European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 07, Issue 03, 2020

A Probabilistic Key phrase extraction approach on large biomedical documents

Jose Mary Golamari[1], D. Haritha[2]

[1]*Research Scholar, Department of Computer Science and Engineering,, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra pradesh.*

[2]*Professor, Department of computer science and engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh.*

Abstract— As the size of the biomedical databases are increasing day-by-day, finding an essential feature set for classification problem is complex due to large data size and sparsity problems. Text feature ranking and clustering is one of the major challenges to scientific and medical researchers due to its high dimensional feature space and limited number of samples. High dimensionality of the feature space is one of the major issues in biomedical document clustering due to large number of candidates sets. Selection of high probabilistic features for clustering is therefore essential for biomedical document analysis such as classification and clustering. In this paper, a novel probabilistic key phrase extraction and preprocessing model is designed and implemented on large number of biomedical documents. In this framework, a novel key-phrase extraction method is used to filter the large biomedical document sets. Experimental results show that the present key phrase extraction approach is better than existing key-phrase extraction approaches in terms of runtime and accuracy are concerned.

Keywords: Biomedical documents, gene. protein entities, probabilistic key phrase extraction.

I. INTRODUCTION

Presently, there exist around 29 million articles in the PubMed database. Every year the volume of my medical literature is increasing about 1 million. Large quantity of important knowledge related to proteins, drugs, diseases and chemicals is present in the unstructured format. Because of this exponential growth of documents, it is very complicated to manually gather and organize information out of biomedical literature just like protein-protein interaction, drug-drug interaction and chemical-protein interaction. Biomedical information extraction can be defined as process responsible to identify biomedical concepts and their relations automatically with the help of advanced language processing and machine learning methods. Every year, biomedical scientists and healthcare professionals publish vast quantity of biomedical research articles. Almost all of these articles are available online. These articles not only help biomedical scientists in their research work, but also help healthcare professionals in their clinical practices.

The volume of information is growing rapidly in different domains with the growth of distributed biomedical repositories. Document preprocessing is a reductive transformation of peer documents to generate summary by selecting an important information in the source documents. However, this has caused the problem of information overload. In order to resolve this drawback, multi-document clustering and feature extraction can be used to minimize the inter-cluster variation. This work considers

the feature extraction strategy and the key phrase clustering and pattern discovery approach to eliminate information redundancy resulting from the multiple original documents. Clustering is one of the vital tools in data mining and knowledge discovery. Due to massive amounts of data collected in databases, cluster analysis has been recently become a highly active topic in data mining. Mostly K-means clustering algorithm is used to group similar objects together. But it requires the number of clusters to be specified in advance which is considered to be one of the problems of this algorithm. The ability to automatically cluster similar items together, allows one to determine hidden similarities and key concepts. It also summarizes a large amount of information into a smaller number of clusters. Therefore, biomedical literature has become more complicated for understanding. Thus, there is a necessity of more efficient approaches in order to extract biomedical information from vast numbers of resources. An appropriate mining approach is required to be implemented in order to discover different types of knowledge from biomedical literature. In biomedical texts we can find a degree of term variation. Apart from this, biomedical term can contain numbers, capital letters inside words, hyphens and different special characters

A group of researchers developed an advanced all path kernel technique in order to retrieve PPIs depending upon several lexical and syntactical features. In the subsequent time, approaches those depend upon deep neural network such as convolutional neural networks and recurrent neural networks have become more popular and widely accepted. In case of support vector machine, we have to select the neighboring word features, bag-of-words features, distance features, keywords feature and shortest-path features. In case of CNN-based approach, sentence sequence and shortest dependency paths are the inputs. But in case of RNN-based approach, there is only one input that is sentence sequence. The word, part of speech, position and embeddings are considered as the input representation in case of CNN and RNN schemes. At last, the majority voting scheme is implemented on SVM, CNN and RNN schemes. Some other researchers combined RNN and CNN approaches in order to present an extended and advanced hybrid approach. We can mention here that, the inputs for this model are sentence sequences and SDPs produced from the dependency graph. The process of relation extraction is considered as the most important category of knowledge discovery. The most common and prime objective of all researchers is to detect relationships in between different biomedical concepts. Different numbers of approaches are implemented in the biomedical relation extraction process such as co-occurrence statistics, rule-based techniques, pattern learning and classification

Let us consider the Human Genome Project that involves sequences of human genes. Human Genome Project is considered as the most complicated and difficult area of research. In the subsequent phase, the process of genome analysis needs to define the function of every individual gene. Apart from this, it also has the responsibility to identify its interactions with other genes. Gene-gene interaction resources gathered from different databases just like MIPS, EcoCyc and KEGG. Large numbers of resources are not at all cataloged. Most of the information are in natural language which is not at all understandable by computers. Therefore, it is very much essential to develop efficient and effective information extraction approaches in order to process vast quantities of unstructured data. Since few decades, research works have been carried out in order to produce patterns of information extraction biomedical documents. The above-mentioned pattern sets depend upon different interaction verbs and gene names. These patterns can give rise to very high precision and poor recall value.

