International Journal of Research in Advent Technology, Vol.7, No.6, June 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Text Mining: Classification of Text Documents using Granular Hybrid Classification Technique

Shiva Prasad KM, Dr.T Hanumantha Reddy

Abstract- Since past many years, a large amount of raw data is getting converted into digital data within the information era. Maintaining and procuring the data is busy task for all the users who are willing to access the information in line with the requirements, however, the digital knowledge that's unbroken throughout this globe is not relevant in line with the need of the users. To overcome this problem classification process plays a major role to classify the data according to the need of the customer and provide relevant information. The classification algorithm is the process of extracting the information from the large data set and classifying the data which helps the customer to get the relevant information. Multi-class classification is the process of classifying more than two outcomes. Most of the algorithms produce good results when the target classes are few but as the target classes increase the accuracy reduces. There are also cases in classification where instead of classifying a category in the target function, we classify a code. Imagine we want to classify a product code from a large corpus based on the text written by a user. In our paper, we study the repercussions of a corpus which outgrows memory after vectorizing and perform a comparative analysis of various algorithms used during the process with our algorithm. We have represented the Granular Hybrid Model algorithm to classify the ocean ship food catalogue data set based on the user need and product code at a granular level and also by taking care of memory constraints which is a major drawback of normal classification algorithms. Our algorithm has represented a good accuracy of around 75% compared to other algorithms by considering the memory constraints of a huge data set of Ocean ship food catalogue.

Keywords— Machine Learning, Natural Language Processing, Tf-Idf, Sklearn Technique, Granular Hybrid Classification Algorithm

I. INTRODUCTION

From the past many years, the large amount of raw data is getting converted into digital data in this internet. All the data can be transformed into digital data which can be considered as information which have to be managed efficiently. Extracting the huge amount of information from the corpus and reverting it into relevant information is the biggest challenge in the area of the research field. In today's generation managing the data and extracting the relevant information have gained large scope for data analysis and represent the data according to the need of the customers. Support Vector machine is kernel based algorithm which maps the data items into high dimensional data space where the information about their position is used for classification [1]. Multi-class classification is the process of classifying more than two outcomes. Most of the algorithms produce good results when the target classes are few but as the target classes increase the accuracy reduces. There are also cases in classification where instead of classifying a category in the target function, we classify a code. Imagine we want to classify a product code from a large catalogue based on the text written by a user. The analysis required in this scenario is different from the conventional manner, and as we deal with a large corpus, which outgrows the memory available of a system. In this paper, we study the repercussions of a corpus which outgrows memory after vectorizing and perform a comparative analysis of various algorithms used during the process with our algorithm and we will try to provide better result compared to other algorithms.

Classification of data has become a common criterion in the text classification process. There are numerous classification algorithms which are already available for classifying the data (But classification process in these algorithms have taken place with outlier structure of the data sets) and provided the good results. But as soon as the data in the data set increases the result analysis of those algorithms have degraded the accuracy and also it does not classify the data using the sub code generation of the data set. This is one of the major issues in the task of classification problems and no work has been carried out on this aspect until today. So here we will try to work on this aspect which will provide a unique representation of our algorithm and it can be cited by various researchers for their similar works.

In this paper we have proposed Granular Hybrid Model to classify the ocean ship food catalogue data set based on the user need and product code at granular level and also by taking care of memory constraints which is a major drawback of normal classification algorithms Till now most of the classification algorithms even though they have given the best results with minimum data as soon as the data in the data set increases the result accuracy is getting degraded drastically. The accuracy of our algorithm have given the better result by getting trained withX-training data set and tested the result with Y- testing data set compared to other traditional classification algorithms.

Manuscript revised June 5, 2019 and published on July 10, 2019

***Shiva Prasad KM**, Assistant Professor Department of Computer Science and Engineering, Rao Bahadhur Y Mahabaleswarappa Engineering College, Affiliated to VTU Belagavi, Ballari Karnataka, INDIA .*

***Dr.T Hanumantha Reddy** ,Professor and Head Department of Computer Science and Engineering, Rao Bahadhur Y Mahabaleswarappa Engineering College, Affiliated to VTU Belagavi, Ballari Karnataka, INDIA*

doi: 10.32622/ijrat.76201910

II. RELATED WORK

In reference [2] Shwetha Joshi Bhawna Nigam have represented the process of categorizing the document using Multi Class Classification by using Naive Bayes Classification model which helped them to obtain better result by classifying the data. They have used Naive Bayes classifier model by assuming all distinctive attributes are independent of one another, and compute the class of a document based on maximal probability by adopting linear and hierarchical based manner and provided a good accuracy. Further research have to be carried out by building the statistical models which provides meaningful and significant hierarchy. In order to carry out the classification of text effectively it is strongly recommended to have efficient hierarchy information for further investigation. Combination of different classification approaches as an hybrid classifier along with hierarchic structure of classes may provide better result in future.

