Ontology Designing Towards Effective Disease Analysis

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), topic modeling, 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.

Information about a disease and every aspect associated with it is available in abundance. However, it is critical to distinguish between the correct and incorrect information. This is particularly a major issue when dealing with bio-medical data since is may impact the well-being of an individual or a community in general.

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. Text processing using the NLP techniques help with various aspects of life as identified in Fig 1. The benefits range from simple translation to deeper analysis of the sentiments associated with the data along with various other applications based on the type, source and use of the data involved.

Fig 1. Applications of Natural Language Processing [14]

We have designed an ontology using several concepts of text data analysis using the NLP techniques that enable in easy understanding and analysis of the information made available. This aim of the ontology developed is to provide optimal knowledge of the subject matter leading to a more efficient management of the disease.

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.

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.

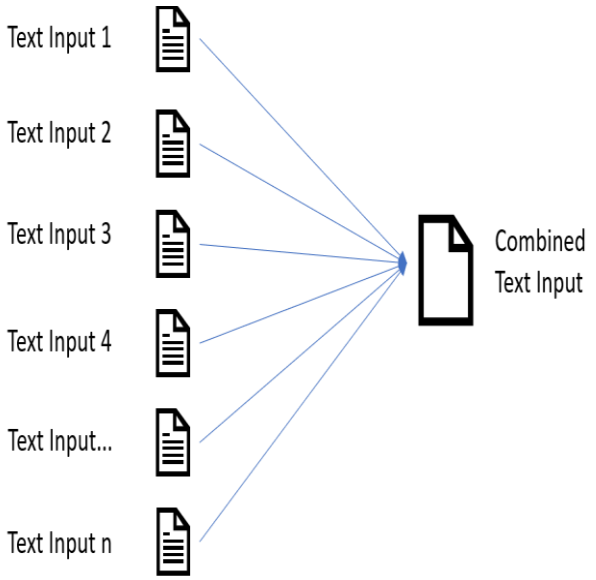

Fig 2. Multiple text-based input files read into a single input variable

Once the data is read, the process of sentence and word tokenization is performed. T*okenization* process involves the splitting of data into independent components, either sentences or words. This process eliminates characters that do not add value to the data such as punctuation marks.

The process of *sentence tokenization* yields a list of individual sentences identified from the text data. The process of *word tokenization* yields individual terms or words obtained from the text data. The process of word tokenization is applicable on the data derived using sentence tokenization or directly on the original data based on user requirements.

The tokenization process enables the identification or extraction of individual sentences (sentence tokenization) and individual words/terms (word tokenization) from the data. The tokens derived may comprise of important keywords, phrases etc. that may help the user with identification of the concept involved.

The process of stemming and lemmatization follows leading to the derivation of the root (source) of the tokens identified. *Stemming* is the process of reducing a given term to its root form i.e. pseudo stem. The terms derived as a result of stemming may not always hold a proper meaning. *Lemmatization* is the process of reducing a given term to its valid linguistic form referred to as lemma. These derived terms provide a valid meaning of its own. Application of stemming and/or lemmatization helps in understanding the origin of the word tokens identified which form an essential part in topic modeling.

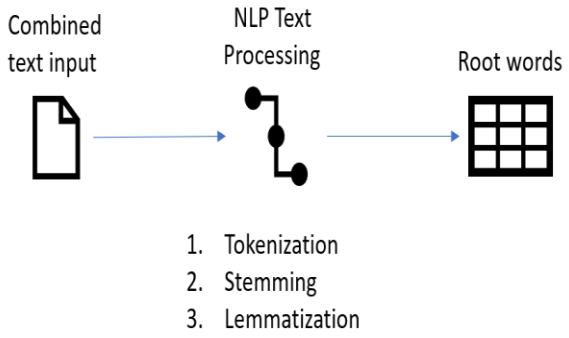

Fig 3. NLP processing of input data

C. Information Extraction & Retrieval

The use of the originating terms derived through the process of stemming and/or 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.

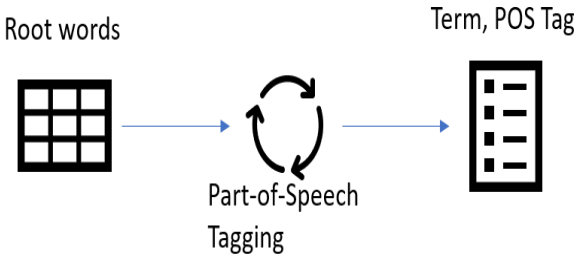

Fig 4. Part-of-speech tags of root words identified

Triplet extraction involves identification of the relationship existing between the words (tokens) identified in the form *predicate (subject, object)*. This process helps in understanding the context of the data along with the attributes that relate to one other. It also enables in identifying and understanding how one word (token) may related to multiple different words (tokens) in similar or unique manner.