In the domain of biomedicine, the production of new knowledge is exponential and continuous. Hence, the biomedical literature is getting much complicated and time consuming to understand in case of non-professionals. Therefore, it is very much necessary to develop an efficient approach in order to extract paramedical information out of the existing resources. In the recent era, there have been different advanced approaches developed in order to discover, access and share knowledge out of medical literature. We can find very high degree of term variation case of biomedical texts. In the lexical perspective, terminology is considered as the most difficult challenge during the process of text mining because of continuous generation of neologisms and presence of term abbreviations. Biomedical term can contain numbers, letters, words, special characters, etc. Implementation of text mining on high altitude diseases of biomedical literature can result potential insights in order to detect potential targets for this sickness. In other words, the process of text mining is defined as a data mining algorithm which can be implemented on textual data. The prime objective of the text mining process is to detect non-trivial, implicit, unknown and very much useful patterns.

Most of the biomedical data are present in heterogeneous format and these data are known as unstructured text data. The quantities of unstructured text data is growing day by day. Hence, it is essential to implement a significant knowledge extraction approach. This knowledge extraction process follows the basic concepts of text mining in order to influence the process of scientific hypothesis generation and knowledge discovery. Some of the mostly used text mining processes are:- named entity recognition, text classification synonym and abbreviation extraction, relationship extraction and hypothesis generation. The named entity recognition process has the responsibility to detect particular names just like gene, protein, drug, chemical out of vast range of text. The process of text classification can be defined as a specific process that can automatically identify importance of a particular document. Apart from all of these, emphasis is also given on recognizing synonyms and term abbreviations. Another important process is the process of relationship extraction. In the process of relationship extraction emphasis is given on the detection of occurrences of a pre-specified relation just like associative, chemical or regulatory relations among various entities. Again, process of hypothesis generation has the responsibility to produce additional interesting relationships those are not directly mentioned before.

2. RELATED WORKS

C. Olsen, K. Fleming, et.al, emphasized on inference and validation of predictive gene networks out of biomedical literature and expression data [1]. Till today they are have been large numbers of approaches introduced in order to explain the inference of biological networks. Again, the validation process of the above models is also a serious issue. In this piece of research work, they introduced an advanced framework for the quantitative assessment of gene interaction networks. They have used knockdown data out of several cell line experiments. By implementing the above presented framework, we can state that, network inference that depends upon combination of pre-existing knowledge is capable enough to enhance the overall quality of the inferred networks.

J. M. Ruiz-Martínez, et.al, proposed a new method of ontology learning out of biomedical language documents with the help of UML [2]. Everyday large numbers of new biomedical knowledge is formed continuously in biomedical domains. Therefore, the quantity of biomedical literature is increasing day by day and hence, the complexity of the knowledge is also increasing. Ontologies have the responsibility to provide vocabulary standardization. Thus, these are very much beneficial during the understanding of

biomedical literature. In this research paper, an advanced technique is presented which is responsible for constructing biomedical ontologies out of biomedical literature. Their presented technique completely depends upon natural language processing and evolutionary knowledge acquisition methods. The prime objective of the above two mentioned approaches are to provide various relevant concepts and relationships among them. Apart from this, they have also developed a new algorithm in order to merge various concepts regions with the help of UMLS. Again, they also introduced different approaches in order to represent and exploit archetypes with the help of OWL. The above-mentioned archetypes can be used in order to guide the complete knowledge extraction process. At last, they tried to enhance their system in order to include axioms. In future, additional research works can be carried out in order to extend the above presented model.

Y. Zhang and Z. Lu explored semi-supervised variational autoencoders in case of biomedical relation extraction process [3]. The biological literature has the responsibility to provide different kinds of knowledge just like protein-protein interactions, drug-drug interactions and chemical protein interactions. The process of biomedical relation extraction has the objective to automatically extract different biomedical relations out of biomedical text. Most of the traditional biomedical relation extraction processes depends upon supervised machine learning technique. Hence, all of the above-mentioned methods are based on labelled data. There exist huge quantities of unlabelled biomedical text in the PubMed database. Efficient computational approaches are beneficial for employing unlabelled data in order to decrease the load of manual annotation. They have presented an advanced semi-supervised technique that depends upon variational autoencoder in order to carry out the complete process of biomedical relation extraction. The above presented model can be divided into three important parts, those are: - a classifier, an encoder and a decoder. Multilayer convolutional neural networks are very much important during the classification phase. On the contrary, bidirectional long short term memory networks along with convolutional neural networks play important role during the encoding and decoding phase.

B. Bhasuran, t.al, developed a new and advanced text mining and network analysis approach in order to detect functional associations of genes in case of high-altitude diseases [4]. This piece of research work has the objective detect in every functional association of genes that take part in high altitude sickness. In this paper, they have detected the gene networks those are responsible for the above-mentioned sickness. They have included the basic concepts of gene co-occurrence statistics out of literature and the analysis process. At first, a mining algorithm is implemented on the text data from PubMed database in order to extract the co-occurring gene pairs. In the subsequent phase, according to the co-occurrence frequency, each and every gene pair is ranked. At last, an efficient gene association network is constructed with the help of various statistical measures in order to explore all kinds of potential relationships. In future, full text articles and advanced text mining systems can be used in order to resolve the current issues of this model.

