International Journal of Scientific Research & Engineering Trends
Volume 6, Issue 5, Sept-Oct-2020, ISSN (Online): 2395-566X

A Survey on XML Data Processing Using Data Mining Techniques

Twinkle Gurjar, Arjun Parihar
Dept. Of CS.,Rajiv Gandhi Proudyogiki Vishwavidyalaya,
Bhopal Sushila Devi Bansal College, Indore, India
Twinklegurjar321@Gmail.Com, Arjun.Parihar@Sdbct.Ac.In.

***Abstract*- The text classification is a classical research domain and used in a number of real world applications i.e. information retrieval, reviews analysis and others. It is complex task to manage different structure or unstructured nature of data in text mining. Additionally the different size of text documents and their feature extraction and accurate classification of document is also a complicated task. In this paper, the key issues in XML based document classification technique are addressed and an improved XML based text classification technique is proposed to classifying the semi-structured data i.e. XML documents, according to their subjects. Therefore, in order to analyze and recognize the XML based document patterns a hybrid classifier is proposed using Bay's classifier and fuzzy logic. This classifier organizes the decisions of fuzzy based classifier and Bays classifier in terms of weights to improve the classification reliability. This paper includes a survey on different recent contributions and offers a data model for future implementation and experimental study.**

***Keywords*- unstructured data mining, text classification, XML documents, feature extraction, weighted classifier.**

I. INTRODUCTION

Text mining is also known as text analytics or text processing. That technique is used to analyze qualitative or unstructured data using data mining or machine learning based computational algorithms [1]. That usages descriptive data which cannot be measured in form of numbers for example, a photograph can be considered as "qualitative data" but when it is transformed into pixel then it can be measured in terms of quantities. However, there are a number of techniques are available for quantity data analysis by still there are not much effective technique exist for unstructured data analysis [2]. In recent years a rich amount of internet based applications are increased which store and use the text as well as multimedia data (i.e. audio, video or image) [3].

- In an estimate 80% of data is unstructured by nature in these platform
- The unstructured data is frequently used in our daily life such as emails, newspaper, web articles, internal reports, transcripts, research papers, blogs, and applications, and many more.

According to English Dictionary "Oxford", Text mining is the process or practice of examining large collections of written resources to generate new information, using specialized computer software.

Text-mining techniques can used to, categorize information, make links between unconnected documents, and providing visual maps. Text mining involves the application of techniques from various different areas such as [4]:

- Information retrieval (IR) systems match a user's query to documents in a collection or database. The first step is to find the body of documents that are relevant to the search query.
- Natural language processing (NLP) analyzes text in structures based on human speech. It allows the computer to perform a grammatical analysis of a sentence.
- Information extraction (IE) involves structuring the data, and obtains the specific kind of data.

Therefore using the text mining techniques we can prepare a number of very essential applications. In this context the proposed work is focused on study of various technique mining techniques and applications. In next it is tried to design and develop a hybrid text mining technique which suitable for XML document classification. This section provides the overview of the work involved in this paper. The next section highlights some key contributions in the domain of text classification. Additionally based on this review we proposed a text classification method which is also reported in this paper.

II. LITERATURE SURVEY

This section provides the recent studies and noteworthy contributions in the domain of text mining some essential articles are reported here.
Text categorization is required to label the documents automatically with the predefined set of topics. In the proposed system, fuzzy rule along with Bayesian

classification method is proposed by **Swathi V et al [5]** for text categorization. The method selects the feature subset for each class. Then, these features are applied for the classification. To achieve this, Baggenstoss's PDF Projection Theorem is followed to reconstruct PDF in low- dimensional feature space and build the fuzzy based Bayes classification rule. The significance of this method is that most feature selection criteria such as information gain and maximum discrimination can be easily incorporated. The classification performance is evaluated on different datasets and with the different feature selection methods. The results illustrate that the effectiveness of the method and indicates its wide applications.

