Enriching medical text classification by Using Machine Learning

Mayuri B. Satpute
Computer Engineering,
Smt. Kashibai Navale College of Engineering, Pune, India,

Dr. V. V. Kimbahune
Computer Engineering,
Smt. Kashibai Navale College of Engineering, Pune, India.

***Abstract* – Health care sector has been noticing some of the most technological advances in the recent years. This has led to the significant improvements in the various aspects of health care such as diagnosis, treatment and also prevention. Technology has been one of the biggest contributors to this rise. This has allowed the patients to achieve effective diagnosis and treatment from the comfort of their homes through the internet platform. There are several websites and web portals for online diagnosis through the internet platform. But most of these approaches have been plagued with inaccuracies as the medical text classification is a highly complicated task. To provide remedy for these issues, the proposed methodology has been detailed in this research article. This article outlines an effective medical text classification technique through the use of NLP techniques such as TF-IDF, noun identification, Bag of Words, etc. The methodology utilizes Logistic Regression and Artificial Neural Networks along with Decision Tree to achieve effective classification of the Medical Text. Extensive Experimentation has been performed to facilitate the extraction of the performance metrics of the system. The outcomes of the experimentation concluded that the proposed methodology significantly outperforms the conventional approaches.**

Keywords— *Natural Language Processing, TF-IDF, Logistic Regression, Artificial Neural network, Decision Tree.*

I INTRODUCTION

Health is one of the biggest concerns that has been significant in a long and fruitful life. There have been large-scale advances that have been implemented in the medical sector to provide improvements to the diagnosis as well as the treatment of different diseases and ailments. This is ongoing research that has been facilitating the decrease in the incidence of various diseases as well as improving the preexisting infrastructure for diagnosis. This is essential for the effective functioning of the whole process of medication which has been significant in improving the average lifespan of a human being considerably over the past few years.

The improvement in the lifespan has led to you a lot more useful researches which have been contributing positively towards technological advancements. Technological advancements have also helped the medical fraternity in achieving enhancement in their various processes. There are types of equipment and other machines that have been fruitful in providing effective diagnosis for various diseases is as well as their treatment effectively. One of the most important processes of treatment is diagnosis. Medical professionals or doctors educate themselves in this field for an extensive number of years which makes them recognize and memorize various medical terms and terminologies.

These terminologies are developed over years of practice and dedication by the medical professionals in their profession. These are not common words but instead are highly specialized terminologies used to refer to various diagnosis and reactions which are used specifically by medical staff. These words are highly difficult to process and quite complicated for an average human being. Therefore the classification of medical documents and reports has been done manually by these medical professionals. Manually performing this task is a time-consuming process that can be detrimental to the medical professional’s routine.

Therefore there is a need to automate this process of classification which can provide a much-needed enhancement in the medical professional’s life. This can reduce working hours which can be utilized for more fruitful purposes other than Medical text classification. The process of automatically categorizing text is a highly complicated process for a computer to understand. The recognition of text and its various modalities comes very naturally to humans as there are social human beings who like to communicate effectively. But the computer treats them as any other combination of ones and zeros that makes it highly difficult to extract the classification of the medical text effectively.

For the purpose of enabling effective classification of text, the paradigm of Natural Language Processing or NLP is widely utilized. The natural language processing approach is designed for automatic text classification through a computer with little or no interference from a human being. There are various processes inside the umbrella of natural language processing that allow for such a complex task to be executed efficiently. The natural language processing approach is one of the leading sources that can effectively be utilized for the classification of Medical text.

But as the medical text is a highly specialized form of communication and terminology it can be difficult to implement the natural language processing approaches on this type of text. Therefore the proposed methodology also implements a specialized biomedical text bag of words for assistant the natural language processing approach in its execution. This bag of words contains an extensive library of medical terminology that can be correlated with the medical test to provide any relationships between these specialized terms and effectively classify the medical text. Therefore the proposed methodology effectively modifies the natural language processing paradigm through the bag of words approach to be applied specifically for medical text processing.

The proposed methodology implements this improved NLP approach that is specialized for application on Medical text along with the implementation of a feature extraction approach to effectively extract the relevant features from the medical text. This supplement the NLP approach to effectively and completely extract the medical terminologies present in the medical text being given as input. This is further improved through the addition of artificial neural networks that perform neuron creation for the hidden layer estimation to effectively realize the classification of the relevant medical terms. The ANN approach is supplemented through Logistic regression which provides effective regression on the dependent and independent values. Finally, the medical terms are accurately classified through the Decision Tree classification module. This module performs the classification based on the naive Decision tree which provides highly precise classifications of the medical text.