J. Chiang, et.al, identified gene-gene relations out of sequential sentence patterns in biomedical literature [5].In this piece of research work, they have introduced a new gene–gene relation browser (DiGG) which has the responsibility to merge sequential pattern mining algorithms and information extraction approaches. This research work has the major objective to extract relevant knowledge out of biomedical literature. The above proposed approach has the responsibility to integrate advanced mining approaches in order to identify frequent gene-gene sequences. Apart from this, this technique has also the goal to identify

different associated gene relations. Graphical presentations can be considered as the most useful method in order to demonstrate every individual relationship among gene products. One of the major characteristics of this technique is that it results an additional and new frequent gene relation. There are three advantages of this approach, those are mentioned below: -

1. It can be very useful during the assimilation of large quantities of biomedical data,
2. It can assist the scientists in order to understand the complete structure and working procedure of biomedical mechanism by gene-gene relations.
3. It also provides insights into relationships in between genes and proteins in case of gene networks. In future, researchers can correlate protein-protein interactions in order to expand the above presented model.

*M. T. Hassan, et.al,*introduced a new method of document clustering with the help of discrimination information maximization [6]. Document clustering approaches usually generate different clusters those are semantically relevant. Again, those clusters can also be defined as group of documents those are part of a specific context. Most of the previously developed classical document clustering approaches never consider the concept of term-document corpus-based semantics. On the other hand, it includes the generic measures of similarity. In this piece of research work, they introduced a new framework in order to carry out the process of partitional clustering. They termed their proposed framework as CDIM. It has the responsibility to maximize the sum of the discrimination information gathered from documents. Furthermore, this framework exploits the semantic which term discrimination information. It also provides easy understanding of various contextual topics.

S. J. Fodeh et.al, used MEDLINE database in order to carry out the process of gene molecular function prediction through the implementation of NMF based multi-label classification scheme [7]. Gene ontology can be defined as a specific presentation of various terms and categories those are used to describe genes and its molecular functions, cellular components and biological processes. Gene ontology can also be called as a standard which is mostly used to describe the functions of a particular gene in case of model organisms. The complete process of gene annotation is a manual and time consuming process. There are numbers of different automated techniques those are developed what the process of annotation. Some of these approaches use the knowledge gathered from the literature. In this research paper, they explained the development and evaluation of an advanced projective system in order to in molecular functions to genes automatically. As we all know, a single gene can be associated with different molecular functions. Therefore, the researchers considered the above molecular function annotation as a multilevel classification issue having numbers of different classes. They have considered non-negative matrix factorization in order to carry out complete process of feature reduction and classification. To classify multilevel data, they have implemented binary relevance approach.

Y. Ji, implemented the MapReduce algorithm in order to extract associations among different biomedical concepts in case of large text data [8]. Biomedical text data are considered as very important source of information. In this research paper, they demonstrated the use of MapReduce method which is actually a parallel and distributed programming paradigm. It has the responsibility to mine each and every association various biomedical concepts those are extracted out of different biomedical articles. Initially, biomedical concepts are gathered through a matching text process to unified medical language system. Unified medical language system can be defined as the most commonly used standard biomedical

database. In the subsequent step, they introduced a MapReduce method which can be implemented in order to evaluate a specific kind of interestingness measures. The above mentioned method can be divided into two sub-methods. In future, the above proposed method can be extended.

H. Kim and S. Chen proposed a new Naïve Bayes classification algorithm in order to automate the linking of gene ontology to MEDLINE documents [9]. They proposed a new and advanced text mining approach and named it as associative naive Bayes (ANB) classifier. This technique is useful to link MEDLINE documents to gene ontology automatically. This technique is actually a non-trivial extension of document classification approach from a specific set of classes acknowledge hierarchy such as gene ontology. As we all know, the complexity of gene ontology is very high, hence, an efficient knowledge representation structure can be implemented here. With the help of the above mentioned structure, they presented the text mining classifier known as ANB classifier. This classifier has the responsibility to link MEDLINE documents to gene ontology.

X. Ling et.al, produced gene summaries out of biomedical literature [10]. In this piece of research work, they emphasised on a thorough study of different approaches those can automatically produce gene summaries out of biomedical literature. In case of most of the previously existing methods, the produced summary is actually a list of extracted sentences. In this research paper, they have given emphasis to produce a semi-structured summary that will contain different sentences including essential semantics of a gene. We can mention here that, the semi structured summary is more useful to describe genes. They presented a two-phase technique in order to produce the above summary for a specific gene. Initially, the articles related to a particular gene are retrieved. After that, each and every sentence for a particular semantic aspect is extracted. Apart from this, they have also considered the problem of gene name variation in the initial phase. In the next phase, they introduced various approaches in order to carry out the process of sentence extraction.

S. Kim and J. Yoon presented an advanced link-topic model for biomedical abbreviation disambiguation [11]. The ambiguity of biomedical abbreviation is considered as one of the major issues in the domain of biomedical text mining. We can mention here that, the management of term variants and abbreviation without proper definition is a very challenging task. In this research paper, they have considered the topic documents and word link in order to disambiguate abbreviations. They have introduced advanced link topic model with the help of latent Dirichlet allocation model. In the above presented model, every individual document is considered as a random mixture of topics. Again, each and every topic is characterized through a proper distribution process. This model is useful for two different modes of word production in order to incorporate semantic dependencies in between words. We can mention that, the semantic dependencies among different words can be defined as a link. For every individual link, a random parameter is assigned to every single word. By analysing the link status, we can say whether a link is present or not.