Textual dissection can be a very useful for the extraction of useful information. The ideology of textual dissection is the way people think about a text. It is the process where given reviews is classified as positive or negative. A huge amount of reviews is present on the web which can be analyzed to make it useful. It can prove to be useful specifically for marketing, business, polity as allow us to do easy analysis of the subjects. Internet has made it possible for us to connect and find out the opinions dissection. Internet has provided a lot of platform such as Forums, Blogs, and Social networking sites**. Sourav Kunal et al [6]** proposes the use of Tweepy and TextBlob as a python library to access and classify Tweets using Naïve Bayes. This technique is ease out the process of analysis, summarization and classification.

The text is nothing but the combination of characters. Therefore, analyzing and extracting information from such data is complex. **N. Venkata Sailaja et al [7]** present different text mining techniques to discover various textual patterns from the different sources. They also deals with the areas i.e. information retrieval, machine learning, statistics, computational data sciences and advanced data mining. They also discuss future challenges using different techniques, particularly rough set based techniques, improvements and directions.

Most of information is stored in the form of texts with the high availability of information. The categorization of documents is a method for managing, organizing information and knowledge. Text classification is the task of assigning predefined categories. The challenge of text classification is accuracy of classifier and high dimensionality. To overcome using Feature Selection is used. It is a process of identifying a subset of the most useful features from the entire set of features**. Pradnya Kumbhar et al [8]** surveys of text classification, several approaches of text classification, feature selection methods and applications of text classifications.

Being one of the most linguistically rich languages, Azerbaijani has been researched less in the context of natural language processing area. The text corpus created from Azerbaijani news articles is designed to apply supervised machine learning approaches for the case of automatic news labeling. Chi-squared test and LASSO methods have been implemented by U. Suleymanov and **S. Rustamov [9]** for feature selection and pre-processing. The application of supervised machine learning approaches to the text corpus allowed us to compare the performance results of well- established supervised machine learning approaches in the domain of Azerbaijani language.

Ashutosh Adhikari et al [10] present, the application of BERT to document classification. A few characteristics of the task might lead one to think that BERT is not the most appropriate model: syntactic structures matter less for content categories, documents can often be longer than typical BERT input, and documents often have multiple labels. Nevertheless, authors show that a classification model using BERT is able to achieve the state of the art across four datasets. To address the computational expense with BERT inference, we distill knowledge from $BERTlarge$ to small bidirectional LSTMs, reaching $BERTbase$ parity on multiple datasets using 30× parameters. The contribution of the paper is improved baselines to provide the foundation for future.

The global pandemic has made it more important than ever to quickly and accurately retrieve relevant scientific literature for effective consumption by researchers in a wide range of fields**. B. J. Gutierrez et al [11]** provide an analysis of multi-label document classification models, a collection of 23,000 research papers regarding the coronavirus. Author find that pre-trained language models fine-tuned on this dataset outperform all other baselines and BioBERT surpasses the others by a small margin with microF1 and accuracy scores of around 86% and 75% respectively. We evaluate the efficiency and generalizability of these models to deal with an urgent situation like the health crisis. Finally, explore 50 errors made by the best performing models and find that they often (1) correlate certain labels too closely and (2) fail to focus on discriminative sections of the articles.

With the continuous development of Internet and rich resources emerging on the Web, information retrieval based on XML has emerged; A new method SCSimilarity is proposed by **R. Xueli et al [12]** to compute similarity of XML documents from structure and content. XML document is expressed as a collection of tuple, the paths are extracted and delete the recurring to improve efficiency, and matching fuzzy path using dynamic programming and WordNet; and then the structure similarity between documents are calculated. The content similarity is estimated by set matching. Finally, the similarity of XML is estimated.
The experiments show that the method is effective, for 1 test structural similarity; the information retrieval is test using automatically generated documentation sets and real

data sets, and results show the accuracy may arrive at 95%.

The world is dealing with heterogeneous, big, diverse and dynamic documents for storing data across clusters of machines. Among them XML documents are store huge amount of information with semi-structure. XML clustering, as a part of knowledge mining requires highly efficient algorithms for managing and analyzing extremely large data. Many researchers have proposed clustering of XML documents according to its content and structure.

