JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 15, 2020

STUDY OF TOPIC MODELLING TECHNIQUE FOR BIOMEDICAL DOCUMENT CLASSIFICATION

Mr. P. Boopathi[1], Dr. V. Kathiresan[2]

[1]Research Scholar, Department of Computer Science, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore.
[2]Director, Department of Computer Applications (PG), Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore..

Received: 14 March 2020 Revised and Accepted: 8 July 2020

ABSTRACT: - The data has been exponentially increased in the field of bioinformatics over the last few decades, so it is a difficult task for a user to find the appropriate data on the basis of the user requirements for decision taking. Bioinformatics problems are the collection, management, and study of large volumes of medical datasets. The automated classification of medical records into predefined groups is increasingly increasing in electronic data repositories, one of the main challenges inspired to help experts locate valuable knowledge from a vast number of centralized databases of documentation.
Topic modelling techniques help to remove unknown topics from a vast range of papers, articles available and discover the distributions of topics for each paper. Theme models discover the topics from documents that are described by word distribution. Topic modelling is a process in which documents are a mixture of topics transformed into probability over the distribution of terms. Subject modelling is an effective technique for biomedical text mining but requires some enhancement as biomedical text documents are terms that are repetitive and redundancy is a negative influence on thematic modelling and text mining. Therefore, in this research, we proposed a topic modelling technique for biomedical documents through machine learning perspective. To ameliorate problems including sparsity, redundancy and high-dimensionality issue for biomedical text documents, in this research, a topic technique with machine learning perspective is proposed. In the proposed approach proposes four stages. Stage one search and retrieval of documents and pre-processing of documents, stage two represents the NLP with LDA, stage three is nothing but an text clustering, and the final thing is for text classification are implemented. The proposed technique shows higher performance for both redundant and non-redundant biomedical text documents. The proposed technique performance is evaluated on different real-world datasets.

KEYWORDS: - Biomedical document, Topic Modelling, Clustering, Classification, machine learning.

I. INTRODUCTION

Data mining has become a crucial research area for applications in the life sciences and healthcare. It focuses on the methodologies and processes for extracting useful patterns, insights, or knowledge from large amounts of data automatically or semi-automatically. Its importance has increased with the rapid growth of online information on the Web as well as the databases used by industries (e.g. different biological or clinical databases), resulting in an increasingly urgent need to derive novel knowledge from the large amounts of data. The collection and search of publications or documents plays a key role in the field of biomedicine because of their unstructured data format and are not grouped according to the keywords [1].

Text mining is a process in which we extract a variety of patterns from textual sources and discover useful facts. It is difficult to extract useful information from large data collections, such as those generated by a surveillance system for residential health care. Unsupervised machine learning algorithms provide powerful tools to find patterns and hidden data structure. In addition, probabilistic models allow for modeling uncertainty in the data and underlying patterns. Specifically, topic modelling algorithms originally developed to discover hidden themes in a large corpus of text documents have gained popularity and used patterns in a variety of data sources [3]. The resulting distributions that were discovered in the data can be used to analyze, summarize, search, cluster and explore the original large, complex dataset.
The process of extracting text goes through several stages. The document is collected from different sources in the first step than the format of the document is verified in the next step and then it goes through the stage of analysis. The analysis stage includes the semantic analysis and other techniques for processing data in accordance with the requirements. The result of this phase can be stored for further processing in the data base management system. Text mining aims to get information from various textual sources that is not available before. A topic model is a frequently used mathematical text mining model for the discovery of abstract subjects

occurring in a body of text. It is mostly used as a systematic tool in many applications to structure huge textual corpora, such as data mining, text mining, and image retrieval [4].
We describe in this work our exploration of a vastly untapped set of biomedical documents. To start by using topic modeling to characterize the whole corpus of notes.

II. RELATED WORK

Several approaches for solving the problem of biomedical document classification research are identified in this literature review. But each approach also has its own traits and deficiencies. This section summarizes the review of the issue by literature.

