CHAPTER 3
PROPOSED SYSTEM

3.1 OVERVIEW

The development of biomedical domain increases the research finding in that field. As the result, there is a significant increase in the biomedical research articles database. The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information. NCBI is a known as National Center for Biotechnology Information which is a part of United States National Library of Medicine (NLM) and it is a division of National Institutes of Health, United States of America[34]. NCBI maintains the series of databases related to biomedicine and biotechnology[26]. NCBI is an important repository for bioinformatics services and tools. The NCBI includes various major databases like GenBank, PubMed where PubMed is a database that maintains biomedical literatures.

3.2 TEXT MINING

The Text Mining plays a vital role in extracting biomedical knowledge from the biomedical data sets and is known as Bio text mining. The text mining process includes various consecutive stages towards the extraction of knowledge, and it includes tasks like:

- Information Retrieval
• Information Extraction

The Information Retrieval involves in constructing the corpus by retrieving the subset of documents from the available databases based on the requirements specified by the user. Whereas Information Extraction involves extracting the knowledge from the documents retrieved by the Information Retrieval system.

Various studies exist in the process of extracting information from the biomedical documents[33]. Each one got its own attainment and limitations. The proposed work “Information Extraction from unstructured documents- A Text mining Approach” tries to overcome the limitations of the existing systems in the same tone.

3.3 PROPOSED SYSTEM

The objective of the proposed work “Information Extraction from unstructured documents- A Text mining Approach” is

- To acquire Biomedical document form Biomedical Databases.
- To extract the biomedical Named Entities automatically from the Biomedical Documents which are available in Unstructured format.
- To identify biomedical Relations automatically from the Biomedical Documents.
The generic pipeline of the proposed system consists of the subtasks (i) Information Retrieval and (ii) Information Extraction. Fig 3.3 illustrates the generic pipeline of the proposed system. The generic pipeline consists of various subsystems like user requirements, Query Generation system, Corpus Generation, Named Entity Recognition system, Relation Extraction and Biomedical Databases[27]. The subtasks categorized into Information Retrieval tasks and Information Extraction tasks.

**Fig: 3.3 Generic pipeline of the proposed system**

The Information retrieval is the process of retrieving documents from the huge collection of database. It will create a subset of documents from the database. Fig 3.3.1(a) exhibits the process of Information Retrieval.
The Information Extraction is the process of extracting information’s or knowledge from the subset of documents. Which is retrieved from the processes of Information Retrieval. Fig 3.3.1(b) exhibits the process of Information Extraction.

![Fig 3.3.1(a) Information Retrieval Fig 3.3.1(b) Information Extraction](image)

### 3.4 ARCHITECTURE OF PROPOSED SYSTEM

The architecture of proposed system consists of a sequence of many tasks to achieve the desired vision[6]. The initial process of the proposed system begins with identifying the user requirements and ends with the solution to the user's requirements. The Major tasks of the proposed system is classified into

- Information Retrieval Process
- Information Extraction Process

The Information Retrieval and Information Extraction Process can be subdivided in to

- Document Acquisition
- Document Structuring
- Validation of Relevance
• Named Entity Recognition

• Relation Extraction

The Document Acquisition, Document Structuring and Validation of Relevance do belong to Information Retrieval process[7]. Named Entity Recognition and Relation Extraction are the parts of Information Extraction Process. All these process are subsequence processes.
3.4.1 INFORMATION RETRIEVAL

The Information Retrieval process is to obtain the user requirements and to generate corpus that fulfills the user requirements. The Information Retrieval process gathers the user requirements using a GUI Design and based on the user requirements, the query will be
generated. The query will be given to the e-Utile a search engine for NCBI, retrieves the biomedical documents that satisfies the user requirements. The corpus of retrieved documents will then be generated.

The documents will be in unstructured format and cannot be understood by the Natural Language Processing tools. So the PDFBox is used to convert the document into a form that can be understood by the Natural Language Processing tools. The relevance of the retrieved documents is validated using the Document – Weight Matrix.

### 3.4.2 INFORMATION EXTRACTION

The Information Extraction Process begins after validation of retrieved corpus. Information Extraction process is a combination of Named Entity Recognition process and Relation Extraction Process.

#### 3.4.2.1 NAMED ENTITY EXTRACTION

The Named Entity Extraction Process begins with Preprocessing of documents, to understand the structure of the documents and following that the document annotated based on the Syntactic structure. This involve with the process like Parts-Of-Speech Tagging and Noun phrase chunking.

The next phase in the Named Entities Recognition is Semantic Annotation, which is done in two phases namely Lexical based Named Entity Recognition (L-NER) and Machine Learning based Named Entity
Recognition (M-NER). The L-NER identifies the entities with the help of lexical resources like Dictionaries, Ontology's, Lookup Tables, Rule sets and Lexical Words.

M-NER uses a hybrid machine learning algorithm to extract the entities. Feature Space based Conditional Random Field is used to identify the Named entities. This over comes the drawbacks of other machine learning algorithms[72]. A set of 60 features are used to identify the Named Entities from the biomedical documents. The system is compared with Support Vector Machine learning algorithm.

### 3.4.2.2 RELATION EXTRACTION

The Named Entity Recognition is the prerequisite for the relation extraction. The Relation Extraction phase extracts the relations from the documents. This phase exploits a hybrid algorithm to identify the relations[9][14]. The mechanism is a twofold Initially the co-occurrence approach is used to identify the relations. Next, Conditional Random Field machine learning algorithm is used to extract the relations[3]. The outcome of the first part is used as the training data for the machine learning algorithm[17]. The system is compared with the Rule based algorithm.

The proposed system outperforms the existing systems with its limitations.