In reference [3] Semuel Franko and Ismail Burak Parlak have presented multiclass text analysis for the classification problem in Spanish documents. Even if Spanish language is considered as one the most spoken language, classification of text is not carried out due to certain issues in multiclass classification. Naive Bayes Classifier and Maximum Entropy was used to perform the classification of text documents by performing the smoothing of parameters and three different document models. Comparative study on various approaches have been analyzed and found that Maximum Entropy Model is more accurate compared to other models. Comparative assesment would be executed with possibility facts. By the way, alternative multiclass Spanish corpus is not unusual as English language. This hassle is a bottleneck in comparative communities. This hassle is a bottleneck in comparative agencies. Beside, Multi-language multi-elegance hassle may be analyzed with the unique techniques. Furthermore, useful resource Vector Machines or Neural Networks with hierarchical structure could be applied with notable language fashions inside the potential comparative assessment

In reference [4], Mondher Bouazizi And Tomoaki Ohtsuki have proposed a unique approach that in addition to the a fore mentioned responsibilities of binary and ternary classifications, is going deeper inside the classification of texts amassed from Twitter and classifies those texts into a couple of sentiment lessons. In this paper they have introduced SENTA tool which helps the user to select the various features that fits the most of the applications that helps to classify the data through graphical user interface. The accuracy of the approach is found good and accurate. Neither the approach proves it will be more accurate in binary and ternary classification.

In reference [5], Sumitra Pundlik, Prachi Kasbekar and Gajanan Gaikwad have represented Multi-Class and Class based analysis of Hindi Language using Ontology process. Here they have used an HINDISENTIWORDNET (HSWN) n order to find the polarity of the various classes and in order to improve the accuracy of the system they have combined the LM Classifier along with Ontology. In future the problem can be represented and implemented using parallel processing technique along with ontology which may help to improve the accuracy of classification.

In reference [6], Dewan Md. Farid, Mohammad Masudur Rahman, M A Al-Mamun have proposed Navie Bayes Tree approach for efficient Multi Class Classification of Documents. The NB Tree is a hybrid classifier of decision tree and navie bayes classifier. In NBTree nodes comprise and cut up as normal decision tree, however the leaves are replaced with the aid of naive Bayes classifier. We tested the overall performance of proposed algorithm towards the ones of the present decision tree and naive Bayes classifiers respectively using the classification accuracy, precision, sensitivity-specifitown evaluation, and 10-fold move validation on 10 actual benchmark datasets from UCI (university of California, Irvine) machine learning repository. The experimental results display that the proposed method progressed the classification accuracy of actual existence tough multi-magnificence troubles.

In reference [7], Guobin Ou1, Yi Lu Murphey, and Lee Feldkamp have discussed major approaches of neural networks for classification of documents in the form of multi classes. They have discussed various algorithms such as One-Again-All, One-Against-One, and P-against-Q. The performance is measured based on NSIT Handwritten digit text.

In reference [8], Rajni Jindal and Swetha Taneja have proposed Multi Label classification of text documents mistreatment quantifiers. They have created eight new quantifiers that calculate the degree of membership of sophistication labels of a specific text document. As a result, we tend to square measure able to perform ranking of sophistication labels in multi label learning.

In reference [9], Quips, Alexander Ocsa1 and Ricardo Coronado have Latent Semantic Indexing for Text classification since the text belongs to multiple number of classes or index they have used Latent Semantic indexing along with neural networks which helped them to extract the required information and classify the document by providing the indexing for the document. The result shows that the accuracy of the approach is good but the precision rate of the model will be poor when the size of the document is small.

In reference [10], Yan Xu has proposed feature selection for unbalanced text classification problem. They have represented dimensionality reduction process for classifying the data based on the feature selection process and performed the comparative process on various feature selection algorithms on both Chinese and English corpus by finding out the document frequency values and information gain values. The comparative results shows that IG and CHI have provided more efficient result compared with other algorithms They have mentioned that in future combination of complex algorithms like bagging and Boosting algorithms can be used for classification of these type of text documents.

