like the feature vectors for documents, the statistical representations of their vocabularies are used. The main objective of our work was to resize the SOM algorithm in order to handle large amounts of high-dimensional data. In a practical experiment, they mapped 6 840 568 patent abstracts in a SOM of 1.002.240 nodes. As characteristic vectors, we use vectors of 500 stochastic figures obtained as random projections of histograms of weighted words [3].

K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, and R. Krishnapuram, describe the "A hierarchical monothetic document clustering algorithm for summarization and browsing search results," in that Organizing Web search results in a hierarchy of topics and secondary topics makes it easy to explore the collection and position the results of interest. In this paper, they propose a new hierarchical monarchic grouping algorithm to construct a hierarchy of topics for a collection of search results retrieved in response to a query. At all levels of the hierarchy, the new algorithm progressively identifies problems in order to maximize coverage and maintain the distinctiveness of the topics. They refer to the algorithm proposed as Discover. The evaluation of the quality of a hierarchy of subjects is not a trivial task, the last test is the user's judgment. They have used various objective measures, such as coverage and application time for an empirical comparison of the proposed algorithm with two other monotetic grouping algorithms to demonstrate its superiority. Although our algorithm is a bit more computationally than one of the algorithms, it generates better hierarchies. Our user studies also show that the proposed algorithm is superior to other algorithms as a tool for summary and navigation [4].

R. Xu and D. Wunsch, describe the "Survey of clustering algorithms," in that Data analysis plays an indispensable role in understanding the various phenomena. Conglomerate analysis, primitive exploration with little or no previous knowledge, consists of research developed in a wide variety of communities. Diversity, on the one hand, provides us with many tools. On the other hand, the profusion of options causes confusion. They have examined the grouping algorithms for the data sets that appear in statistics, computer science and machine learning and they illustrate their applications in some reference datasets, the problem of street vendors and bioinformatics, and a new field that attracts intense efforts. Various closely related topics, proximity measurement and cluster validation are also discussed [5].

S. Dumais, J. Platt, D. Heckerman, and M. Sahami, describe the "Inductive learning algorithms and representations for text categorization," in that Text categorization the assignment of natural language texts to one or more predefined categories based on their content is an important component in many information organization and management tasks. They compare the effectiveness of five different automatic learning algorithms for text categorization in terms of learning speed, real-time classification speed and classification accuracy. They also examine training set size, and alternative document representations. Very accurate text classifiers can be learned automatically from training examples. Linear Support Vector Machines (SVM) are particularly promising because they are very accurate, quick to train and quick to evaluate [6].

R. Kohavi and G. H. John, describe the "Wrappers for feature subset selection, "In that the feature subset selection problem, a learning algorithm is faced with the problem of selecting a relevant subset of features upon which to focus its attention, while ignoring the rest. To achieve the best possible performance with a particular learning algorithm on a particular training set, a feature subset selection method should consider how the algorithm and the training set interact. They explore the relation between optimal feature subset selection and relevance. Our wrapper method searches for an optimal feature subset tailored to a particular algorithm and a domain. They study the strengths and weaknesses of the wrapper approach and show a series of improved designs. They compare the wrapper approach to induction without feature subset selection and to Relief, a filter approach to feature subset selection. Significant improvement in accuracy is achieved for some datasets for

the two families of induction algorithms used: decision trees and Naive-Bayes [7].

T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela, describe the "Self-organization of a massive document collection," This paper describes the implementation of a system that is able to organize vast document collections according to textual similarities. It is based on the self-organizing map (SOM) algorithm. As the feature vectors for the documents statistical representations of their vocabularies are used. The main goal in our work has been to scale up the SOM algorithm to be able to deal with large amounts of high-dimensional data. In a practical experiment we mapped 6 840 568 patent abstracts onto a 1 002 240-node SOM. As the feature vectors we used 500-dimensional vectors of stochastic figures obtained as random projections of weighted word histograms [8].

Q. Mei, X. Shen, and C. Zhai, describe the "Automatic labeling of multinomial topic models," In this paper, they propose probabilistic approaches to automatically labelling multinomial topic models in an objective way. They cast this labelling problem as an optimization problem involving minimizing Kullback-Leibler divergence between word distributions and maximizing mutual information between a label and a topic model. Experiments with user study have been done on two text data sets with different genres. The results show that the proposed labelling methods are quite effective to generate labels that are meaningful and useful for interpreting the discovered topic models. Our methods are general and can be applied to labelling topics learned through all kinds of topic models such as PLSA, LDA, and their variations [9].