A. Thulasi et al [13] provide an exhaustive review of structure-based XML clustering and the research gaps.

Classification models are used in component content management to identify content for retrieval, reuse and distribution. Intrinsic metadata, used as assigned class. With the increasing demand for efficient classification of content, the sector of technical documentation needs mechanisms that allow for automation. Vector space model based approaches can lead to sufficient results, while maintaining good performance, but they must be adapted to the peculiarities that characterize modular technical documents.

J. Oevermann et al [14] will present domain specific differences, as well as characteristics, to the field of technical documentation and derive methods to adapt classification and retrieval techniques. Authors verify the approach with data provided from companies in the sector of manufacturing and mechanical engineering.

III. PROPOSED WORK

The information is available in different formats. Using Information Retrieval (IR), it is possible to obtain unstructured information. Zulfany Erlisa Rasjid et al [15] focused on data classification using two approaches of classification kNN (k-Nearest Neighbors) and Naïve Bay's for XML. It is downloaded from TREC Legal Track more than three thousand documents. The six types of text documents are chosen and processed using RapidMiner. The result shows that k-NN at k=13 offers accuracy of 55.17 %, which is better than Naïve Bayes which offers 39.01% accuracy. That is a significant achievement for XML format data. By using this article some key issues are addressed for this system improvement:

- XML is semi-structured format which need additionally attention to deal.
- Need to be obtain better feature selection strategy.
- Improvement on accuracy of the existing system.

In order to deal with the above discussed issues for improving the existing text classification model a new methodology is proposed. The proposed data model is demonstrated in figure 1, additionally their functional aspects are discussed as:

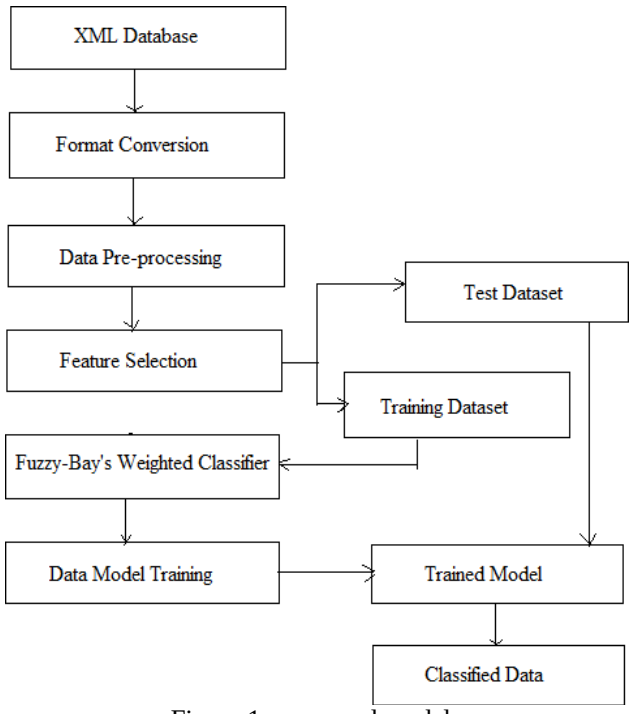

Figure 1. proposed model.

The proposed model is aimed to classify the XML documents, therefore the similar source of data is going to use for preparing the XML document database. The used dataset is a labeled database which contains data as well as the class labels, which is need to recognize during classification. In next step the XML data is transformed into fully unstructured format. Therefore the parsing of XML data is performed and only contents of the XML is extracted. The extracted content is free from the structural information of XML with class labels. Thus the content is used with preprocessing techniques to make it clean and noise free.

The preprocessed or enhanced quality of data is used in next process to select the features from the text contents. So, TF- IDF (term frequency – inverted document frequency) is calculated. Using the TF-IDF the weights of document tokens are calculated for preparing the feature vectors. The calculated feature vector is split into two parts namely training and testing. The 70% of data is used for training and remaining 30% of data is used for testing. The training vector is used for training of the fuzzy model and also bay's model. Both the trained models are used for classifying the 30% test data. The made decisions based on both the models are used for calculating the weights for decisions. The finally computed weights are used to predict the class labels for the test data on that behalf we compute the performance of system.