In reference [11], Sowmya B J and Chetan K.G.Srinivasa have proposed two algorithms such as Rocchio and KNN algorithm which is used for classification of text in the form

doi: 10.32622/ijrat.76201910

of hierarchal structure. Hierarchical information is changing into more and more distinguished, particularly on the online. Wikipedia is one such example wherever there are numerous documents that are classified into multiple categories in an exceedingly class-conscious fashion. This provides rise to a remarkable downside of automating the classification of latest documents. Because the size of the data set grows, thus will the amount of categories. The sparcity of the document is also major issue which can be overcome with the help of these two algorithms. When used in text categorization, the nearest centroid classifier is also known as the Rocchio classifier because of how similar it is to the Rocchio algorithm for relevance feedback. An elongated version of the nearest centroid classifier has found many uses in the medical domain, specifically classification of tumors.

In reference 12, Nafiseh Forouzideh, Maryam Tayefeh Mahmoudi and Kambiz Badie have proposed an artificial Immune Recognition system foe text classification. The AIRS method performs well on either large dimensioned feature space problems or problems with many classes or both. In this paper various versions of AIRS along with KNN algorithm is used to classify the text document which helps for user and organization tasks. Here they have chosen Low, Medium and High nominal values are chosen to organize and classify the text in the form of functional classes.
As final discussion, the conferred classification algorithms supported AIRS will be notably helpful for organizing texts in call support environment, wherever enriching the prevailing texts for supporting the human components with their selections is of specific significance.
In reference 13, Kamran Kowsari, Donald E. Brown and Mojtaba Heidarysafa represented that the traditional supervised classifiers performance gets degraded as the number of documents increases which gets failed to classify the text. So in order to overcome the problem they have proposed new method called Deep Learning (HDLTEXT) method. This paper approaches this drawback otherwise from current document classification ways that read the matter as multi-class classification. Instead we tend to perform ranked classification using an approach we call Hierarchical Deep Learning for Text classification (HDLTex). HDLTex may be applied to untagged documents, like those found in news or alternative media retailers. Marking here can be performed on tiny sets mistreatment human judges.

III. MOTIVATION

Classifying the huge amount of data according to the need of the customers plays a vital role in today's generation and gained lot of attention towards this area of research. Classifying the data with the minimum amount of data can be done with the traditional algorithm but classifying the huge amount of data by considering the memory constraints and sub-topics of data set brought us lot of attention towards the process of granular classification of data by considering the memory constraints.

IV. PROBLEM STATEMENT

We have proposed the Granular Hybrid Model to classify the ocean ship food catalogue data set based on the user need and product code at granular level and also by taking care of memory constraints which is a major drawback of normal classification algorithms Till now most of the classification algorithms even though they have given the best results with minimum data as soon as the data in the data set increases the result accuracy is getting degraded drastically. The accuracy of our algorithm have given the better result by getting trained with X-training data set and tested the result with Y-testing data set compared to other traditional classification algorithms.

V. BACKGROUND KNOWLEDGE

5.1. Definition of Text Analytics:

Text analytics is the process of exploring and analyzing the unstructured text data provided by the user that can identify the patterns, topic models, keywords and other attributes. Now a days text mining is playing the major role in the area of research for analyzing the huge amount of data with various machine learning and deep learning algorithms. Text mining helps in identifying the relevant information which is required for the user by converting the unstructured data into structured data by exploring the various sort of inputs needed for the user [2]. It is similar to data mining but it concentrates more on unstructured text compared to structured forms of data mining. Text mining starts the process by considering the unstructured dataset as input and converts it in the form of structured data representation which is most essential for qualitative and quantitative analysis which is involved with Natural language processing techniques, Machine learning algorithms and other linguistic algorithms which helps to perform different operations like retrieval, classification and discovering the various patterns based on the requirement of the user.

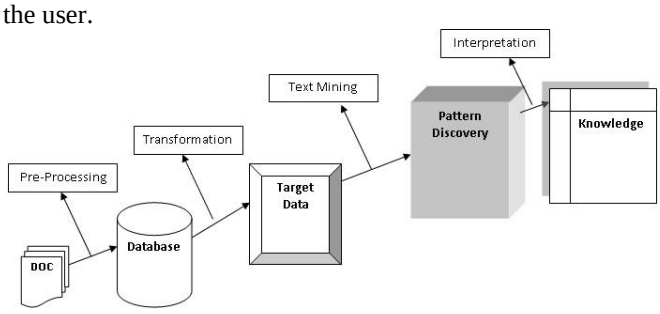

Figure 1: Text Analytics Architecture

5.2. Definition of Text Classification

Text classification is the process of assigning one or more classes based on the content present in the document. Classification process takes care of all preprocessing tasks such as stemming, extracting, removal of stop words and organizing into different classes which are required for automatic classification of a huge amount of data. It is capable of filtering the spam data, identifying the positive and negative remarks of the text documents and multi-class classification like selecting one category by considering several alternatives of information from the corpus [2]. The different statistical algorithms like Naive biased, Decision Tree, Neural network Classifier Centroid-based classifier,

doi: 10.32622/ijrat.76201910

Linear classifier and so on are used for performing the process of classification based on the data set and requirement of the user.