Generating substantial quantities of low-cost, high-quality next-generation sequences (NGS) has empowered mainstream researchers to address various problems in biological and medical research. The vast information that NGS technologies deliver presents a major challenge for data processing, analysis, data mining, and text mining [2]. The existing work uses Probabilistic Latent Semantic Analysis (PLSA) to propose a novel topic modeling technique for the NGS data analysis. The proposed method has four tasks: NGS dataset construction, data preprocessing, theme modeling, and text mining using topic outputs from PLSA. In this procedure the NGS data of Salmonella enteric strains was used as the dataset. The performance of the topic modeling is measured using standard clustering comparison measures such as Adjusted Rand Index, Normalized Mutual Information, Standardized Distance to Information and Normalized Variation of Information. [6].

The author explore electronic patient records in the form of unstructured clinical notes and genetic mutation test results using a variety of techniques including Topic Modeling, Principal Component Analysis, and Bi-clustering. The ultimate goal is to gain insight into a unique set of clinical data, specifically the topics discussed within the content of the note and the relationships between patient clinical notes and their underlying genetics. [7].

Digital libraries, journals, and repositories of conference proceedings are a great source of information. These sources are very useful for research and development purposes. The existing work presents an overview of text mining and its use for the extraction of information from literature. In this study, to extract information from multi-domain research articles, we used word cloud, term frequency analysis, similarity analysis, cluster analysis and topic modeling. Cloud computing and big data are new trends which are emerging. It is therefore important to extract useful patterns and knowledge in these domains from published articles, and to discover the relationship between them. Finding the latest trends, related topics, tools, terms, and author-affiliation from extracted data is a cross-domain analysis in cloud computing and big data domains. This study uses cloud computing to identify the ten major areas of big data, fourteen factors for cloud adoption and hurdles in adoption. [8].

Text data plays an essential role in the field of biomedicine. As patient data includes an enormous amount of text documents in an un-standardized format. The text documents pose a lot of challenging issues to data processing in order to obtain the relevant data. Topic modeling is one of the most popular information retrieval techniques based on themes from biomedical papers. Discovering the exact subjects from the biomedical documents is a challenging task in topic modeling. In addition, redundancy places a negative impact on the quality of text mining in biomedical text documents as well. The rapid growth of unstructured documents therefore involves machine learning techniques for topic modeling which are capable of discovering precise topics. In this paper we proposed a topic modeling technique for text mining through the hybrid inverse document frequency and the clustering algorithm for machine learning fuzzy k-means. The technique proposed improves the issue of redundancy and discovers precise topics from the biomedical text documents. The existing work technique generates frequencies of local and global terms via the Bag-of - Words (BOW) model. The global term weighting is calculated using the proposed hybrid inverse frequency of documents and the term frequency is calculated for local term weighting. The robust analysis of main components is used to eliminate the negative impact of higher dimensionality on global term weights. Subsequently, text mining classification and clustering is performed with a likelihood of topics in the documents. Classification is performed by discriminating analysis classifier whereas the clustering is performed by the clustering of k-means. Clustering performance is assessed using the internal validation method Calinsiki-Har-abasz (CH) index. [9].

Topic modeling becomes a popular area of research that shows us a new way of searching, browsing and summing up a large amount of texts. The topic modeling methods try to uncover the hidden thematic structure in the collections of documents. Topic modeling in connection with social networks, which is one of the strongest communication tools and produces a large amount of opinions and attitudes about world events, can be useful for analysis in case of crisis situations, elections, market launch of a new product etc. This is why we are promoting in this paper a tool for topic modeling over text streams from social networks. Proposed tool description is extended with practical experiments. Realized experiments showed promising results when compared with state-of-the-art methods, when using our tool on real data. [10].

The author develops a novel online algorithm, namely the moving average stochastic variational inference (MASVI), which applies the results obtained from previous iterations to smooth out noisy natural

gradients. The researcher analyzes the convergence property of the proposed algorithm, and performs a set of experiments on two large-scale collections containing millions of documents. Experimental results indicate that our algorithm achieves a faster convergence rate and better performance, as opposed to algorithms called 'stochastic variational inference' and 'SGRLD'. [11].