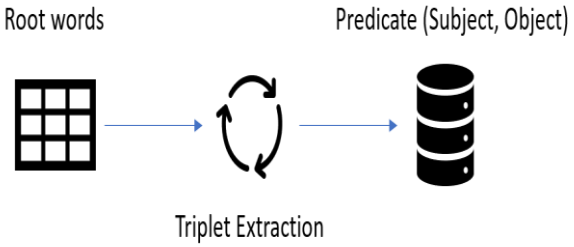

Fig 5. Triplets extraction using refined terms

D. Topic Modeling

Topic modeling is a statistical unsupervised machine learning approach used in text analysis. It helps with identification of the different topics from the text data by identifying patterns existing within. It helps in understanding the context and the concept associated with the data.

Using the data set, topic modeling enables in the identification of various concepts focused on the data. Considering cancer research related data, the topic modeling helps in identification of various aspects such as causes, symptoms, diagnosis, treatment, medication and so forth.

E. *Ontolgy Creation*

Ontology, a philosophical study of nature of being is a branch of metaphysics. It is a concept of understanding what exists and what may exist. In the context of text analysis, it helps us understanding the existence of concepts (topics) and how they relate to one another.

IV. IMPLEMENTATION AND EVALUATION

The implementation of the project involved a combination of the various concepts identified in section III. The primary focus involves the identification of content associated with causes, symptoms and treatments, along with diagnosis and other identifiable concepts.

To initiate the text analysis process, it is essential to identify the various machine learning and deep learning modules and/or libraries enabling the process involved as identified in Fig 6.

```
# Import libraries
from __future__ import print_function
import spacy
import textacy
import os
import re
import nltk
import gensim
import pyLDAvis.gensim
from gensim import corpora, models, similarities
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.stem import WordNetLemmatizer
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('wordnet')
```

Fig 6. Deep Learning/Machine Learning Libraries

A. *Read Text Data Input*

The data has been collected from several research articles and medical journals available online and the abstract of each is made available in text format. The input is read and brought together as a list, which is further combined to be read as a single string.

```
# Read input available as a collection of abstracts from research material and articles into a lis
new_list = []
for root, dirs, files in os.walk("/Users/GowthamYesG/Desktop/KDM/Project/Abstracts"):
    print(files)

# Grouping the input data into multiple sections (3) for processing efficiency
def chunks(lst, n):
    for i in range(0, len(lst), n):
        yield lst[i:i+n]

chunked_list = list(chunks(files, 7))

# Processing the input reading abstracts from Abstracts folder( 7 at a time)
for file_list in chunked_list:
    for file in file_list:
        if file.endswith('.txt'):
            with open(os.path.join(root, file), 'r') as f:
                text = f.read()
                new_list.append(text) # Stored all the abstracts data in new_list

    print("File list: ",file_list)

    # Data cleansing and pre-processing by converting list of data into a single string
    clean_data = ''.join(str(e) for e in new_list)
```

Fig 7. Input Reading Algorithm

B. *Text Processing*

The data once read need to be cleansed for its optimal utilization. It involved the application of various techniques available within NLP associated following the concepts of lexical, syntactic and semantic analysis. Application of a combination of these techniques results in the output in the required format relating to the business use case.

The pre-processing begins with separating the data into individual sentences and words (tokens) using the process of tokenization following the concept of lexical analysis. The process helps to identify the individual sentences and further the individual words present within the data.

```
# Sentence Tokenization
clean_data_sent = sent_tokenize(clean_data)
print("Sentence Tokenizer results:")
print(clean_data_sent)

# Word Tokenization
clean_data_word = word_tokenize(clean_data)
print("Word tokenizer results:")
print(clean_data_word)
```