Additionally, it has the responsibility to check whether a particular word is forming a unigram or a bigram. In the subsequent time, the above proposed model can be modified and extended.

K. Liu, proposed natural language processing approaches and systems in case of biomedical ontology learning [12]. Now-a-days, the popularity of domain ontology is increasing but, there also exist numbers of different challenges. The most vital need of domain ontology is that, it must result a very high degree of

coverage about the domain concepts and their relationships. We can also say that, the complete development process of the above said ontologies is manual, time consuming and not completely error free. Restricted numbers of resources may lead to missing concepts and their relationships. Hence, it may create problem while updating the ontology at the time of knowledge modifications. The approaches of natural language processing, information extraction, information retrieval and machine learning quite helpful in order to automate enrichment of a particular ontology. In this piece of research work, they thoroughly studied and surveyed most of the previously developed existing approaches. Apart from this, they have also discussed their pros and cons.

T. Theodosiou, carried out the process of gene functional annotation through statistical analysis of different biomedical articles [13]. Process of functional annotation of genes is considered very vital process because it has the capability to affect the characterization of genes relationships. Numbers of different gene functions can be explained through standardized and structured vocabularies which are known as bio-ontologies. The assignment of bio-ontology terms to genes is performed through implementation of several approaches those include data sets gathered from various articles. Most of these approaches are generated from data mining and machine learning techniques. The above mentioned approaches involve maximum entropy or support vector machines algorithm. The prime objective of this research work is to present an alternative approach for functionally annotating genes. This approach includes construction of efficient classification schemes, validation models and graphical representation of the outcomes. Apart from this, dimension reduction of the dataset is also another prime concern of this research work. The classification schemes are developed by considering the basic concepts of linear discriminant analysis approach. On the other hand, the validation models depend upon the concepts of statistical analysis and interpretation of the outcomes.

Z. Wang introduced semantic relation extraction aware of n-gram features out of unstructured biomedical text data [14]. Semantic relation extraction is considered as the most vital phase in order to build a knowledge graph automatically. The above-mentioned graph is built out of unstructured biomedical text. This approach can be implemented in numbers of different real-world applications. We can mention here that, the implementation popularity of unsupervised relation extraction techniques, generative probabilistic schemes, Rel-LDA and Type-LDA is growing day by day. Both of the above-mentioned approaches inherit the bag of word assumption from the classical LDA approach and it restricts the exploitation of n-gram features. In order to resolve the above-mentioned issue, two new approaches are proposed which are known as Rel-TNG and Type-TNG. Topic N Grams approach play important role during the allotment of the above two proposed model.

E. Yan and Y. Zhu tried to track work semantic changes within biomedical literature [15]. The aim of this research work is to describe and analyse word semantic modifications in the biomedical domain. They have also detected couple of representative words in medical literature depending upon word frequency and word topic probability distributions. They have implemented a word2vec scheme in order to detect words to identify different semantic modifications. Based on word level, this research work presents to separate methods in word semantics by analysing distance metrics. We can mention here that, words are capable to generate clusters along with their semantic neighbours. Again, words as a cluster after capability to coevolve semantically. Apart from this, words are capable enough to drift apart from their

semantic neighbours. In future, further researches can be performed in order to support the law of parallel modification or the law of conformity.

N. Zong emphasized on inter-topic entity search for biomedical linked data that depends upon heterogeneous relationships [16]. Most of the keyword-based entity search methods limit search space depending upon preference of search. If the keywords and preferences are not linked to the relevant topic, in that case most of the traditional biomedical linked data search engines become inefficient [17]. The prime objective of this research work is to overcome type of mentioned problem through performing and inter topic search that will no doubt enhance the search process with inputs, keywords and preferences. In this piece of research work, they presented an advanced approach by implementing which relations among by medical entities can be used in tandem with a keyboard-based search [18]. It is the modified and extended version of personalised page rank algorithm. In case of extended keyword-based entity search system, the search preferences play significant role filters in order to restrict relations in linked data [19]. In this piece of research work, they have proposed PRRank algorithm which uses each and every relation presenting linked data. Further works can be performed in order to extend this approach in future [20-21].

3.Proposed Model

Text normalization is responsible for removing the punctuation symbols, tokenizing the terms and splitting the identifiers. Filtration of non-essential words from a stop-word list then follows, and it serves to reduce cases of noisy matches and increase the accuracy. An example could be the words "goes" and "going" whose original forms are the word "go." Stemming represents like words with a common root form are represented using the same word.

In the proposed model, initially all the biomedical gene/protein entities are extracted from the biomedical repositories such as PubMed/Medline. All the input biomedical documents are pre-processing using tokenization, stemming and stop word removal. Each document is pre-processing and its gene/protein features are extracted using the Abner tagger as shown in fig 1.