Section 2 of this publication details the evaluation of the previous researches. The section 3 outlines the proposed system. Section 4 explains the experimental results and section 5 furnishes the conclusion to the paper and defines the future research approaches.

II LITERATURE SURVEY

E. Tutubalina explains that the concept of mining scientific libraries and its text has been a very useful tool for diagnosis virus diseases and ailments [1]. For the purpose of mining the text, the concept of Natural Language Processing has been utilized extensively. Therefore to perform this task the authors in this approach utilize bidirectional long short term memory for the purpose of medical concept normalization on the text that is posted on social media using recurrent neural networks. Experimental outcomes suggest that the proposed methodology provides a significant improvement over conventional approaches. The major limitation noticed in this approach is that authors have not utilized the integration of linguistic knowledge into their model.

N. Pattisapu states that the process of extracting meaningful knowledge from the text is a highly useful concept that can be very useful in various different contexts. Text is one of the largest formats in terms of the information and knowledge contained in them. Therefore it's absolutely necessary for the realization of a system that can effectively extract this knowledge. To this end, the researchers in this approach propose an effective model for the purpose of medical concept normalization through target knowledge encoding [2]. Delimitation observed in the proposed methodology is that the authors have not utilized knowledge base embeddings for the purpose of encoding target knowledge.

C. Carbery introduces the concept of learning models as they have been getting highly popular and mainstream as they are being applied in various different fields. The learning models are highly lucrative as they employ the use of deep learning approaches that can effectively in hands any system significantly [3]. Therefore the researches in this approach have utilized the deep learning approaches for the purpose of achieving a future Computer-Based system. The researchers have utilized a deep dynamic Bayesian network for the implementation of this approach.

S. Gupta elaborates on the process of relation classification that is performed on textual material. The analysis of the text has been significantly useful for the purpose of extracting useful knowledge from a corpus of text [4]. This is highly useful in the medical context as it can provide significant improvements in the diagnostic capabilities of the medical institution. The research article outlines an effective utilization of machine learning approaches for the purpose of effectively classifying the relation contained in a structured medical text. The major limitation noticed in this approach is that the authors have not improved the performance of the CNN algorithm using the feature set in the methodology.

A. Zalewski discusses the large-scale advances that have been implemented in the medical field which have been highly useful in achieving a healthy lifestyle for everyone [5]. These technological advances have been supplementing the medical field to improve their systems and provide much better care. But most of these approaches and equipment generate a large amount of data that needs to be effectively e processed to gain meaningful conclusions. To achieve this the authors utilized latent structure which is inferred through the utilization of clinical text and time series. The authors implemented a hierarchical Dirichlet process and a Bayesian nonparametric framework for the purpose of inferring the patient's health status.

V. Plachouras explains the concept of drug testing and other effective utilization of Pharmaceutical drugs that require the monitoring of various side effects that are exhibited to its users. There are various organizations that are set up to effectively analyze the various side effects and adverse drug reactions to effectively provide feedback to the pharmaceutical companies which can be highly useful for their research and development purposes [6]. But due to the inefficiency of most of these approaches, the authors proposed an innovative Framework for quantifying adverse drug reactions through self-reporting on the Twitter social media platform. The authors utilized SVM based classifiers which achieved significant improvements in the experimental results.

B. Parlak states that Medical text is a highly Complex and professional approach that is being designed to be used by highly qualified medical professionals in effectively diagnosing and classify various effects and diseases [7]. But this is a highly time-consuming process that can take up the useful time of the doctors and other medical professionals who are already overworked. Therefore provide solutions to this lack of Medical text classification the authors have proposed and effective utilization of medical documents for the purpose of disease classification.

K. Denecke introduces the concept of natural language processing that is being utilized for effective and in-depth classification and processing of textual data. This is a highly useful paradigm that can be utilized for enabling automatic classification and text extraction approaches. These techniques are utilized by the researchers in this approach for the purpose of extracting medical concepts and other terminology from Medical social media websites through the utilization of clinical NLP tools [8]. The drawback observed in this approach is that the authors have not utilized relation extraction in their methodology.