The researcher analyzes the feelings deriving from the social networking conversations. The main aim is to identify users' feelings within the social network through their conversations. In order to conduct a study to determine if social network users (especially twitter) tends to gather according to the similarity of their feelings. In the proposed framework, (1) we use ANEW, a lexical dictionary to identify affective emotional feelings associated with a message according to Russell's affection model; (2) To design a topic modeling mechanism called Sent LDA, based on the generative Latent Dirichlet Allocation (LDA) model, which allows us to find the topic distribution in a general conversation and we associate topics with emotions; (3) To detect communities in the network according to the density and frequency of messages between users; and (4) To compare the feelings of the communities using the Russell model of affect versus polarity and we measure the extent to which the distribution of topics reinforces similarity in the feelings of community users. This work contributes to the analysis of the sentiments in conversations taking place in social networks with a topic modeling methodology. [12].

III. STUDY OF EXISTING APPROACH

Text data plays an essential role in the field of biomedicine. As patient data includes an enormous amount of text documents in an un-standardized format. The text documents pose a lot of challenging issues to data processing in order to obtain the relevant data. Topic modelling is one of the popular information retrieval techniques based on biomedical paper themes. Discovering the exact topics from the biomedical documents is a challenging task in the topic modelling [5]. In addition, redundancy places a negative impact on the quality of text mining in biomedical text documents as well.

A topic modelling technique for text mining by means of hybrid inverse document frequency and machine learning Fuzzy k-means clustering algorithm in the existing approach. The existing technique improves the issue of redundancy and discovers precise themes from the biomedical text documents. The current methodology produces frequencies on local and global terms through the Bag-Of - Words (BOW) model. The global term weighting is calculated using the proposed hybrid inverse frequency of documents and the term frequency is calculated for local term weighting. The robust analysis of main components is used to eliminate the negative impact of higher dimensionality on global term weights. Subsequently, text mining classification and clustering is done with a probability of topics in the documents. Classification is done by discriminating classification classifier while the clustering is conducted by the k-means clustering.

Biomedical text documents are constantly growing nowadays, thus the study of these documents is very important for the discovery of the useful information tool. Archives of scientific text records such as PubMed provide important Scientific Community resources. Topic Modelling is a common method that discovers in unorganized biomedical text documents the hidden theme and structure. The layout of these records is used for the scan, indexing and description of records [13].

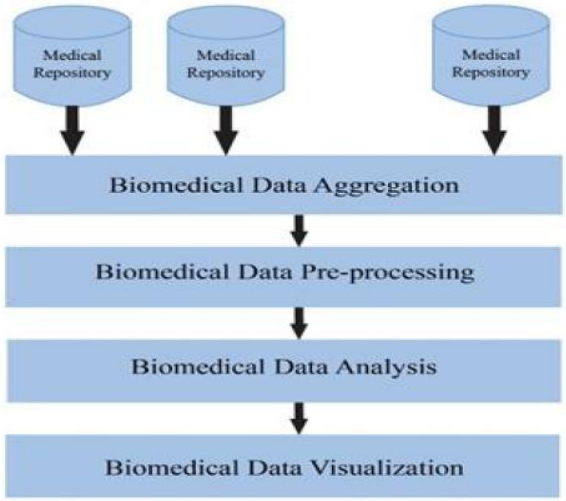

Figure 1: - Existing Architecture

The main contributions of this research are listed below.

- A new inverse hybrid document frequency for global term weighting is proposed. The current modelling methodology for the subject induces weighting of local and global terms via the bag-of - words model. Global term weighting process helps in the probabilistic filtering of rising high-frequency terms.
- With the robust master component analysis dimension reduction method, the high dimensionality negative effect on global term weighting is removed. A fuzzy k-means clustering algorithm is employed after that.
- Compared to state-of-the-art theme modelling techniques, the latest topic modelling technique discovers the more reliable and relevant topics from biomedical text documents.

3.1. Disadvantages of Existing Approach

Techniques for topic modelling are used to summarize a large collection of text documents. Probabilistic modelling techniques are used to classify the key topics from the set of documents in the biomedical text. The new strategy is comprised of the following setbacks.

- Existing approach focuses on linear algebra and statistical distribution methods in the subject modelling techniques.
- The clustering of Fuzzy k-means is highly prone to successful initialization.
- The current algorithm that generates coincident clusters as the columns and rows of the typicality matrix are independent from each other. When each row's initialization is not sufficiently distinct it may lead to clusters of coincidences.
- The current topic modelling technique is very weak in removing redundancy in biomedical text documents and its log-like efficiency is much lower..