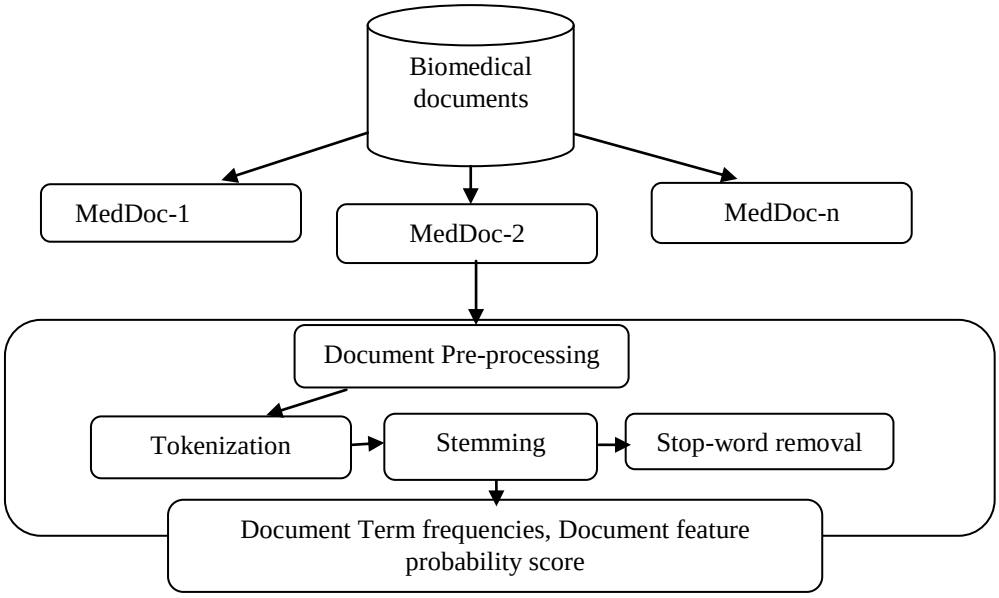

Figure 1: Proposed Framework

After the preprocessing phase, each document and its corresponding gene/protein tags are extracted to find the highest probability biomedical features for the graph initialization process. Here, each biomedical token, gene protein tags, synonym tags are used to find the feature probability ranking.

Hybrid Biomedical Document Preprocessing Algorithm:

Input : Biomedical XML documents D, λ min threshold;

Output: Document frequency terms, document entity probability scores.

Procedure:

1. Read Biomedical PubMed/Medline data in xml format.
2. Extract each document PMID in the xml file.
3. For each document d_i in D[]
4. Do
5. Dt_i=Tokenize(d_i)
6. Sdt_i=stemming(Dt_i)
7. PBD[]=stopwordremoval(Sdt_i)
8. Done
9. Gene_proteins=GP[]=Abner(PBD);// Extract gene_protein tags using Abner library.
10. GP Tokens=GPT[]=Tokenize(GP[],D);// Extract gene_protein tokens in each document D.
11. for each gpt in GPT
12. Compute Gene_Protein probability in each PMID document as
13. $GP\,Prob(D[i], gpt) = Max\{\frac{Prob(gpt/\ D[i])}{Prob(gpt)}\}; i = 01, 2...N$
14. Find the similarity between gene tags to the synonym dataset SD.
15. For each gene_protein gp_m in GPT[]
16. Do

    Synonym gene_protein set =SGP[]=Sim(gp_m,SD)

    $SGP[] = Sim(gp_m, SD)$

    $Sim(gp_m, SD_j) = Max\{\frac{Prob(gp_m / SD_j)}{|SD_j| . Prob(gp_m)}\};, m = 1, 2...|G|$

    $j = 1, 2...|SD|$

    $\text{Document Weight=}DW(Dt_i, gp_m, SD_j) = Sim(gp_m, SD_j) * \sum_{i=1}^{|Dt_i|} Prob(Dt_i / SD_j) . Prob((gp_m \cap Dt_i) / SD_j)$

    $\text{Add } WSGPD[]=\{DW(Dt_i, gp_m, SD_j), Dt_i, gp_m, SD_j\}$
17. Done
18. Mark gene_protein tags and synonym terms as bio keyterms.
19. For each biomedical document from D[]

    Do

    For each wd_i in WSGPD[]

    Do

    If($wd_i > \lambda$)

    Then

    Bio key terms BKT[i]=$\{Tk_i, gp_m, SD_j\}$

    End if

    End for
20. End for

In the algorithm1, each document is filtered by using the tokenization, stemming, stopword removal etc. After tokenization, gene/proteins are extracted using the java library ABNER tagger. These gene/proteins are used to find the probability computation to each document for document entity scoring and graph weight initialization process.

4 Experimental Results

Experimental results are evaluated on large collection of TREC document sets taken from the repository [13]. Different biomedical datasets such as Pubmed and medline xml datasets are used for document clustering process. Here, each dataset is pre-processed to remove the uncertain features or noisy content. After the documents pre-processing phase, each document is analyzed its performance using the accuracy and the runtime metrics.