The text classification process takes place based on the set of categories and collection of text documents in order to find the relevant topics based on the instances given by the user for each document. Let us consider the set of text documents d and its constraints as c where P(d,c)=>(0,1). The value of P(d,c) will be considered as 1 if the document is relevant otherwise it will be set as 0.

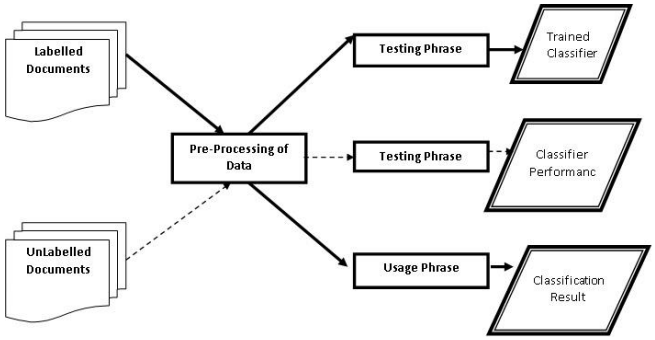

Figure 2: Text Classification Architecture

5.3. Multi-Class Classification:

Multi-class classification is the process of predicting more than two classes. It forms a part of the supervised learning section of machine learning. There are dependent attributes say X and independent attributes say Y. We run a series of operations on X to enable the algorithm to run and produce a set of results. Let this be Y'. So basically f(X) = Y which is the actual values and f(X) = Y' which becomes the predicted values. The difference between our predicted value and the actual is the error. So f(Y') = Y becomes our error functions. Normally multi-class algorithms have been seen providing decent results when it comes to a lower number of target functions and the algorithm gets enough records for each category to run on but as the number of the target increase as in the case of NLP(Natural Language Processing) the error increases. So basically

f(X) = Y (Actual)
f(X) = Y (Predicted)
E(Y) = (Error function)

Now , E is in a decent range when we have a large number of records R and lower number of target classes say T but when the records R are less and the target classes are more in that cases the E increases exponentially.

5.4. Random Decision forest Classifier:

Random Forest is a supervised classification algorithm which was first introduced properly in [3] by Leo Brieman. It is the method that operates by constructing multiple decision trees during training phrase. The decision of majority of the trees is choosen from the random forest as final tree. Random forest Classifiers uses bootstrap aggregating techniques that repeatedly selects random samples of the coaching set with replacement and uses this sample to create trees learn since bootstrap technique decreases the variance and so ends up with higher performance. It is an ensemble learning method process which performs various operations such as classification regression and other tasks by constructing multiple decision trees. Random Forest classifier is used in outlier detection and exchange missing information. It's climbable because it will run on massive information sets.
The Relationship between the nearest neighbors is calculated using the formula

$$y = \sum_{i=1}^{n} W\ (x\ i, x) yi$$

5.5. Naive Baisen Classifier:

Naive Baisen classifier is a popular method for text classification problems. It is also considered as probabilistic classifier which makes use of Bayes theorem for classification of various features of the text documents. For a particular document or article the classifier will decide the category of the article. It was first proposed by D.Lewis [4].This classifier works on the assumption of independence between the various features. It is a scalable classifier and can run efficiently with large data sets.

Naive Bayes classifier is fast as compared to other classifiers and thus is used as a baseline for text classification problems. The probability of given dataset D which contains all the words WI for given class C is derived as

$$P\ (D/C) = \pi\ P\ (W_i/C)$$

The probability of given dataset D with respect to class is derived as

$$P\ (D/C) = P\ (D \cap C)/P(C)$$

5.7. Support Vector Machine Classifier

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges.SVM (Support Vector Machine) [5] works on the principle of Supervised Learning. SVM requires a training set and labels associated with it. After training, if a test data is fed in, the model assigns it to one category or the other. It performs well with linear classification. It can even work efficiently on a nonlinear classification using a kernel trick by mapping the inputs into high dimensional feature space. It constructs a hyper-plane for classification [6]. The hyper-plane is chosen such that the distance between the nearest data point on either side is maximized [7]. One possible form of the prediction equations for SVM classification is:

$$h(x) = sign(\sum_{j=1}^{s} aj, yiK(xj, xi) + b)$$

5.7. Linear Support Vector Classifier:

doi: 10.32622/ijrat.76201910

In the linear classifier model, we assumed that training examples plotted in space. These data points are expected to be separated by an apparent gap. It predicts a straight hyper plane dividing 2 classes. The primary focus while drawing the hyper plane is on maximizing the distance from hyper plane to the nearest data point of either class. The drawn hyperplane called as a maximum-margin hyper plane [8].