IV. STUDY OF PROPOSED APPROACH

With the exponential rise in the number of papers published per year in the biomedical domain, the main goal of the proposed solution is to develop automated systems to retrieve unknown knowledge from the published articles. Text mining techniques allow for the extraction from unstructured documents of unknown information. Text mining is the process of extracting new knowledge from a collection of documents on a given subject. Text mining is useful where the data is in the form of an unstructured text (document) that cannot be processed using conventional methods such as data mining methods. Text mining is distinct from standard search queries because it is often useful when missing information from a collection of documents is discovered. Text mining is based on the processing technique of natural languages which helps computers to understand and process human language.

Biomedical researchers have started using text mining techniques in the form of journal papers, case reports, Electronic Health Records (EHRs), and so on, owing to the vast amount of unstructured knowledge available in the biomedical domain. Regardless of their unstructured data format, gathering and searching articles or documents plays a key role in the field of bioinformatics and they are not classified by the keywords. The data has been exponentially increased in the field of bioinformatics over the last few decades, and it is a difficult task for a user to find the appropriate data on the basis of the user requirements for decision taking.

In bioinformatics an ecosystem turning case-based experiments into large-scale, data-driven big data analysis. Bioinformatics problems are the collection, management, and study of large volumes of medical datasets. The automated classification of medical records into predefined groups is that rapidly online data repositories, one of the biggest challenges that help experts find valuable knowledge from a vast number of distributed databases of records.

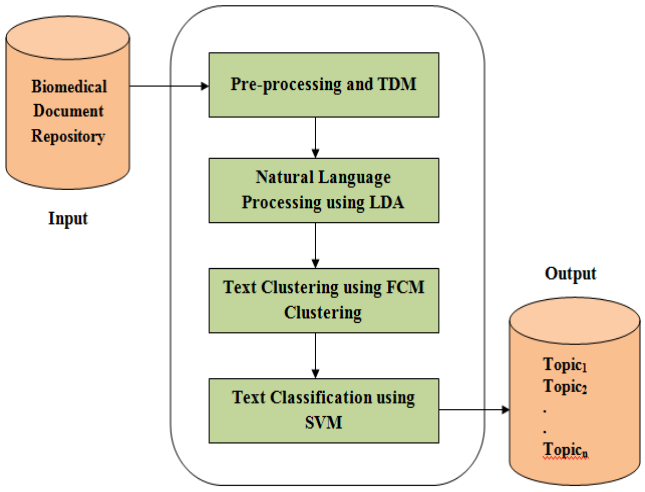

Figure 1: - Workflow of the Proposed Approach

The proposed topic modelling methodology explores the more reliable topics that shape biomedical text documents and eliminates the issue of redundancy from those documents. In addition, it can be used for classification of biomedical documents and clustering tasks in text mining. It also reduced the expense of time use when finding topics from twitter's major health news dataset. The method suggested includes the following phases.

Phase 1: - Pre-processing of Biomedical Documents

`Report pre-processing includes stopword deletion and stemming measures. The classic method excludes predefined stop terms, and the Zip law method eliminates terms with the value of high Term Frequency-Inverse Document Frequency (TF-IDF) and words that only appear once in the text. The next step after the elimination of stop words is 'stemming,' which allows one to use the roots of terms only. Upon completion of the pre-processing, the next step is to prepare the term document matrix (TDM), in which words are represented by rows, and columns represent documents.

Phase II: - Topic Modelling using LDA

Topic modelling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body. The latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. It is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modelled as Dirichlet distributions. Here we are going to apply LDA to a set of documents and split them into topics.

Phase III: - Text Clustering using FCM Clustering

Documents are grouped according to their document vector, and each cluster is denoted by the document vector name. Clustering of documents can be carried out using technique called FCM clustering is one of well-know unsupervised clustering techniques. However FCM algorithm requires the user to pre-define the number of clusters and different values of clusters corresponds to different fuzzy partitions.