Sample Data in xml format:

```
          <DescriptorName UI="D011092" MajorTopicYN="N">Polyethylene Glycols</DescriptorName>
        </MeshHeading>
        <MeshHeading>
          <DescriptorName UI="D011189" MajorTopicYN="N">Potassium Chloride</DescriptorName>
        </MeshHeading>
        <MeshHeading>
          <DescriptorName UI="D012995" MajorTopicYN="N">Solubility</DescriptorName>
        </MeshHeading>
        <MeshHeading>
          <DescriptorName UI="D013552" MajorTopicYN="N">Swine</DescriptorName>
        </MeshHeading>
      </MeshHeadingList>
    </MedlineCitation>
    <PubmedData>
      <History>
        <PubMedPubDate PubStatus="pubmed">
          <Year>1975</Year>
          <Month>12</Month>
          <Day>15</Day>
        </PubMedPubDate>
        <PubMedPubDate PubStatus="medline">
          <Year>1975</Year>
          <Month>12</Month>
          <Day>15</Day>
          <Hour>0</Hour>
          <Minute>1</Minute>
        </PubMedPubDate>
        <PubMedPubDate PubStatus="entrez">
          <Year>1975</Year>
          <Month>12</Month>
          <Day>15</Day>
          <Hour>0</Hour>
          <Minute>0</Minute>
        </PubMedPubDate>
      </History>
      <PublicationStatus>ppublish</PublicationStatus>
      <ArticleIdList>
        <ArticleId IdType="pubmed">85</ArticleId>
        <ArticleId IdType="pii">0005-2795(75)90038-0</ArticleId>
      </ArticleIdList>
    </PubmedData>
  </PubmedArticle>
  <PubmedArticle>
```

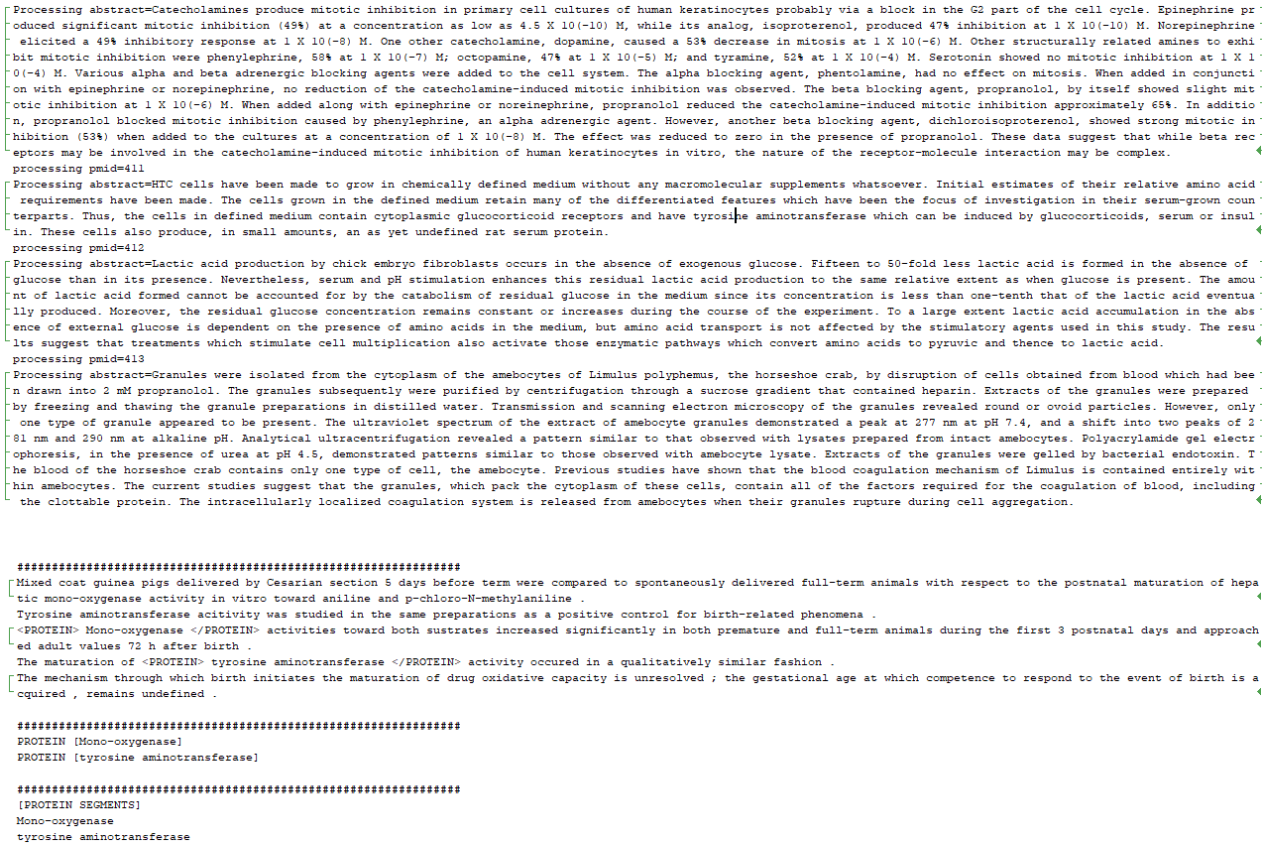

Figure 2: Above screenshot represents the abstract extraction and gene/protein entity extraction.

