Paper Title* (To-be-decided)

*Note: Sub-titles are not captured in Xplore and should not be used

Poonam Kankariya
Computer Science & Engineering
University of Missouri – Kansas City
Kansas City MO USA
pkkdg@mail.umkc.edu

Rohita Goparaju
Computer Science & Engineering
University of Missouri – Kansas City
Kansas City MO USA
rgmh5@mail.umkc.edu

Madhuri Sarda
Computer Science & Engineering
University of Missouri – Kansas City
Kansas City MO USA
ms6dt@mail.umkc.edu

***Abstract*—Ontology is an effective and efficient way to understand a concept and any and everything related to it. It is a great tool in the world of deep learning and machine learning, where a system is trained to understand, interpret, analyze and derive results. The machine learning models built using Natural Language Processing (NLP) techniques resulting in knowledge graphs have evolved over the years making it applicable on text data pertaining to any field possible, containing information in various text formats. The world of medicine is no different and the NLP techniques have facilitated the understanding of the existing research and its evolution over the years. With focus on breast cancer, we shall develop a machine-learning model using NLP techniques that will identify the various aspects of breast cancer such as causes, symptoms, diagnosis, treatment and its side effects.**

Keywords—ontology, machine learning, deep learning, Natural Language Processing (NLP), breast cancer

I. Introduction

With an increase in the number of diseases emerging and their impact on human well-being, there has been an increasing demand for faster diagnosis to ensure timely treatments. There is an enormous amount of bio-medical data available today. The world of big data, machine learning and deep learning is a great source to analyze this data and obtain results.

Availability of research work in the field of bio-medical analyzing and understanding the causes, symptoms, diagnosis, cure and effects of various diseases has eased the process of finding cures for newly evolving diseases in a better, faster way. All these information present in various forms of text data may be analyzed to identify various ontologies helping one understand, analyze and fight these diseases. The natural language processing (NLP) techniques available in the world of deep learning and machine learning process to be an efficient and effective means of analyzing the information (data) available to identify solutions.

II. Related Work

There exists a tremendous amount of research in the area of ontology creation in the bio-medical domain. A prototype system has been proposed using the approach of semantic technology leveraging the staging manual's automatic parsing of data. Further, additional biomarkers were included when developing the cancer staging manual. Development of the ontology involved the use of generic terms used within the community to map the terms associated with breast cancer.

III. Proposed Work

We propose to build a machine learning model using NLP techniques for the identification of causes, symptoms, diagnosis, treatment and/o side-effects related to breast cancer. The model obtained shall be used to build ontologies relating to various aspects of the disease.

The process involves several steps as described below.

A. Data Collection/Extraction

The data associated with the research of this paper belongs to the category of 'Cancer' with a focus on breast cancer. The data set used comprises of abstracts of several research papers available related to the category selected. The data has primarily been sourced from IEEE research papers along with a few other publications presenting promising content.

The process of data collection involved extensive search of the research material available associated to the topic under consideration and accessing each appropriately. The process of data extraction involves manual extraction of the 'abstract' section of the research material identified as our input data presented in individual text format.

B. Data Processing

The text data is processed using the techniques associated with Natural Language Processing (NLP). The initial processing of the data using NLP techniques shall identify the prominent, relevant information available within the data set resulting in elimination of noise and other irrelevant content with no value-addition.

The process begins with the reading of the input (multiple text files of research paper abstracts) into the system as a single input. This is achieved through the append process bringing together multiple data of the same format.

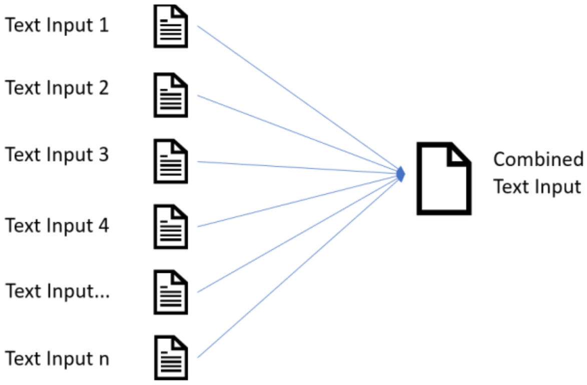

Fig 1. Multiple text-based input files read into a single input variable

Once the data is read, the process of sentence and word tokenization is performed. This enables the identification or extraction of individual sentences (sentence tokenization) and individual words/terms (word tokenization) from the data.

The process of stemming and lemmatization follows leading to the derivation of the root (source) of the tokens identified.

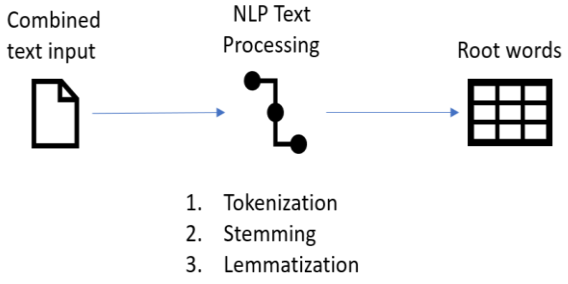

Fig 2. NLP processing of input data

C. *Information Extraction & Retrieval*

The use of the originating terms derived through the process of lemmatization helps in multiple text analysis.

Part-of-speech (POS) tagging on the terms identified enables the identification of the part of speech associated with each term identified.

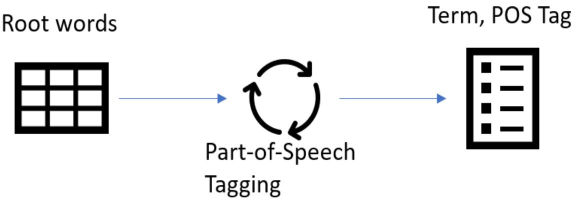

Fig 3. Part-of-speech tags of root words identified

Extraction of triplets using the terms identified enables the identification of the *predicate (subject*, *object)* relationship between the terms.

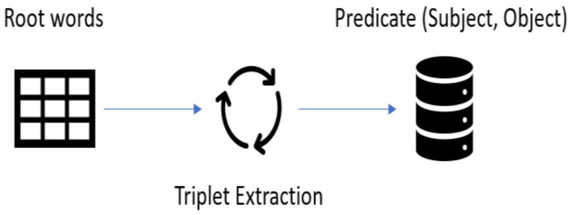

Fig. 4 Triplets extraction using refined terms

D. *Ontolgy Creation*

Add context here describing what information will this section provide.

IV. Implementation and Evaluation

Add context here describing what information will this section provide.

Acknowledgment

Add context here describing what information will this section provide.

References

[1] Seneviratne O. et al. (2018) Knowledge Integration for Disease Characterization: A Breast Cancer Example. In: Vrandečić D. et al. (eds) The Semantic Web – ISWC 2018. ISWC 2018. Lecture Notes in Computer Science, vol 11137. Springer, Cham.

[2] Add references

[3] Add references

[4] Add references

[5]