B. Parlak discusses the concept of classification of medical documents that are being highly researched in recent years. Most of the text classification and processing research has been focused on Medical text which has been highly critical in its approach [9]. There have been numerous approaches to effectively classify and extract Useful information from Medical text for which the authors have researched in realizing the impact feature selection on these approaches. The researches implemented pattern classified structure is C 4.5 decision tree and Bayesian networks. The drawback of this approach is that the authors have only considered a singular dataset for their approach.

Z. Pella explains the concept of text processing that is gaining significant traction and getting highly popular in recent years. Text processing is a highly complex and complicated approach that it is tossed with the extraction of semantic knowledge from a set of text given as input. But due to the fact that there has been an ever-increasing size of text and textual data, it becomes even more difficult to follow this approach [10]. To provide a remedy for this approach the authors have designed an innovative approach for the processing of text in cardiovascular medical records. The limitation noticed in this approach is that the researches have only utilized cardiovascular disease text for their methodology.

H. Yoon states that the implementation of an effective system for the purpose of extracting information from text a highly Complex task which is very difficult to perform by computer. There has been an increase in the research that is being performed for the purpose of extracting useful information from a large amount of text [11]. The researches in this article have proposed that innovative model for the extraction of information from cancer pathology reports through the utilization of convolutional neural networks and hyperparameter Optimization on a high-performance computing environment. The major drawback of this approach is that there is an increased computational complexity that is observed in this approach.

N. Pattisapu introduces the concept of automatic text summarization and recognition by a computer which is a highly difficult task that can be easily performed manually by humans. Due to the large amount of data that is being generated by various social media websites online, this becomes almost impossible for segregating the text manually [12]. Therefore the authors in this approach proposed and innovative concept for the application of automatic Medical concept normalization in social media posts through the implementation of learning semantic representation. The experimental outcomes reveal a satisfactory performance in comparison with the conventional approaches.

H. Al-Mubaid discusses the concept of classification of business documents that are being largely researched for the application of text processing approaches. Automatic text processing is a complicated approach that can be done easily manually but can be highly difficult for a computer to perform. To provide a solution to this problem the authors have proposed the utilization of Bayesian classification based on the naive Bayes theory for the effective and complete classification of disease documents [13]. Improved Bayesian classification performs exceptionally well on the classification task for medical text. The problem with this approach is that the authors have not normalized the weight based on the probability of each attribute that is considered.

S. Han explains the concept of classification of text that is being increasingly gathering the interest of a lot of researches for various purposes especially medical-based implementations. Due to the rise in various social media websites and their frequent and continuous use by a large section of users it has been a very useful source of gathering experiences of users related to medication intake for adverse drug reactions [14]. For utilizing this unique concept the authors have proposed the implementation of a traditional linear model along with hierarchical long short term memory and convolutional neural network for the purpose of automatic classification of medication intake posts on the Twitter social media platform.

III PROPOSED METHODOLOGY

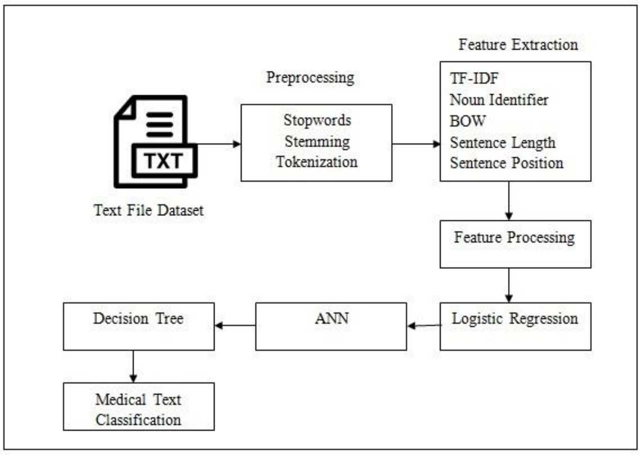

Figure 1: System overview Diagram.

The proposed methodology for the medical text classification in biomedical texts has been elaborated in figure 1 above. There are a collection of steps in the procedure that are performed to attain the goals in the presented system. The steps that are realized in the system are depicted in the process given below.