Table 1 Average MESH term context similarity on different training Documents

Medlinesize	BioNER	SVDD	NMF	NER	ProposedModel
#50	0.786	0.796	0.813	0.906	0.969
#25	0.796	0.802	0.795	0.916	0.961
#75	0.818	0.817	0.807	0.907	0.971
#100	0.796	0.818	0.847	0.925	0.978
#125	0.816	0.885	0.946	0.935	0.987

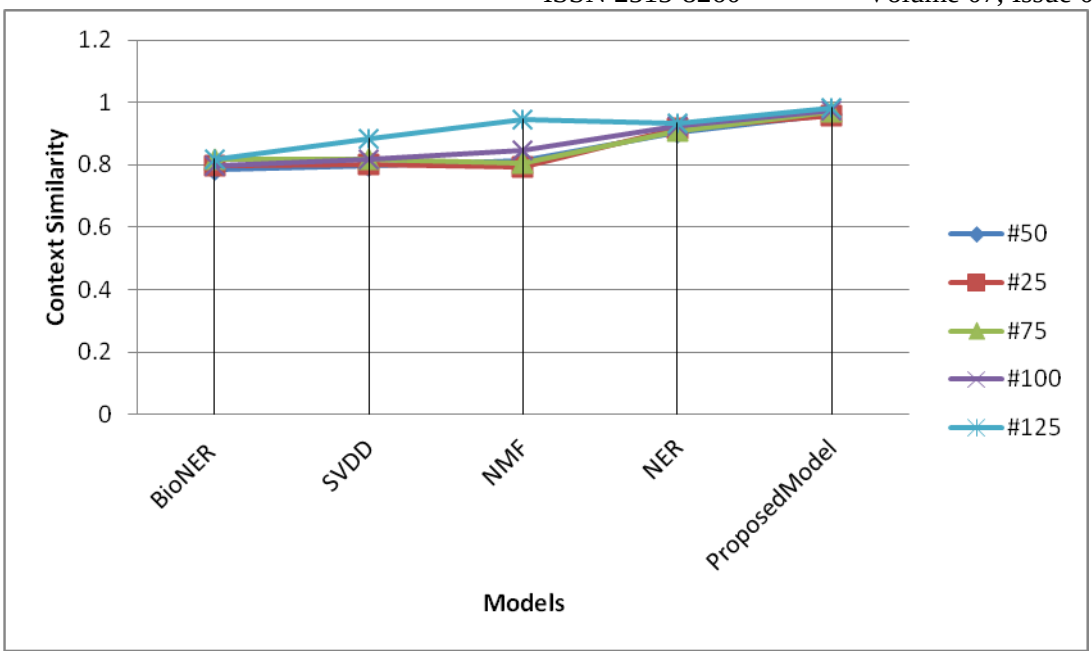

Figure 3: Average Synonym context similarity on different Medline Documents

Table 1 and Fig 3 represent the average gene/protein context similarity accuracy on different Medline document sets. From the figure, it is clear that proposed model has high preprocessing rate compare to traditional model in terms of key-phrase identification and document feature similarity.

Table 2: Runtime (secs) comparison of Medline preprocessing models on Medline Datasets

Medline size	NER	SVDD	NMF	BioNER	Proposed Model
#25	198.3	189.5	218.9	239.9	171.7
#50	267.7	243.2	324.8	318.7	182.3
#75	406.7	389.2	419.1	448.9	289.7
#100	578.2	589.7	572.1	593.6	301.9
#125	723.3	683.5	729.5	679.3	387.3

Table 2 describes the runtime comparison of proposed model to the existing models in terms of milli-secs. From the table , it is clear that proposed model has less time complexity compared to traditional models.

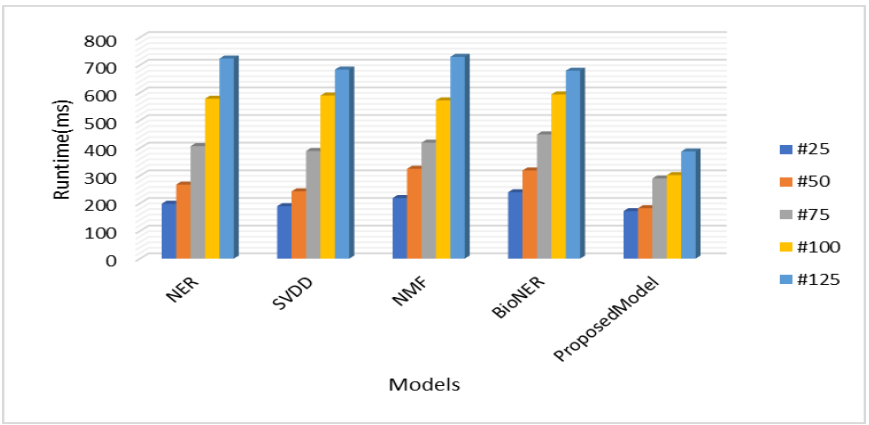

Figure 4:Performance of runtime measure on the bimedical document sets.

Figure 4 describes the runtime comparison of proposed model to the existing models in terms of milli-secs. From the figure, it is clear that proposed model has less time complexity compared to traditional models.

References

[1] C. Olsen, K. Fleming, N. Prendergast, R. Rubi, F. Emmert-Streib, G. Bontempi, B. Haibe-Kains and J. Quackenbush, "Inference and validation of predictive gene networks from biomedical
a. literature and gene expression data", "Genomics 103 (2014) 329–336".

[2] J. M. Ruiz-Martínez, R. Valencia-García, J. T. Fernández-Breis, F. García-Sánchez and R. Martínez-Béjar , "Ontology learning from biomedical natural language documents using UMLS", "Expert Systems with Applications 38 (2011) 12365–12378".

