Analysis of Text Classification With ARM on Various Data Mining Metrics

Meghna Utmal[1], Dr. R. K. Pandey[2]

[1]*Research Scholar , Deptt. of Computer Science , RDVV Jabalpur , India*

[2]*Dean and Professor, UICSA, RDVV Jabalpur, India*

Abstract: In today's era of data mining where we get enormous data from heterogeneous sources in structured as well as unstructured format, so our present paper is based on the classification techniques involved for finding different parameters for processing unstructured data . Till now work on association and classification have been carried out independently but our work proposes technique involving association and classification altogether. We have taken FIRE Dataset (Forum of Information Retrieval) for the proposed work and our experiment has carried out on Rapid Miner which is a data mining tool. FP-growth an algorithm of association rule mining is used by the proposed work for obtaining frequent set and Naive bayes classifier is a model of classification for the purpose of constructing a class. The results have been evaluated with the standard measures. The experimental result shows increase in efficiency while comparing with other traditional text classification methods.

Keywords: Naive Bayes Classifier, FP-growth , Data Mining, FIRE, Rapid Miner, Association Rule Mining, Classification.

I. INTRODUCTION

The data from various fields are being generated by humans everyday . The data generated are in both structured as well as unstructured form. It may be in the form of documents, may be graphical formats ,may be the video ,may be records (varying array) . The magnitude of data generated and shared by businesses, public administrations numerous industrial and not-to-profit sectors, and scientific research, has increased immeasurably. These data include textual content (i.e. structured, semi-structured as well as unstructured), to multimedia content (e.g. videos, images, audio) on a multiplicity of platforms (e.g. machine-to-machine communications, social media sites, sensors networks, cyber-physical systems, and Internet of Things [IoT]).We have to take proper action with unstructured data so as to get some important results. Proper analysis helps in good decision making with these data. It is actually the process of finding the hidden information/pattern of the repositories[1]. The important reason that attracted a great deal of attention in information technology the discovery of useful information from large collections of data industry towards field of "Data mining" is due to the perception of "we are data rich but information poor". There is huge volume of data but we hardly able to turn them in to useful information and knowledge for managerial decision making in business.

The term data mining is used for methods that analyze data with the objective of finding rules and patterns describing the characteristic properties of the data.

Text mining: It is a multidisciplinary field, involving information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, machine learning, and data mining [2]. Here we generally deal with the unstructured text documents using natural language processing techniques to extract keywords labeling the items in that text documents. Then, classical data mining techniques are applied on the extracted data (keywords) to discover interesting patterns. Starting with a collection of documents, a text mining process would retrieve a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted [3].

For mining large document collections, it is necessary to pre-process the text documents and store the information in a data structure, which is more appropriate for further processing than a plain text file [4]. Text pre-processing classically means tokenization and then Part of Speech Tagging or in a bag of words approach word stemming and the application of a stop word list. Tokenization is the process of splitting the text into words or terms.

The concept of text mining generally deals with unstructured or textual information for the extraction of meaningful information and knowledge from huge amount of text. Most of the times, data that we gather from different sources is so large that we cannot read it and analyze it manually so we need text mining techniques to deal with such data. Identifying and separating out any specific type of information from the given text requires text mining techniques or methods. These methods also help in clustering the data into different groups on the basis of specific requirements. In the field of education, text mining techniques helps to explore and analyze data coming from new discoveries and researches that are made on daily basis in large amount.

Many approaches have been proposed to discover useful information from structured data. Among these approaches, association rule mining **(ARM)** plays an important role. The ARM algorithms aim to discover hidden rules among enormous pattern combinations based

on their individual and conditional frequencies. The traditional ARM algorithms first generate all of the possible patterns from the data while pruning out non-frequent ones and then produce rules from these frequent patterns. Once the rules are generated, some interesting measures (IMs) are applied to obtain interesting rules that can be used in decision making.

In a few decades, management of the electronic document based on the content, has received a distinguished reputation in information system domain. This is due to the extended accessibility of text data in the digital form [5]. An important component of text classification system is selection of good quality of document collection, representation model and adaption of suitable classification techniques [4]. At its best, a text file series can be any grouping of text content based files. Algorithmically, solutions of the text mining applications depend on the selection of patterns in the large datasets [6]. The total number of (text files) text documents in such datasets may range from a few thousands to millions [7-8]. Text documents, is a collection of terms (words), which is difficult to understand and challenging to be interpreted by a classifier. Due to this, the unstructured text data is transformed into machine understandable form. This article presents the text classification by considering score level fusion techniques [9-10]. Two different classifiers are designed based on the two different representation models of the text documents.

II. RELATED WORK

Several work has been carried out in the field of text mining of which work done by Hussein Hashimi [11] includes categorization of text, summarization, topic detection , extraction search and retrieval. Each of these techniques can be used in finding some non-trivial information from a collection of documents. Text mining can also be employed to detect a document's main topic/theme which is useful in creating taxonomy from the document collection. Areas of applications for text mining include publishing, media, telecommunications, marketing, research, healthcare, medicine, etc.