Step 1: Dataset Collection and Preprocessing – The Dataset containing biomedical text is obtained from the link https://ml4ai.github.io/BioContext/. The dataset is downloaded in the word document or .doc format. This doc file is furnished as an input to the system. A single string is formed using the biomedical text of the dataset file that is being read by the system. The string is given as an input to the preprocessing module after being segregated into sentences. The preprocessing is performed in 4 steps given below.

Special Symbol Removal – Special symbols are utilized in the English language to provide grammatical structure to the sentences. These symbols realize flow of speech such as !,?,., etc. These symbols are not influential for the semantics of the sentences. Such symbols are purged from the string completely.

Tokenization – The document has to be given as an input for an effective execution by the proposed system. Which is difficult to perform on a string format. Therefore, the tokenization procedure transforms the string into a well-indexed string. This enables the string to be easy to be converted into an array list.

Stopword Removal – Stopwords are certain words in the English language that are implemented to provide aesthetics to sentences. These words do not dispense any additional meaning to the sentence. Such words are not important for the execution of our methodology. These words have the ability to increase the processing time of the system significantly. These words are eliminated from the document in this step.

For example, the phrase “going to walk” is used to perform stopword removal. The stopword recognized in this phrase is “to” which will be purged during the execution and converted into “going walk”. This example details that the stopwords removal process does not modify the meaning of the sentence.

Stemming – Stemming replaces the words existing in the document to their root format. This is important as it can significantly reduce the size of the document and also due to the fact that stemming does not have any effect on the meaning of the sentence. Stemming significantly diminishes the resources required to process the document.

For example, “sleeping” will be stemmed into “sleep” through the reduction of the substring “ing” which is replaced with an empty character. It can be noticed that the semantic difference between “sleeping” and “sleep” is not significant but it can have a considerable impact on the time taken for the execution of the document.

Step 2: Feature Extraction –Five features are extracted in the presented technique for the input document provided. These steps are performed in an exact sequence utilizing the input document which is executed for both the training as well as testing purposes. The process is defined in the steps given below.

Noun Identification – The pre-processed string from the previous step is supplied as an input in this step. This module utilizes the pre-processed string to isolate the individual words present in the string. This is achieved through the splitting the string on the space character. The isolated words that are extricated are then compared using an Oxford Dictionary. The words in the dictionary are provided in a list which are matched to the words that are isolated. If the word does not match to any of the words in the dictionary, it is declared as a noun which is then dispensed with a score that is stored along with the word in a separate list.

Term Frequency and Inverse Document Frequency (TF-IDF) – The TF-IDF measurement is one of the most essential modules in the presented system. In this step the detection of the essential words is being executed through one of the most in-depth paradigms of natural language processing, viz. TF-IDF. A complete list of the words incorporated in the dataset doc files which are previously pre-processed and stored in the preceding steps. The lists of words along with their frequency are stored the in two columns of a double dimension list. Here one column is used for storing the frequency extracted from the document and another stores the respective word. The resultant list with the relevant frequency is stored as the Term Frequency list (TF).

The extracted values in the term frequency list are utilized to execute frequency analysis of the entire document and the existence of these terms is calculated. The logarithmic value of the documents is evaluated by inversing the count extracted previously. This is executed along with the ratio of the number of documents and the outcome of these values is realized as the Inverse document frequency (IDF). The TF-IFD is measured through the scalar product of TF or Term frequency and IDF or Inverse Document Frequency of the word. This is achieved through the utilization of the equation mentioned below in Equation 1.

TF-IDF=TF X Log (Number of Documents)/ (Number of Documents Containing Word W) ______ (1)

The extracted TF-IDF for a word is appended to the previous list containing the other features that are previously extracted.

Bag of Words – The obtained string which was pre-processed in the previous step is used to execute the Bag of Words approach. The Bag of Words approach matches all the words occurring in the list with a medical dictionary. If the word matches with the word in the dictionary then a score of 1 is provided. If the words that are compared from the list are not encountered in the dictionary, then a score of 0 is provided to that specific word. These BoW scores are appended to the list of features achieved in the previous steps.