[3] Y. Zhang and Z. Lu, "Exploring Semi-supervised V ariational Autoencoders for Biomedical Relation Extraction".

[4] B. Bhasuran, D. Subramanian and J. Natarajan," Text Mining and Network Analysis to Find Functional Associations of Genes in High Altitude Diseases", "Computational Biology and Chemistry".

[5] J. Chiang, H. Liu, S. Chao and C. Chen, "Discovering gene–gene relations from sequential sentence patterns in biomedical literature", "Expert Systems with Applications 33 (2007) 1036–1041".

[6] M. T. Hassan, A. Karim, J. Kim and M. Jeon, "CDIM: Document Clustering by Discrimination Information Maximization", "Information Sciences 316 (2015) 87–106".

[7] S. J. Fodeh and A. Tiwari , "Exploiting MEDLINE for Gene Molecular Function Prediction via NMF based Multi-Label Classification", "Journal of Biomedical Informatics".

[8] Y. Ji, Y. Tian, F. Shen and J. Tran, "Leveraging MapReduce to efficiently extract associations between biomedical concepts from large text data", "Microprocessors and Microsystems 0 0 0 (2016) 1–9".

[9] H. Kim and S. Chen , Associative Na¨ve Bayes classifier:Automated linking of geneontology to medline Documents, Pattern Recognition42(2009)1777—1785.

[10] X. Ling, J. Jiang, X. He, Q. Mei, C. Zhai and B. Schatz, Generating gene summaries from biomedical literature: A study of semi-structured summarization, Information Processing and Management 43 (2007) 1777–1791.

[11] S. Kim and J. Yoon , Link-topic model for biomedical abbreviation disambiguation, Journal of Biomedical Informatics 53 (2015) 367–380.

[12] K. Liu, W. R. Hogan and R. S. Crowley , Natural Language Processing methods and systems for biomedical ontology learning, Natural Language Processing methods and systems for biomedical ontology learning, Journal of Biomedical Informatics 44 (2011) 163–179

[13] T. Theodosiou, L. Angelis, A. Vakali, G. N. Thomopoulos , Gene functional annotation by statistical analysis of biomedical articles, international journal of medical informatics 7 6 (2 0 0 7) 601–613

[14] Z. Wang, S. Xu and L. Zhu , Semantic Relation Extraction Aware of N-Gram Features from Unstructured Biomedical Text, Journal of Biomedical Informatics.

[15] E. Yan and Y. Zhu, Tracking word semantic change in biomedical literature, International Journal of Medical Informatics 109 (2018) 76–86.

[16] N. Zong, S. Lee, J. Ahn and H. Kim , Supporting inter-topic entity search for biomedical Linked Data based on heterogeneous relationships, Computers in Biology and Medicine 87 (2017) 217–229

[17] Hashwanth A., Sanghavi P.V.N., Satya Vara Prasad B.B.V., Tenali R.K. (2019), 'Priority based retrieval of information with documents domain division approach', International Journal of Recent Technology and Engineering, 8(1), PP.562-566.

[18] Kousar Nikhath A., Subrahmanyam K. (2019), 'Feature selection, optimization and clustering strategies of text documents', International Journal of Electrical and Computer Engineering, 9(2), PP.1313-1320.

[19] Mnssvkr Gupta V., Phani Krishna V. (2019), 'Key node selection network analysis and centrality measurements on a dataset of cancer documents', ARPN Journal of Engineering and Applied Sciences, 14(5), PP.1051-1061.

[20] Haritha Donavalli, Balaji Penubaka,'Identification of Opinionated Features Extraction from Unstructured Textual Reviews', International Journal of Recent Technology and Engineering(IJRTE)ISSN:2277-3878, Volume-7,Issue-6s, PP.674-677, March 2019.

[21] Prakash K.B., Dorai Rangaswamy M.A. (2019), 'Content extraction studies for multilingual unstructured web documents', Advances in Intelligent Systems and Computing, 749(), PP.653-664.

[22] Syamala M., Nalini N.J., Maguluri L., Ragupathy R. (2017),'Comparative analysis of document level text classification algorithms using R',IOP Conference Series: Materials Science and Engineering,225(1).

[23] Jose Mary G., Haritha D. (2017),'A survey on best keyword cover search',Journal of Advanced Research in Dynamical and Control Systems,9(Special issue 14),PP.2217-2231.

[24] Sharma N., Yalla P. (2019), 'A conceptual dependency graph based keyword extraction model for source code to API documentation mapping', International Journal of Recent Technology and Engineering, 8(2), PP.5888-5895.

[25] Rizwana S., Challa K., Rafi S., Imambi S.S. (2019), 'Enhanced biomedical data modeling using unsupervised probabilistic machine learning technique', International Journal of Recent Technology and Engineering, 7(6), PP.579-582.

[26] Vasantham V,Haritha D, 'A Survey on cost minimization techniques for big data processing', Journal of Advanced Research in Dynamical and Control Systems(2018).

[27] Joseph Sastry K.S.S., Ketaraju V.D.S. (2018),'A document ranking approach based on weighted-gene/protein in large biomedical documents using mapreduce framework',International Journal of Simulation: Systems, Science and Technology,19 (6),PP. 261