IV. CONCLUSION & FUTURE WORK

The proposed work is motivated to explore the techniques of text data mining and designing an accurate method for classifying semi structured data. in this context first a survey is carried out which provide the understanding about the developed techniques and methods. Additionally based on the observational facts of reviewed articles some issues in text processing is addressed in this paper. Furthermore to improve the existing models a hybrid weighted classifier is proposed for implementation. The proposed model includes the fuzzy rule based classification as well as Bayesian classifier for more accurate classification. This model is implemented in near future and their performance is reported. The proposed work can also deliver the following outcomes after successfully implementation of the proposed methodology.

- Enhanced accuracy for semi-structured data.
- An improved feature selection strategy to deal with semi-structured data.
- Reducing the efforts and complexity for handling the semi-structured data.

REFERENCES

[1] R. Talib, M. K. Hanif, S. Ayesha, F. Fatima, "Text Mining: Techniques, Applications and Issues", International Journal of Advanced Computer Science and Applications, Vol. 7 No. 11, 2016

[2] D. Strohmeier, S. Jumisko-Pyykko, K. Kunze, "Open Profiling of Quality: A Mixed Method Approach to Understanding Multimodal Quality Perception", Hindawi Publishing Corporation Advances in Multimedia Volume 2010, Article ID 658980, 28 pages

[3] B. B. Gupta, Shingo Yamaguchi, Dharma P. Agrawal, "Advances in Security and Privacy of Multimedia Big Data in Mobile and Cloud Computing", Multimed Tools Appl (2018) 77:9203–9208 DOI 10.1007/s11042-017- 5301-x

[4] S. Sathya, Dr. N. Rajendran, "A Review on Text Mining Techniques", International Journal of Computer Science Trends and Technology (IJCST) – Volume 3 Issue 5, Sep-Oct 2015

[5] V Swathi, S. S Kumar, Dr. P. Perumal, "A Novel Fuzzy- Bayesian Classification Method for Automatic Text Categorization", 2017 IJSRST | Volume 3 | Issue 1 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X

[6] S. Kunal, A. Saha, A. Varma, V. Tiwari, "Textual Dissection Of Live Twitter Reviews Using Naïve Bayes", Procedia Computer Science 132 (2018) 307–313

[7] N. Venkata Sailaja, L. Padmasree, N. Mangathayaru, "Survey of Text Mining Techniques, Challenges and their Applications", International Journal of Computer Applications (0975 – 8887), Volume 146 – No.11, July 2016

[8] P. Kumbhar, M. Mali, "A Survey on Feature Selection Techniques and Classification Algorithms for Efficient Text Classification", International Journal of Science and Research (IJSR), Volume 5 Issue 5, May 2016

[9] U Suleymanov and S Rustamov, "Automated News Categorization using Machine Learning methods", Aegean International Textile and Advanced Engineering Conference (AITAE 2018)

[10] A. Adhikari, A. Ram, R. Tang, J. Lin, "DocBERT: BERT for Document Classification", arXiv:1904.08398v3 [cs.CL] 22 Aug 2019

[11] B. J. Gutierrez, J. Zeng, D. Zhang, P. Zhang, Y. Su, "Document Classification for COVID-19 Literature", arXiv:2006.13816v2 [cs.IR] 9 Sep 2020

[12] R. Xueli, D. Yubiao, "Research on Similarity for XML Information Retrieval", International Conference on Education, Management, Computer and Society (EMCS 2016) © 2016 - Published by Atlantis Press

[13] A. Thulasi, K T V Remya, G. Raju, "Structure Based XML Document Clustering: A Review", 978-1-5386-0514-1/17/$31.00 ©2017 IEEE

[14] J. Oevermann, W. Ziegler, "Automated Intrinsic Text Classification for Component Content Management Applications in Technical Communication", DocEng '16, September 12 - 16, 2016, Vienna, Austria c 2016 Publication rights licensed to ACM. ISBN 978-1-4503- 4438-8/16/09

[15] Z. E. Rasjid, R. Setiawan, "Performance Comparison and Optimization of Text Document Classification using k- NN and Naïve Bayes Classification Techniques", 2nd International Conference on Computer Science and Computational Intelligence 2017, ICCSCI 2017, 13- 14 October 2017, Bali, Indonesia, Procedia Computer Science 116 (2017) 107–112