Sentence Length – In this step the pre-processed string acquired in the previous step is implemented for the isolation of words. These individual words are the used to identify the original sentence. The length of the sentence is evaluated and appended to the respective word in the feature list. The information attained from the length of the sentence is significantly more than that of the other sentences. Therefore, the length of a sentence is very important feature and must be regarded as critical for the approach. The length of the sentence can be calculated using equation 2 below.

$$\text{SL}f = \frac{\text{Sentence Length}}{\mathit{Biggest\ Sentence\ Length}} \_\_\_\_\_\_ (2)$$

Sentence Position – The relevance of a sentence can be evaluated through the position of that sentence in the document. The position can ascertain the quality of the content that is incorporated in the sentence. Therefore, for the assessment of the sentence position a score is provided to the first five sentences in the document. The scores given in the following sequence according to the position of the sentence as 1, 0.8, 0.6, 0.4, and 0.2 respectively. Any sentences encountered henceforth are given a score of 0. This score is also appended to relevant words in the feature list.

The values achieved in the feature extraction process are stored simultaneously in the database and referred to as the trained data.

Step 3: Logistic Regression – After the consolidation of the training data in the previous steps, a test document is given as an input to the methodology. This test document is subjected all the steps performed previously in the feature extraction process. The five features are unravelled from this input document and stored in the form of a list. The extracted features from the training data and the list testing data list are compared with one another. To achieve this, the trained feature value is considered as an independent value x and the testing feature value is regarded as a dependant value y. These x and y values are provided as an input to the logistic regression technique. The logistic regression is calculated using equation 3 given below.

Y=Mx+B ____ (3)

Where:
x = how far up (Array of Attribute)
M = Slope or Gradient (how steep the line is)
B = the Y Intercept (where the line crosses the Y axis)
Y=Intercept value

The regression values which are calculated provide an output through the utilization of the x and y values given as input. These x and y values are also referred to as the intercept values. These intercept values are appended to the relevant word in a separate list called the regression list. The list is eventually sorted in a descending order using the bubble sort approach and the topmost 80% values are extracted and given as an input for the next step in the approach in the form of a regression list.

Step 4: Artificial Neural Network – The regression list created in the previous step is given as an input to the ANN module. The sorted regression list is made use of for the assessment of the two maximum and minimum target values, i.e., Target 1 and Target 2 respectively.

For this step, 5 attributes are assigned to 25 random weights. The weights are assigned in such a way that their mean is approximately 0.5.

The measured values are then classified depending on the boundary condition stipulated by the Target 1 and Target 2 values to assess the ANN probability score. The score is then appended to the specific row of an array list. The array list created is then provided as an input to the Decision Tree module for executing the classification approach. The process of ANN is performed based on the equation 6 and 7 given below.

$$T=\left(\sum_{k=0}^{n} AT * W\right) + B \_\_\_\_(6)$$

$$H_{LV}=\frac{1}{(1+\exp(-T))} \_\_\_\_\_\_\_ (7)$$

Where,
n- Number of attributes
A_T- Attribute Values
W- Random Weight
B- Bias Weight
H_{LV} – Hidden Layer Value

And this process of ANN is depicted in algorithm 1 given below.

ALGORITHM 1: Artificial Neural Networks

```
//Input: Regression List $R_L$, Weight set $W_S$= { }
//Output: Hidden Layer value list $H_{LV}$
hiddenLayerEstimation ( $R_L$, $W_S$)
Start
$H_{LV}$ =∅, index=0
  for i=0 to size of $R_L$
    ROW=$R_{L[i]}$
    for j=0 to 5
      sigma=0
      for k=0 to size of ROW
        ATR=ROW[k]
        sigma=sigma+ (ATR* $W_{S[index]}$)
        index++
      end for
      sigma=sigma+1
    Val=1/ (1+$e^{-sigma}$)
      $H_{LV=}$ $H_{LV}$+val
  end for
  end for
  return $H_{LV}$
Stop
```

Step 5: Decision Tree Classification – Here in this step the words which relate to the specific levels in this list are accumulated to form their respective clusters. An interactive user interface is then implemented to furnish the users with the classified words in addition to the respective Sentences. This representative of the medical terms that are established as the results of the methodology.

IV. RESULT AND DISCUSSIONS

The presented technique for detection medical terms in biomedical text data is achieved on a machine which is furnished with an Intel core i5 processor assisted with the 6 GB of RAM. The Machine is equipped with Windows operating system and the methodology is powered by Java Programming Language. For the implementation of the model NetBeans IDE is selected and database responsibilities are fulfilled by the MYSQL database server.

For the experimental assessment the presented model employs a dataset from the URL: https://ml4ai.github.io/BioContext/.