Further it as suggested by S.N. Bharat Bhushan [12] that text document classification is a well known theme in the field of Information Retrieval. Selection of most desired features in the text document plays a vital role in classification problem. Two different classifiers are designed based on the two different representation models of the text documents. Score level fusion is applied on two proposed models to find out the overall accuracy of the proposed model. A word level representation model for semantic information preserving of the text document and score level fusion approach.

Next the problem of mining structured data to find useful patterns were dealt by Ömer M. Soysal [13] where he found potentially useful patterns by association rule mining. The proposed approach requires less computational resources in terms of time and memory requirements while generating a long sequence of patterns that have the highest co-occurrence.

The paper by Rahman [14] discusses the technique for classifying mining text using classification with association rule. Pre-classified text document are formed by Association rule mining technique which is then used by Naïve Bayes classifier for final classification.

Kamruzzaman [15] in his work uses word relation instead of using words to derive feature set from pre-classified document. Naive Bayes Classifier is then used on derived features along with Genetic Algorithm.

III. METHODOLOGY

In this paper we used FIRE dataset for experimentation. The experiment has been performed with data sets taken from different domains such as sports, news, politics, education etc. FIRE dataset adapts TREC document style format.

There are three different fields in this document such as header file represented by DOC which is the starting part of the document. The next field comprise of DOCNO that has unique identifier. Finally we have a TEXT field which contains the data in unstructured format. One of the snapshots of the FIRE doc is shown in the following figure,

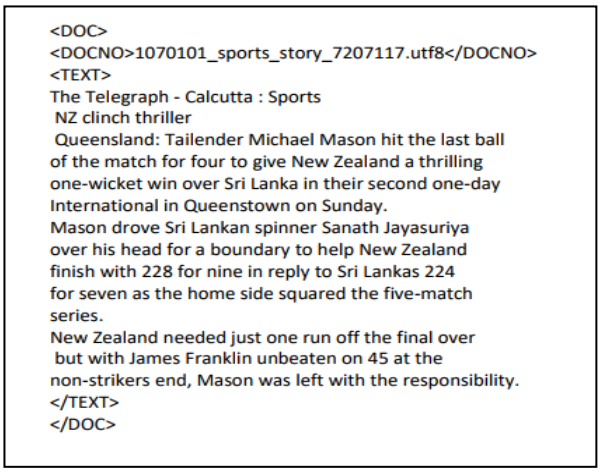

Figure 1: Sample Text File FIRE

Out of the various tools for data mining like R, WEKA etc. We have used RapidMiner for our experimental results.

As our data comprise of unstructured data so it is necessary to perform preprocessing of the given data. The various steps involved in preprocessing of the given document comprise of :

1. Tokenization – It divides the entire string into tokens.

2. Case Transformation – it transform all the words in lower case as RapidMiner is case sensitive.
3. Stop word removal – it removes the stop words like is, as, an, they etc.
4. Stemming – it reduces the words to its root form
5. N-gram – there are 2 n-grams character and term , here we are using term wise n-gram.

Word	Attribute Name	Total Occurences	Document Occurences	a
actor	actor	1	1	1
actor_try	actor_try	1	1	1
actor_try_dodg	actor_try_dodg	1	1	1
affili	affili	1	1	1
affili_member	affili_member	1	1	1
affili_member_nation	affili_member_nati...	1	1	1
allow	allow	3	2	3
allow_icc	allow_icc	1	1	1
allow_icc_full	allow_icc_full	1	1	1
allow_yard	allow_yard	2	1	2
allow_yard_circl	allow_yard_circl	2	1	2

Figure 2: Word list result of Preprocessing

The various steps involved after the preprocessing of the dataset are as follows:

1. Firstly we will perform the association rule mining.
2. Next we perform classification based on the result of association rule mining.
3. Further we have studied our results on the parameters such as accuracy, recall and precision on the different classification methods such as KNN, Random Forest, Naive Bayes.
4. In our results we have combined two functions of data mining association mining which comprised of FP-growth and Naive bayes algorithm in classification function.

IV. ANALYSIS OF RESULTS

The methodology and experiment is based on the dataset taken from FIRE (Forum of Information Retrieval) [22]. The result of the proposed approach have been evaluated with of various parameters of data mining metrics such as accuracy, precision and recall. The details of the measures are summarized as following,

(i) Accuracy is one of the measures which evaluate the performance of the models.Correctly classified examples are defined by accuracy. It is calculated by taking % of correct predictions over total number of examples. It can be express through the given formula

$$\textbf{Accuracy} = (TP + TN) / (P + N)$$

Where TP= True positive, TN= True Negative,P= Positive and N= Negative

(ii) Precision is the number of the predicted positive values that were correct.

$$\textbf{Precision} = TP / (TP + FP)$$

Where TP= True Positive, FP= False Positive

$$\textbf{Recall} = a/(a+b)$$

Where a is number of relevant records retrieved, b is number of relevant records not retrieved.