K. Lagus and S. Kaski, describe the "Keyword selection method for characterizing text document maps," in that Characterization of subsets of data is a recurring problem in data mining. They propose a keyword selection method that can be used for obtaining characterizations of clusters of data whenever textual descriptions can be associated, with the data. Several methods that cluster data sets or form projections of data provide an order or distance measure of the clusters. If such an ordering of the clusters exists or can be deduced, the method utilizes the order to improve the characterizations. The proposed method may be applied, for example, to characterizing graphical displays of collections of data ordered e.g. with the SOM algorithm. The method is validated using a collection of 10,000 scientific abstracts from the INSPEC database organized on a WEBSOM document map [10].

III. EXISTING APPROACH

Lot of work has been done in this field because of its extensive usage and applications. In this section, some of the approaches which have been implemented to achieve the same purpose are mentioned. These works are majorly differentiated by the algorithm for Text Classification.

In another research, to access the relevant information from mass of data is very difficult and time consuming task as every day mass of information increases because of digital world. Every day, the mass of information available to us increases. This information would be irrelevant if our ability to efficiently access did not increase as well. Automated text classification provides us with maximum benefit that allow us to search, sort, index, store, and analyze the available data. It also allows us to find in desired information in a reasonable time.

As my point of view when I studied the papers the issues are related to Text Classification. The challenge is to addressing automatic text classification problem using machine learning algorithms.

IV. PROPOSED APPROACHES

In Proposed System training is creation of train data set using which classification of unknown data in predefined categories is done. Here a learning system is created using regression. It is a supervised learning where unlabeled data is classified using labelled data. Training data is always a labelled dataset based on its features.\\

Project had considered no of scientific papers form different publication of different domains for creating training dataset. These papers are input for creating training dataset. This input is first pre-processed and most informative features are extracted using TF/IDF algorithm. Ten different domains from market are identified and then extracted feature and have to put to corresponding domain where each domain is considered as one class that which is used for labeling test dataset in testing part and features are considered as nodes. Once training part is completed, all features of respective domains are get updated in corresponding tables in database.

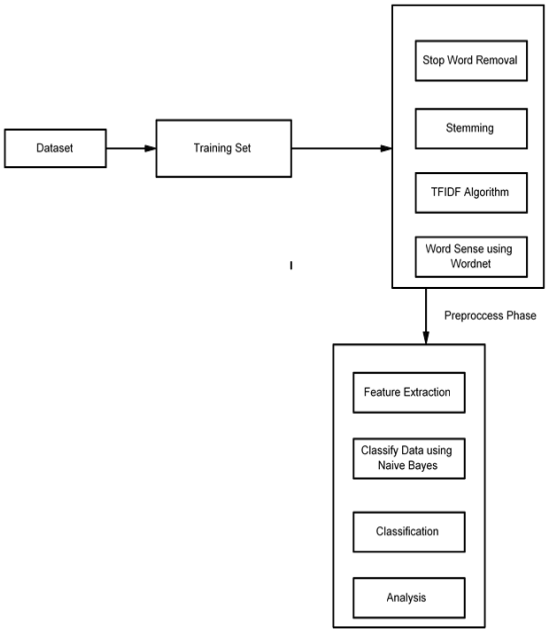

Figure 1: System architecture

Algorithms

Preprocessing Algorithms:

Stop word Removal-This technique remove stop words like is, are,they,but etc.
Tokenization-This technique remove Special character and symbols.
Stemming remove suffix and prefix and Find Orignal word

TFIDF Algorithm:

The tfidf score for term Iij is calculated using the term frequency and the document frequency. Term frequency (tf) and inverse document frequency (idf) are the foundations of the most popular term weighting scheme in IR. The tfidf score of Iij , tfidf(Iij).

Semantic Score Calculation:

In semantic Calculation, the keywords that are selected from the preprocessing techniques are applied to the word net ontology to extract the semantic relation of every keyword. A synonym is a word, which can be used to substitute another word without a change in the meaning of the words based on the semantic processing, unique keywords are selected in association with the extracted synonym words. The semantic processing is the effective way of text classification with robustness, reliability, and effectiveness. The organizational diagram of the semantic keyword processing

Classification

- Given training dataset D which consists of documents belonging to different class say Class A and Class B
- Calculate the prior probability of class A=number of objects of class A/total number of objects
- Calculate the prior probability of class B=number of objects of class B/total number of objects
- Find NI, the total no of frequency of each class
- Na=the total no of frequency of class A
- Nb=the total no of frequency of class B
- Find conditional probability of keyword occurrence given a class:
- P (value 1/Class A) =count/ni (A)
- P (value 1/Class B) =count/ni (B)
- P (value 2/Class A) =count/ni (A)
- P (value 2/Class B) =count/ni (B)
- P (value n/Class B) =count/ni (B)
- Avoid zero frequency problems by applying uniform distribution
- Classify Document C based on the probability p(C/W)
- Find P (A/W) =P (A)*P (value 1/Class A)* P (value 2/Class A)……. P(value n /Class A)
- Find P (B/W) =P (B)*P (value 1/Class B)* P(value 2/Class B)……. P(value n /Class B)
- Assign document to class that has higher probability.