This URL contains the index page for all the data utilized in the biological context researches. The dataset downloaded from this URL contains a corpus of biomedical text accumulated on the PubMed central web portal. This dataset consists of 22 papers containing text based on biomedical context in word document format. The presented technique utilizes a small section of the papers for the purpose of training the methodology. Conversely, some papers were utilized for the testing purposes.

For the evaluation and the assessment of entropy each words which are used in the Decision tree are measured using the Shannon information gain. Through the Shannon information gain each word depicted as 'W' is counted for its existence in all the subsequent sentences and it is stored as 'A'. 'C' contains the total number of sentences and B provides the total number of sentences where the word has not occurred, so B can be written as (C-A). Therefore, according to the Shannon information gain as depicted in the equation 8, the entropy of each word is evaluated.

$$IG = -\frac{A}{C}\log\frac{A}{C} \;-\frac{B}{C}\log\frac{B}{C} \_\_\_\_\_\_(8)$$

Where

IG = Gain of the word

The entropy of each word is obtained and stored in the form of a double dimension list, where one column consists of the word and another stores the gain factor. This gain factor values have a maximum of 1 and a minimum value of 0. Any value closer to 1 symbolizes that the word is very important and if the values is closer to 0 then the word is not a significant. Therefore, this gain list is sorted in descending order and then its uppermost 80% of the data is regarded as the model word bag, which is subsequently utilized for the measurement of the sentence relationship using the precision and recall performance metrics.

Precision and recall performance metrics are one of the most important assessment parameters that extract the system's accuracy. The precision is calculated as the relative accuracy of the model. And Recall can be depicted as the absolute accuracy of the model. For a detailed depiction of precision, recall and F-Score the equations 9, 10 and 11 can be observed.

$$\text{Precision} = \frac{X}{X+Y}\text{------------ (9)}$$

$$\text{Recall} = \frac{X}{X+Z}\text{------------ (10)}$$

$$\text{F} - \text{Score} = \frac{(2*Precision*Recall)}{(Precision+Recall)}\text{--------- (11)}$$

Where,
X = the number of accurately Identified Medical terms.

Y= the number of inaccurately Identified Medical terms
Z = the number of accurate Medical terms are not Identified.

If the identified medical terms are encountered in the Information gain bag then it is deemed as an accurate identification or else it is regarded as an inaccurate identification. So based on this approach some experiments are performed and their outcomes are listed in the table 1 and it is graphically represented in figure 2.

S No	Dataset File Name	X	Y	Z	Precision	Recall	F-Score
1	A Novel Neural	8	3	3	72.72727273	72.72727	72.72727
2	COT MAP3K8	7	3	2	70	77.77778	73.68421
3	Cells and the role	9	4	4	69.23076923	69.23077	69.23077
4	Spatially Restricted activate	13	5	5	72.22222222	72.22222	72.22222
5	Structure of STING bound to c	12	5	5	70.58823529	70.58824	70.58824

Table 1: Precision, Recall and F-Score Reading

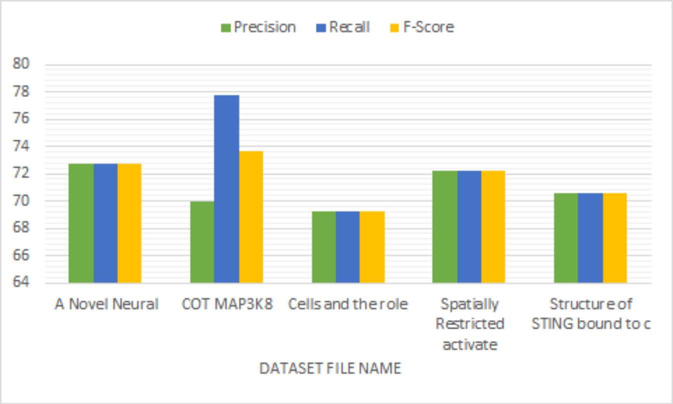

Figure 2: Precision, Recall and F-Score Measurement

On scrutinizing the plot in figure 2 we can discern that the average precision, recall and F-Score values are almost 71.71 %. When the obtained results are correlated with that of [16], which is predominantly concerned with the extraction of chemical disease sentence relationships through the utilization of Convolution neural network. The experimental outcomes in [16] have achieved Average precision of 51.9%, Recall of 7.0% and F-Score of 11.7 %. These parameters are calculated by comparing the obtained results of medical text identification with the GloVe vocabulary [17]. This is distinctly demonstrates that in [16] the performance is lower in comparison to our model. Our approach utilized articulate Machine learning algorithms such as Logistic Regression, Neural networks and Decision tree for detailed identification of the Medical words. Whereas on the other hand, the authors of [16] declared in their research article that their system has implemented rich neural networks. This ultimately culminates that the presented technique works competently in the first attempt of execution.