5.8. Tf-Idf Vectorizer:

Term Frequency and Inverse Document Frequency is a numerical statistic method which is used to reflect how much important a particular word is to the document in the collection of data set. The Tf-Idf is a well know method to evaluate how important is a word in a document. Tf-idf vectorizer is a very interesting way to convert the textual representation of information into a Vector Space Model (VSM) or into sparse features.

The Term Frequency (Tf) of the particular data set or corpus is used to calculate the number of times a particular word has occurred in the document. The Inverse document frequency (Idf) is used to identify the no of documents consisting a particular word by considering and dividing the total no of documents. After completing the process of calculating the values of term frequency and Inverse Document frequency separately we can calculate the Tf-Idf by multiplying the values if Tf and Idf together.

VI. METHODOLOGY

The proposed architecture shown in the figure below represents the process of classification of text at granular level using a hybrid technique. This model is uploaded with the natural language data set which contains the information of Ocean Ship food Catalogue system and tried with classifying the data based on the product code and customer description which is considered as Natural language. This document is then analyzed based on the requirements for classification of text from the data set and performed with general preprocessing of data in order to convert the unstructured data into structured representation which will be more helpful for further processing of data. Based on the customer description and product code of the catalogue we have performed a granular classification of text which provided an effective result compared to other classification algorithms.

The algorithm takes data from a file or a database. The first step is to do analysis of the data received. There can be various different kinds of processing that needs to be done before the actual machine learning algorithm performs the task. There are basically four different kinds of data that can be received. Numerical or floating point data, Categorical data, Date based and Text based Data.

Most of the machine learning algorithms can work directly on the Numerical data so normally not much processing has to be done on it. But empty fields can be updated by following the basic preprocessing approaches. Replacing the empty fields can with zeros is one option, getting the mean or median and replacing the empty fields can be another option. A third way is to predict the value using regression machine learning techniques.

Categorical data are a small set of different classes that can be converted to columns which in turn gives us a sparse matrix. This columns contains zeros for the values which are not existing and contains ones for if the filed exists in the given record. Sometimes when there are a fair amount of classes it gives rise to "curse of dimensionality" which causes many dimensions by which there are chances of going out of memory. There are some optimization techniques which can be used to reduce the dimensions if the weight of the dimensions of the overall result is not lesser.

Date based data offers a fair amount of information and can be used to convert to month of year, quarter of year, day of week etc. to find various patterns that may exist in data. This step gains importance if the weight of the date is more for the classification and impacts the final outcome of the record.

Text based data requires some special importance as they do not come in either of the above fields. Since algorithms works on numerical data and this text based data are not categorical there are special ways to handle them like creating vectors of the text which are dependent on the overall text in a given dataset.

Algorithm of Granular Hybrid Model

1. Initialize input values and output data to zero
2. Analysis of Data set D
3. Preprocessing of data using Tokenization and Lemmatization process
4. For each intermediate node, classify nodes as static nodes and dynamic nodes.
5. For each intermediate node random assign weights
6. For each sample D in d where d in h
   a. Calculate output of each samples
   b. Evaluate the customer description of dynamic intermediate data
   c. Adjust the dynamic data based on the customer requirement percentage
7. For each d in D perform
   a. Calculate in sampling

For each intermediate node in network back propagate and adjust the weights of nodes

$$\textbf{Recall}(t_i, r_i) = \frac{|\{d \in P_i | f_i(d)=t_j, t(d)=t_j\}|}{|\{d \in P_i | t(d)=t_j\}|}$$

$$\textbf{Precision}(t_j, r_i) = \frac{|\{d \in r_i | f_i(d)=t_j, t(d)=t_j\}|}{|\{d \in P_i | f_i(d)=t_j|}$$