As mentioned above in the methodology section the comparison results of the two data mining functions

with their different techniques and parameters are shown in the below table followed by the graph.

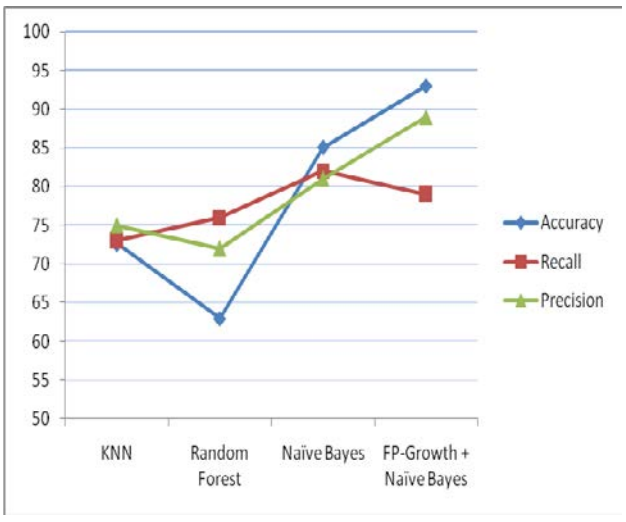

Figure 3 : Comparison Result of different Classification Techniques

The results predicted by the graph mentioned above concludes that accuracy of the proposed method is the highest as compared to the traditional methods of text classification. The association between different frequent words is an efficient reason to improve the accuracy of the system. The proposed method gives higher result in both

the parameters as it correlates the term according to their association based on the FP-growth algorithm. Based on the various evaluation parameters we find that the proposed method outperforms in most of the metrics.

Table 1: Comparison of all the Measures

Parameters	Classification methods			Proposed method
	KNN	Random Forest	Naïve Bayes	FP-Growth + Naïve Bayes
Accuracy	72.6	63	85	**93**
Recall	73	76	82	**79**
Precision	75	72	81	**89**

V. CONCLUSION

To discover hidden pattern from unstructured large textual collection we use the technique of text mining. The present paper is based on the two functionalities of data mining that is association and classification. Our paper has well proposed the technique of text processing moreover we have worked with the data from FIRE set and our work was conducted on Rapid Miner a well known data mining tool. The results have been evaluated with the standard measures. The present method is useful for unstructured large amount of database for analysis in less time.

VI. REFERENCES

[1] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.

[2] Dunham, M. H., Sridhar S., “Data Mining: Introductory and Advanced Topics”, Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006 .

[3] Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., “From Data Mining to Knowledge Discovery in Databases,” AI Magazine, American Association for Artificial Intelligence, 1996.

[4] Vishal Gupta and Gurpreet S.Lehal, "A Survey of Text Mining Techniques and Applications", Journal Of Emerging Technologies in Web Int.

[5] Nasukawa and Nagano, "Text Analysis and Knowledge Mining System", IBM Systems Journal, Vol.40, No.4, pp.967-984, October 2001.

[6] Raymond J. Mooney and Razvan Bunescu, "Mining knowledge from text using information extraction", ACM SIGKDD Explorations Newsletter, Vol.7,No.1, pp.3-10, 2005.

[7] Hotho, Nurnberger and Paass, "A Brief Survey of Text Mining Export", LDV Forum, Vol.20, No.2, pp.19-62, 2005.

[8] Manning and Schütze, “Foundations of statistical IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 3, May 2012 ISSN (Online): 1694-0814 www.IJCSI.org 550 Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved. Natural language processing”, MIT Press 1999.

[9] C.C. Aggarwal, C. Zhai, in: A Survey of Text Classification algorithms. In mining Text Data, Springer, 2012, pp. 163–222.

[10] A.K Pujari, Data Mining Techniques, University press, 2004.

[11] Hussein H., " Selection criteria for text mining approaches ", College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia.

[12] Bhushan B. “Classification of text documents based on score level fusion approach”, Jawaharlal Nehru National College of Engineering, Shivamogga – 577204.

[13] Soysal O.M. “Association rule mining with mostly associated sequential patterns”, 7 Highway Safety Research Group, Louisiana State University, 3535 Nicholson Ext., Baton Rouge, LA, USA.

[14] Rahman C.M. et al., “Text classification using the concept of association rule of data mining” arXiv preprint arXiv:1009.4582.

[15] Kamruzzaman S. M. et al., “Text classification using association rule with a hybrid concept of naive Bayes classifier and genetic algorithm”, arXiv preprint arXiv:1009.4976.