V. MATHEMATICAL MODEL

To decide which terms are minor or major information elements in the document, term weighting schemes using three measures are introduced. The two measures are (1) tfidf Score, and (2) Semantic Score

First, the tfidf score for term I is calculated using the term frequency and the document frequency. Term frequency (tf) and inverse document frequency (idf) are the foundations of the most popular term weighting scheme in IR. The tfidf score of Iij , tfidf(Iij), is computed as:

Tf-Idf(Iij) = Log (tf (Iij , dj) + 1) * Log (—D—/1 + df (Iij , D))

Where tf (Iij , dj) is the frequency of term Iij within document j and df (Iij ,D) is the no. of documents that contain term Iij in the document collection D. Thus, terms with a high tf and low df will get high tf-idf scores.

Second, the semantic score of a term Iij, SSr(Iij) is computed as:

$$SS_R(I_r^j) = \frac{|\{ Ikj \mid SD(Iij, Ikj) \le R \}|}{|dj|, (1 \le k \le |dj|)}$$

Finally classification using Naïve Bayes Algorithm.

VI. EXPERIMENTAL RESULT

In experimental results, we evaluate the proposed system on student conference papers datasets this available on internet. We compare the accuracy of existing system results with proposed system.

The experimental result evaluation, we have notation as follows:

TP: True positive (correctly predicted number of instance)
FP: False positive (incorrectly predicted number of instance),
TN: True negative (correctly predicted the number of instances as not required)
FN false negative (incorrectly predicted the number of instances as not required),
On the basis of this parameter, we can calculate four measurements
Accuracy = TP+TN÷TP+FP+TN+FN
Precision = TP ÷TP+FP
Recall= TP÷TP+FN
F1-Measure = 2×Precision×Recall ÷Precision+ Recall.

A. Comparison Graph:

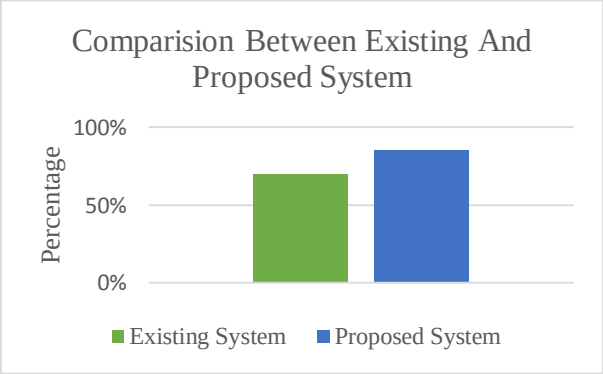

Figure 2: Graph

B. Comparison Table:

Table 1.comparative result

Sr.No	Logistic Regression Result	Naïve Bayes Result
1	75%	87%

VII. CONCLUSION

Proposed Text classification as two coupled predictions activity choose a grouping that is predictive of features. Use predictive performance as a goal criterion, classification parameters the number of function: they are chosen from the model selection. With the result solution, each domain is described by a minimum subset of features necessary to predict if an instance belongs to the data our hypothesis is that even a user will be able to predict domain in the group of documents using the features selected by the clustering algorithm. Given Some relevant requirements, a user can quickly identify that probably contain relevant documents.

VIII. REFERENCES

[1]. J.-T. Chien, "Hierarchical theme and topic modeling," IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 3, pp. 565–578, 2016.
[2]. Bernardini, C. Carpineto, and M. D'Amico, "Full-subtopic retrieval with keyphrase-based search results clustering," in IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intelligent Agent Technol., 2009, pp. 206–213.
[3]. T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela, "Self-organization of a massive document collection,"IEEE Trans. Neural Netw., vol. 11, no. 3, pp. 574–585, 2000.
[4]. K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, and R. Krishnapuram, "A hierarchical monothetic document clustering algorithm for summarization and browsing search results," in Proc. Int. Conf. World Wide Web, 2004, pp. 658–665.
[5]. R. Xu and D. Wunsch, "Survey of clustering algorithms," IEEE Trans. Neural Netw., vol. 16, no. 3, pp. 645–678, 2005.
[6]. S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in Proc. Int. Conf. Inform. Knowl. Manag., 1998, pp. 148–155.
[7]. R. Kohavi and G. H. John, "Wrappers for feature subset selection,"Artif. Intell., vol. 97, no. 1, pp. 273–324, 1997.
[8]. T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela, "Self-organization of a massive document collection,"IEEE Trans. Neural Netw., vol. 11, no. 3, pp. 574–585, 2000.
[9]. Q. Mei, X. Shen, and C. Zhai, "Automatic labeling of multinomial topic models," in Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2007, pp. 490–499.
[10]. K. Lagus and S. Kaski, "Keyword selection method for characterizing text document maps," in Int. Conf. Artificial Neural Networks (ICANN), 1999, pp. 371–376.