doi: 10.32622/ijrat.76201910

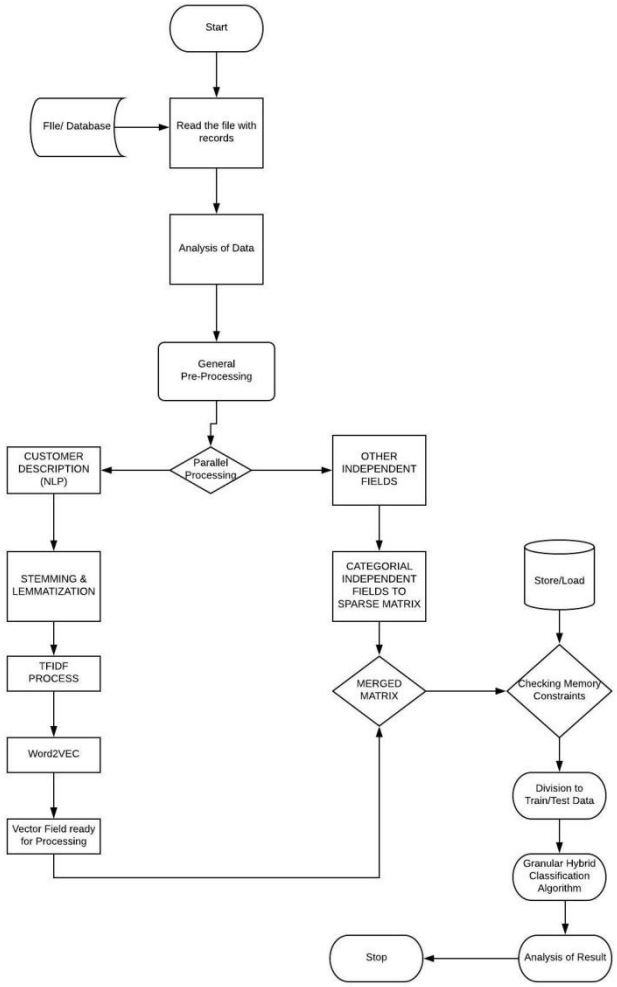

Figure 3: Architecture of Granular Hybrid Model

7. Experimental Results

a. Description of data set:

A ship in the ocean needs various resources and has to dock to a port to get the supplies. A shared catalogue is not available and the caption of the ship describes his requirements by writing down in English, which is sent over to an agency to provide the required items. So based on the Category, Party, Branch and customer description the product code the classification process have to be carried out at granular level.

Category	Party	Branch	Customer Description	Product Code
Provision	Ocean Catering LTD	A	Assorted Chocolates	PA2002 0060
Provision	Ocean Catering LTD	B	Pork Belly	PF10120 100
Provision	Ocean Catering LTD	C	SHRIMP RAW	PA2102 0012
Provision	Ocean Catering LTD	D	Clustered Fish	PA6534 1729

Table 1: Representation of Ocean Food Catering Ltd Data Set

b. Mathematical Representation:

The mathematical representation of our model which is used to calculate the precision and recall is represented below.
Here

S is normal set of data,
r is the partition

$$\mathbf{Ti= \{d \in T/d \notin r :\}}$$

For each pair of task $t_j w/w_j$={ 1,2,….32} & pold ri, i={1,2,3,….,10} Precision and recall values are computed as follows

Where t (d) and f_i (d) are original and the predicted task labels of data sentence d respectively.

c. Result :

The process of classifying the text documents at granular level depends upon the trained data set which is more essential for training our algorithm to work in an effective manner. In recent of the research works carried out by different researchers we have identified the process of classification taken place at modular level with limited number of classes. But the granular level of classification was not yet done so we tried to classify the document at granular level and found effective compared to other classical algorithm. The various parameters were considered for our setup before the actual evaluation result acquired using our algorithm. Our projected model Granular Hybrid Classification algorithm is a classifier which is used to classify the document at minute level which provide n number of categories based on the product code and client description that is described under natural language Processing. This model has provided an efficient result with good accuracy. Granular Hybrid classification formula result compared with alternative classical classification formulas and located that our algorithm is way effective and correct. The comparative result of various classical algorithms along with our proposed algorithm and graphical representation of results is represented below