Phase IV: - Text Classification using SVM

Text Classification is an automated process of classification of text into predefined categories. Documents can be automatically classified into specific categories using classifier algorithm SVM. SVM is an effective technique for classifying high – dimensional data. SVM avoids the costly similarity computation in high-dimensional feature space by using a surrogate kernel function. It is known that support vector machines (SVM) are capable of effectively processing feature vectors of some 10 000 dimensions, given that these are sparse.

The proposed topic modelling technique reduces numbers of features and finds meaningful words in topics. Optimization is the process of determining points that reduce the actual value function, called the objective function.

V. PERFORMANCE EVALUATION

To evaluate the performance of the proposed approach is the mandate one. The following performance evaluation metrics are used to the entire approach [14].

- **Precision:** the percentage of texts the classifier tagged correctly out of the total number of texts it predicted for each topic.
- **Recall:** the percentage of texts the model predicted for each topic out of the total number of texts it should have predicted for that topic.
- **F1-Score:** It is the harmonic mean of precision and recall.
- **Accuracy:** It represents the number of correct predictions by a model.
- **Cohen's Kappa Score:** It gives an estimate of how much better a classifier does over another one that just predicts and classifies randomly based on frequencies.
- **Mathew's Correlation Coefficient (MCC):** It gives the correlation between two values. It ranges between -1 to 1 where -1 indicates a very bad prediction and 1 represents a very good prediction.

VI. CONCLUSION

This research introduces the essential needs of Supervised and Unsupervised learning for medical data classification. The proposed a multistage approach based on consensus function clustering for medical data classification. The system involves three main stages. In the first phase implies dimensionality reduction algorithm, and the second phase implemented the multiple unsupervised clustering algorithms and finally, these fast classification algorithms classified the whole data set in such a way that efficient and accurate profiling of very large and highly dimensional medical data sets can be achieved.

VII. REFERENCES

[1] Mirończuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, *106*, 36-54.

[2] Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, *3*(Mar), 1289-1305.

[3] Sriurai, W. (2011). Improving text categorization by using a topic model. *Advanced Computing*, *2*(6), 21.

[4] La Rosa, M., Fiannaca, A., Rizzo, R., & Urso, A. (2015). Probabilistic topic modeling for the analysis and classification of genomic sequences. *BMC bioinformatics*, *16*(6), 1-9.

[5] Wang, F., Orton, K., Wagenseller, P., & Xu, K. (2018). Towards understanding community interests with topic modeling. *IEEE Access*, *6*, 24660-24668.

[6] Annavarapu, C. S. R., & Mohapatra, A. (2019, February). A novel method for next-generation sequence data analysis using PLSA topic modeling technique. In *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)* (pp. 1-6). IEEE.

[7] Chan, K. R., Lou, X., Karaletsos, T., Crosbie, C., Gardos, S., Artz, D., & Rätsch, G. (2013, December). An empirical analysis of topic modeling for mining cancer clinical notes. In *2013 IEEE 13th International Conference on Data Mining Workshops* (pp. 56-63). IEEE.

[8] Haq, M. I. U., Li, Q., & Hassan, S. (2019). Text Mining Techniques to Capture Facts for Cloud Computing Adoption and Big Data Processing. *IEEE Access*, *7*, 162254-162267.

[9] Rashid, J., Shah, S. M. A., Irtaza, A., Mahmood, T., Nisar, M. W., Shafiq, M., & Gardezi, A. (2019). Topic Modeling Technique for Text Mining Over Biomedical Text Corpora Through Hybrid Inverse Documents Frequency and Fuzzy K-Means Clustering. *IEEE Access*, *7*, 146070-146080.

[10] Smatana, M., Paralič, J., & Butka, P. (2016, September). Topic Modeling over Text Streams from Social Media. In *International Conference on Text, Speech, and Dialogue* (pp. 163-172). Springer, Cham.

[11] Li, X. M., Ouyang, J. H., & Lu, Y. (2015). Topic modeling for large-scale text data. *Frontiers of Information Technology & Electronic Engineering*, *16*(6), 457-465.

[12] Naskar, D., Mokaddem, S., Rebollo, M., & Onaindia, E. (2016, May). Sentiment analysis in social networks through topic modeling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 46-53).

[13] Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in bioinformatics*, *6*(1), 57-71.

[14] McCallum, A., Corrada-Emmanuel, A., & Wang, X. (2005). Topic and role discovery in social networks.