Fig 8. Tokenization

The clean tokens (words) identified are further processed using the concepts related to syntactic and semantic analysis. These concepts help in the derivation of the root of the word being processed adding meaning toward the analysis. The techniques include stemming and lemmatization following similar concepts. The primary difference between stemming and lemmatization is the root word derived is more meaningful in lemmatization when compared to stemming.

```
# Stemming of tokens identified
stemmer = PorterStemmer()
stemming = [stemmer.stem(str(e)) for e in
            clean_data_word]  # word in clean_data_word is a char array and converting that to string
stemming = [word for word in stemming if len(word) > 1]  # removing single letter words
print("Stemming words:")
print(stemming)

# Lemmatization of tokens identified
lemmatizer = WordNetLemmatizer()
lemmatizing = [lemmatizer.lemmatize(str(e)) for e in clean_data_word]
print("Lemmatization output is : ");
print(lemmatizing);
```

Fig 9. Stemming – Lemmatization Algorithm

C. *Information Extraction & Retrival*

The clean, pre-processed data is used as the input to extract meaningful information useful for the analysis and understanding of the content of the data. This is achieved using the technique of triplet extraction which reveals the various combination of words along with their relationship.

```
# Triplets Extraction
nlp = spacy.load("en_core_web_sm")
tuples_list = []
for sentence in clean_data_sent:  # used sentence tokens
    val = nlp(sentence)
    tuples = textacy.extract.subject_verb_object_triples(val)
    if tuples:
        tuples_to_list = list(tuples)
        tuples_list.append(tuples_to_list)
```

Fig 10. Triplets Extraction Algorithm

The concept of Latent Dirichlet Allocation (LDA) is used to identify the topics present within the data using the triplets identified. This process helps in identifying the key concepts present within the data which may be associated to several different tokens or concepts through varying logic.

```
# LDA processing to identify topics
ldamodel = gensim.models.ldamodel.LdaModel(corpus, num_topics=6, id2word=dictionary, passes=15)

# Save LDA topic modeling for repurpose
ldamodel.save('model_combined.gensim')
topics = ldamodel.print_topics(num_words=4)
print('\n')
print("Now printing the topics and their composition")
print("This output shows the Topic-Words matrix for the 6 topics created and the 4 words within each topic")
for topic in topics:
    print(topic)
```

Fig 11. Latent Dirichlet Allocation (LDA) Algorithm

```
# Topic modeling visualization
lda_viz = gensim.models.ldamodel.LdaModel.load('model_combined.gensim')
lda_display = pyLDAvis.gensim.prepare(lda_viz, corpus, dictionary, sort_topics=True
pyLDAvis.show(lda_display)
```

Fig 12. Topic Modeling Visualization Algorithm

D. Statistical Analysis

The topic obtained from the LDA process are used for the analysis of the content helping with the identification of the concepts related to the same. This is achieved using statistical method involving the analysis of similarity score.

```
# Similarity score analysis from abstracts with topics identified
get_document_topics = ldamodel.get_document_topics(corpus[0])
print('\n')
print("The similarity of this abstracts with the topics and respective similarity score are ")
print(get_document_topics)
```

Fig 13. Similarity Analysis Algorithm

The topics derived from the data are significantly influenced by the *num_topics* and *num_words* parameters used for the analysis. Variation of these parameters reveal varying topics being identified prominently. The topics identified are identified using genism modeling techniques. The models derived help identify the top topics associated with the data. The multidimensional scaling of the data showing the correlation among the topics and the detail of the terms or classes associated with the top topics provides an added advantage.

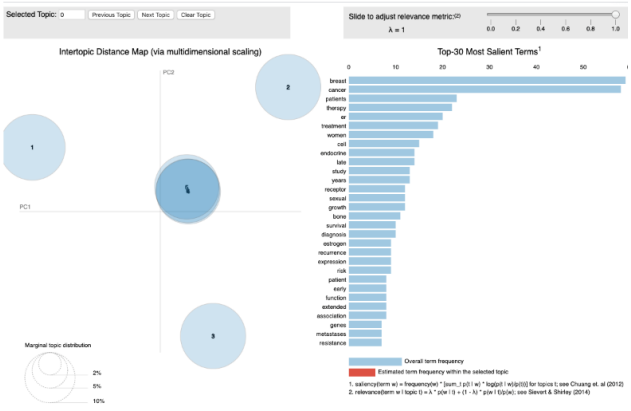

Fig 14. Topic Model Visualization

E. Ontology

Ontology, the concept of the nature of being is the key to understanding and analyzing various components of any topic that gives a broader understanding of the concepts and its related features. The topics obtained using LDA defines the classes and the parameters associated with the classes defined. The tool titled Protégé forms the basis for the creation of the classes, their relations and functions using the results obtained from topic modeling.