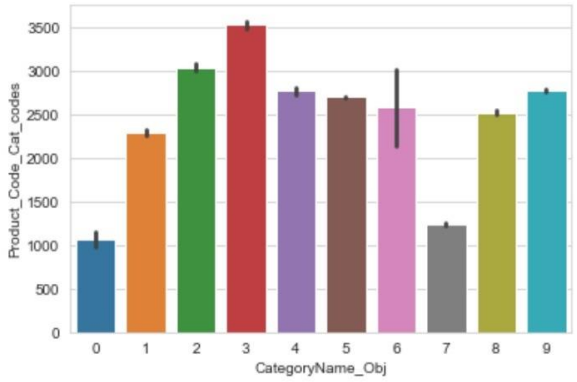

Figure 4: Graphical Representation of Product Code with Category Name

doi: 10.32622/ijrat.76201910

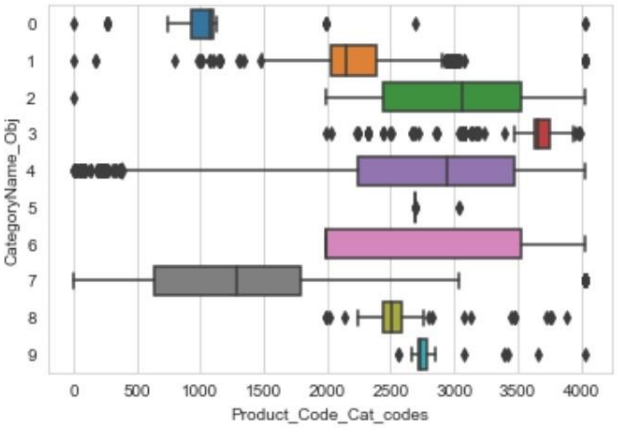

Figure 5: Graphical representation of Accuracy of Product code with different samples

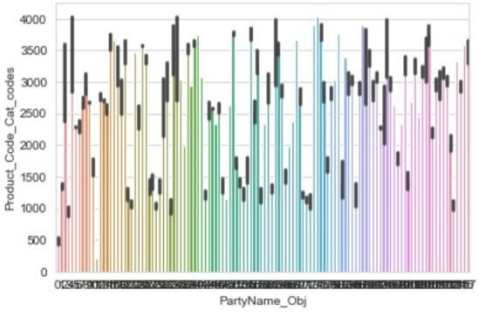

Figure 6: Graphical representation of Party Name with its Product code

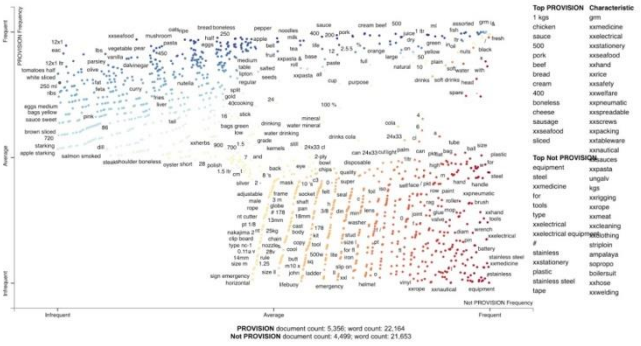

Figure 7: Graphical representation of Major Provision and Minor Provision Products

	TOTAL NO OF SAMPLES TAKEN		
Algorithms	**10000**	**50000**	**100000**
Navie Basie Classifier	24%	38%	50%
Random Forest Classifier	22%	36%	45%
Decision Tress Classifier	35%	42%	55%
Granular Hybrid Classifier	46%	60%	79%

Table 2: Comparative Table of Different Algorithm

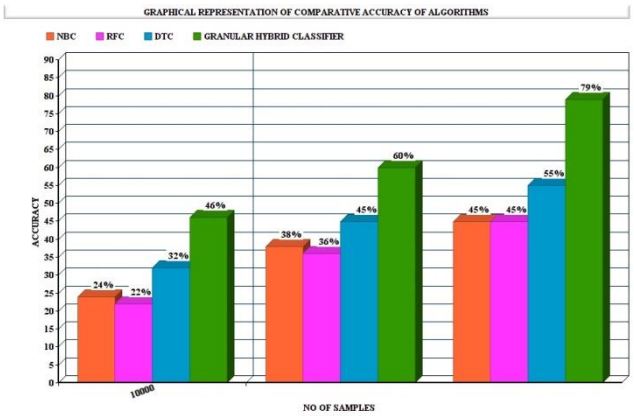

Figure 8: Graphical representation of Comparative result of Classification Algorithm

VII. CONCLUSION AND FUTURE WORK

Text classification is one among the fore most necessary tasks of text mining. During this paper, we've got planned the Granular hybrid classification model for classification of text at a granular level that depicts the assorted phrases through the building of automatic text classification and illustration of relationship among them. The analysis of the planned model was trained and tested with one large integer knowledge samples of Ocean ship food catalogue system that contain the info of product name, client description, Product code and then on. The model has been constructed with a hybrid technique that provided an improved end in terms of granular classification method compared to alternative classical algorithms of Machine learning. The accuracy of our model is 68% for one lakh data samples and found more efficient compared to other alternative classification algorithms.

REFERENCES

[1] **"**Lee, Chung-Hong, Hsin-Chang Yang, and Sheng-Min Ma. "A novel multilingual text categorization system using latent semantic indexing." In *Innovative Computing, Information and Control, 2006. ICICIC'06. First International Conference on*, vol. 2, pp. 503-506. IEEE, 2006.

[2] Joshi, Shweta, and Bhawna Nigam. "Categorizing the document using multi class classification in data mining." In *Computational Intelligence and Communication Networks (CICN), 2011 International Conference on*, pp. 251-255. IEEE, 2011.