V. CONCLUSION AND FUTURE SCOPE

The proposed methodology for the detection of medical terms classification in biomedical texts has been defined in this research article. A dataset consisting of biomedical texts is provided to the system for the purpose of performing the NLP approach. Firstly, the dataset given as input is pre-processed to reduce any redundancy present in the dataset. After which the preprocessed string is given as an input to the NLP technique which executes a set of operations in a sequence to evaluate the relevant features. The NLP module executes the Noun identification and Bag of Words approach along with TF-IDF calculations, and Sentence position approaches. The output data is then stored as the training data and a test document is subjected to the system using the test data through logistic regression. The Artificial Neural Networks are provided the regression list for hidden layer estimation through effective neuron creation. The probability list then segregated using the Decision Tree classification to extract the relevant medical terms. Extensive experimentation has been executed to evaluate the performance metrics of the approach. The outcomes of the experimentation describe efficient performance for a first time implementation of such a system for identification of medical term classification from biomedical text dataset.

Future Research can be achieved through the realization of the proposed methodology in a real time set up like Chabot, Auto reply robots and etc. for effective and efficient medical term identification.

REFERENCES

[1] E. Tutubalina et al, "Medical concept normalization in social media posts with recurrent neural networks", Journal of Biomedical Informatics 84, 2018.

[2] N. Pattisapu et al, "Medical Concept Normalization by Encoding Target Knowledge", Proceedings of Machine Learning Research, 2019.

[3] C. Carbery et al, "Proposing the deep dynamic Bayesian network as a future computer-based medical system", IEEE 29th International Symposium on Computer-Based Medical Systems, 2016.

[4] S. Gupta and A. Manjhvar, "Relation Classification from Unstructured Medical Text using Feature Based Machine Learning Approach", International Conference on Trends in Electronics and Informatics, ICEI 2017.

[5] A. Zalewski et al, "Estimating Patient's Health State Using Latent Structure Inferred from Clinical Time Series and Text", IEEE EMBS Int Conf Biomed Health Information, 2017.

[6] V. Plachouras et al, "Quantifying Self-Reported Adverse Drug Events on Twitter: Signal and Topic Analysis", Proceedings of the 7th 2016 International Conference on Social Media & Society, 2016.

[7] B. Parlak and A. Uysal, "Classification of Medical Documents According to Diseases", 23nd Signal Processing and Communications Applications Conference (SIU), 2015.

[8] K. Denecke, "Extracting Medical Concepts from Medical Social Media with Clinical NLP Tools: A Qualitative Study", Proceedings of the Fourth Workshop on Building and Evaluation Resources for Health and Biomedical Text Processing, 2014.

[9] B. Parlak and A. Uysal, "The impact of feature selection on medical document classification", 11th Iberian Conference on Information Systems and Technologies (CISTI), 2016.

[10] Z. Pella et al, "Application for Text Processing of Cardiology Medical Records", World Symposium on Digital Intelligence for Systems and Machines (DISA), 2018.

[11] H. Yoon et al, "Model-based Hyperparameter Optimization of Convolutional Neural Networks for Information Extraction from Cancer Pathology Reports on

HPC", IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 2019.

[12] N. Pattisapu et al, "Medical Concept Normalization by Encoding Target Knowledge", Proceedings of Machine Learning Research, 2019.

[13] H. Al-Mubaid and M. Shenify, "Improved Bayesian-based method for classifying disease documents", World Symposium on Computer Applications & Research, 2016.

[14] S. Han et al, "Team UKNLP: Detecting ADRs, Classifying Medication Intake Messages, and Normalizing ADR Mentions on Twitter", Second Social Media Mining for Health Applications and Research Workshop, 2018.

[15] Jinghang Gu, Fuqing Sun, Longhua Qian and Guodong Zhoum ,"Chemical-induced disease relation extraction via convolutional neural network",Oxford University Press,2017.

[16] Pennington,J., Socher,R., and Manning,C.D. (2014) GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, 1532–1543.