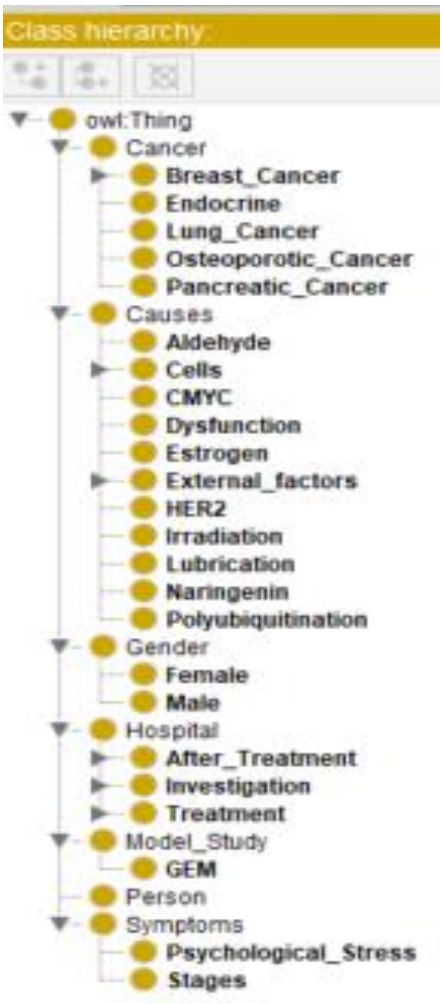

Fig 15. Classes

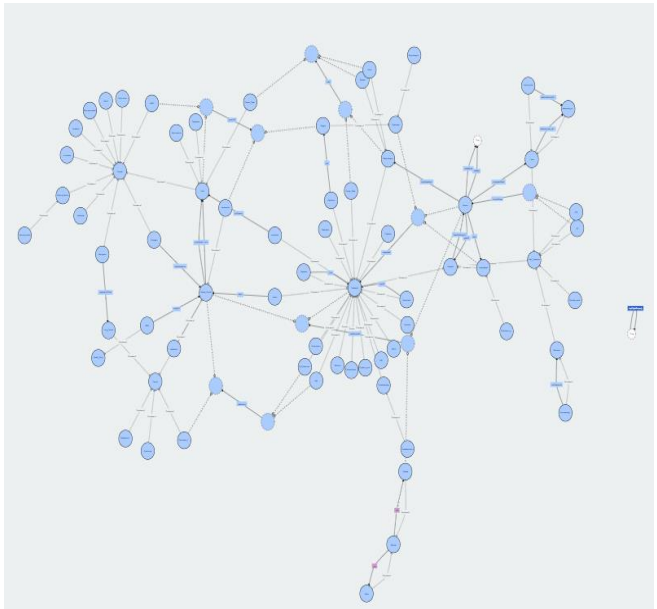

Fig 16. Cancer Ontology Degree 0

CONCLUSION

The main objective of the research involved the development of an effective ontology helping in understanding the various concepts associated with cancer. It is aimed at getting better insights to understand the disease and its related components. The process helps in retaining relevant and meaningful information using the NLP techniques resulting in a robust analysis.

The research also revealed the availability of tremendous amount of information related to the topic of cancer. As there is increasing information available, the need to process deriving only the true and relevant information to process gets critical with the advancements in the technology. Thus, the scope for improvement of the process is open and flexible.

ACKNOWLEDGMENT

We thank Syed Jawad Shah and Sai Sree Narne for the guidance and support towards helping us improve our work.

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] https://pubs.rsna.org/doi/full/10.1148/radiol.14132832

[3] https://www.researchgate.net/publication

[4] https://ieeexplore.ieee.org

[5] https://arxiv.org/pdf/1607.08074.pdf

[6] https://arxiv.org/pdf/1807.07991.pdf

[7] https://ieeexplore-ieee-org.proxy.library.umkc.edu/document/4562013/

[8] https://ieeexplore.ieee.org/document/6834613/

[9] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5174013/

[10] http://dline.info/fpaper/jdim/v15i5/jdimv15i5_3.pdf

[11] https://webprotege.stanford.edu/

[12] http://www.visualdataweb.de/webvowl/

[13] https://en.wikipedia.org/wiki

[14] https://www.kdnuggets.com/2018/10/main-approaches-natural-language-processing-tasks.html