[3] Franko, Semuel, and Ismail Burak Parlak. "A comparative approach for multiclass text analysis." In *Digital Forensic and Security (ISDFS), 2018 6th International Symposium on*, pp. 1-6. IEEE, 2018.

doi: 10.32622/ijrat.76201910

[4] Bouazizi, Mondher, and Tomoaki Ohtsuki. "A pattern-based approach for multi-class sentiment analysis in twitter." *IEEE Access* 5 (2017): 20617-20639.

[5] "Pundlik, Sumitra, Prasad Dasare, Prachi Kasbekar, Akshay Gawade, Gajanan Gaikwad, and Purushottam Pundlik. "Multiclass classification and class based sentiment analysis for Hindi language." In *Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on*, pp. 512-518. IEEE, 2016.

[6] Farid, Dewan Md, Mohammad Masudur Rahman, and M. A. Al-Mamuny. "Efficient and scalable multi-class classification using naïve Bayes tree." In *2014 International Conference on Informatics, Electronics & Vision (ICIEV)*, pp. 1-4. IEEE, 2014.

[7] Ou, Guobin, and Yi Lu Murphey. "Multi-class pattern classification using neural networks." *Pattern Recognition* 40, no. 1 (2007): 4-18.

[8] Jindal, Rajni, and Shweta Taneja. "Ranking in multi label classification of text documents using quantifiers." In *Control System, Computing and Engineering (ICCSCE), 2015 IEEE International Conference on*, pp. 162-166. IEEE, 2015..

[9] Quispe, Oscar, Alexander Ocsa, and Ricardo Coronado. "Latent semantic indexing and convolutional neural network for multi-label and multi-class text classification." In *Computational Intelligence (LA-CCI), 2017 IEEE Latin American Conference on*, pp. 1-6. IEEE, 2017.

[10] Xu, Yan. "A comparative study on feature selection in unbalance text classification." In *Information Science and Engineering (ISISE), 2012 International Symposium on*, pp. 44-47. IEEE, 2012.

[11] Sowmya, B. J., and K. G. Srinivasa. "Large scale multi-label text classification of a hierarchical dataset using Rocchio algorithm." In *Computation System and Information Technology for Sustainable Solutions (CSITSS), International Conference on*, pp. 291-296. IEEE, 2016.

[12] Forouzideh, Nafiseh, Maryam Tayefeh Mahmoudi, and Kambiz Badie. "Organizational texts classification using artificial immune recognition systems." In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2011 IEEE Symposium on*, pp. 1-8. IEEE, 2011.

[13] Kowsari, Kamran, Donald E. Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S. Gerber, and Laura E. Barnes. "Hdltex: Hierarchical deep learning for text classification." In *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on*, pp. 364-371. IEEE, 2017.

[14] Korde, Vandana, and C. Namrata Mahender. "Text classification and classifiers: A survey." *International Journal of Artificial Intelligence & Applications* 3, no. 2 (2012): 85

[15] Breiman, Leo. "Random forests." *Machine learning* 45, no. 1 (2001): 5-32.

[16] Lewis, David D. "Naive (Bayes) at forty: The independence assumption in information retrieval." In *European conference on machine learning*, pp. 4-15. Springer, Berlin, Heidelberg, 1998.

[17] Mining, What Is Data. "Data Mining: Concepts and Techniques." J K M Han (2006).

[18] Yu, Hwanjo, and Sungchul Kim. "SVM tutorial—classification, regression and ranking." In *Handbook of Natural computing*, pp. 479-506. Springer, Berlin, Heidelberg, 2012.

[19] Ee, Chee-Hong Chan Aixin Sun, and Peng Lim. "Automated online news classification with personalization." In *4th international conference on asian digital libraries*. 2001.

[20] http://dataaspirant.com/2017/01/13/support-vector-machine-algorithm.

AUTHORS PROFILE

Mr.Shiva Prasad KM is presently working as Assistant Professor Department of Computer Science and Engineering, Rao Bahadhur Y Mahabaleswarappa Engineering College Ballari Karnataka. His area of interest is Data Miningand Machine learming. He have published numerous papers in Conferences and Journals. He his having professional membership in IEAE,ISC,IAENG

Dr. T Hanumantha Reddy is presently working as Professor and Head of Computer Science and Engineering, Rao Bahadhur Y Mahabaleswarappa Engineering College Ballari Karnataka. His area of interest is Data Mining Artificial Intelligence and Digital Image Processing. He have published numerous papers in Conferences and Journals. He his having professional membership in IEAE,ISC, IAENG.

doi: 10.32622/ijrat.76201910