International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Discovering Text classification for medical terms using the medical Social media text

Satpute Mayuri B. and Dr. Vinod V. Kimbahune

Computer Engineering Smt. Kashibai Navale College Of Engineering, Pune, India

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

Recently there has been a significant increase in the number of medical help websites that provide medical assistance to the people visiting for a variety of symptoms and illnesses. This platform's popularity can be attributed to the increased convenience offered to the patients who can get diagnosed from the comfort of their homes. The medical professionals on the website provide professional counseling and also take appointments for various specializations. As the interaction takes place on an online portal mostly in the form of text, to the medical professional or a chat bot, the deficiency in the patient's medical terminology can be detrimental as there is a significant risk that the doctor missed a part of the symptom's explanation. This is a highly problematic occurrence and the solution to this problem is the formulation of an efficient medical text classification technique. Therefore, this paper outlines an innovative technique that utilizes the Bayesian Classification model along with the addition of Artificial Neural Networks (ANN) coupled with Natural Language Processing (NLP) which achieves highly accurate Medical Text Classification.

Keywords: *Natural Language Processing, TF-IDF, Artificial Neural Network, Bayesian classification.*

Introduction

Following the extreme success of the internet, it was eventually opened up for public use, ever since the internet has grown in size exponentially, it is now a network that encompasses the whole planet which interconnects the whole world together. The internet then gave birth to a lot of facilities that utilized the internet platform to achieve their goals. One such offering of the internet is Social media. It has been used all over the world to communicate with one another.

The internet users have increasingly shared every aspect of their life on social media platforms to their loved ones, friends and relatives. This data also includes various different symptoms that they have been facing when contracting a disease or an ailment etc. The medical terms in this type of data are very specific and sometimes highly vague and can only be understood by a trained medical professional. This is where the paradigm of NLP or Natural Language Processing comes into the picture.

NLP or Natural Language Processing is a paradigm that aims to create an understanding of the human language for computers. Due to the fact that the computer only understands binary and high-level instructional languages, there is a need to facilitate the computers so that they can understand the language humans' speck to enable much better integration and interaction between humans and computers. NLP is a subsection of the Artificial Intelligence paradigm, which allows the computer to make sense, understand and decipher the traditional human language.

The NLP techniques are highly difficult as human speech contains a large number of nuances such as sarcasm, parts of speech, idioms and other different varieties that make it extremely difficult for the computer to understand. Therefore, extensive techniques to decompose and preprocess the text are deployed on the medical text. These techniques effectively break the language down to its fundamental components that can be used for the purpose of identifying the meaning behind the sentence. Along with the use of Natural Language Processing techniques, Artificial Neural Network has also been deployed along with the logistic regression to achieve effective and accurate medical text classification.

Logistic regression is a very useful concept that is applicable in the efficient extraction of the medical terminologies in the text. Logistic regression is one of the earliest known methods of regression analysis and has been in popular use since the early 20th century. Logistic regression belongs to a subsection of mathematics called Statistics. The machine learning platform has borrowed heavily from the statistical

mathematics as it has efficient techniques that can be used to decompose and process large amounts of data of different varieties effectively. The logistic regression is primarily used for the purpose of binary classification containing 2 class values.

The Logistic regression is called as such because of the sigmoid curve defined by the logistic function. The early use of logistic regression consisted of biological sciences and calculation of the population growth dynamics. The most prominent indication that logistic regression can be applied to a problem is when the dependent variable in the problem is categorical. The binary dependent variable is used for the analysis of the logistic function. Logistic regression has been used predominantly used in a plethora of fields such as social sciences, medical fields and most importantly machine learning. Logistic regression is utilized mainly for accurate predictions on the observable characteristics.

Artificial Neural Networks are also highly useful in understanding the various nuances of the text and be able to accurately identify medical terms in the texts. Artificial neural networks need to be trained extensively for this purpose to increase the accuracy of the system. The artificial neural networks have been highly accurate due to the fact that they are designed through inspiration from the human brain. Artificial neural networks form a rough draft of the human brain including all the different parts of the brain.

The human brain is a complex network of interconnections of billions of neurons that allow for such efficient processing of different types of data in real-time. This is due to tiny processing powerhouses called neurons that are in charge of the various different computations in the brain. The neurons are a fundamental part of the human brain. The neuron is also the smallest possible unit of the brain and it is where the artificial neural network gets its name from. The neurons are numerous and are interconnected with each other through synapses. The neurons are excited through sensory input, subject to a threshold or a limit. When the input surpasses the threshold, the neuron “fires”.

The firing of the neurons is a process by which the neuron generates a tiny electrical charge that flows down the neuron and into the adjacent neurons through synapses. The synapse is a small gap between two adjacent, connected neurons. The electrical charge jumps across the synapse in a process called synaptic transmission. The combinations, location, timing and the pattern of the neurons firing are what is considered as memory, perception or recall of the brain. The human brain is highly adept at storing information in this way, as billions of neurons and their connections form an extremely large number of patterns that are unique putting the memory capacity of the brain in the excess of a few terabytes.

This is exactly how the Artificial Neural networks are modeled, the neurons are created and then each of the neurons is given threshold value for excitation, if the input exceeds the threshold value, the neuron fires and excites the other neurons. These neurons work in large groups and can help identify patterns and model predictions. The artificial neural networks are predominantly used in applications where there is a need for human-like analysis is required such as text classification and prediction.

The use of Artificial Neural Networks or ANN for the purpose of medical text classification has been highly useful. The ANN must be trained beforehand for the identification of the medical terms. Artificial neural networks emulate the human brain and analyze the resultant text in detail to extract the relevant medical terms from it. Therefore, it is highly useful for this type of application where human expertise or way of thinking and analyzing is necessary for accurate application.

Ever since the introduction of the internet, there has been an increased number of individuals that have been utilizing this innovative platform for various different purposes. The internet was primarily designed to facilitate the researchers to communicate and share their data and other resources without having to actually be present in the location to utilize them for their research. The internet was developed by the united states military and was deployed all across the United States. This allowed the military to effectively communicate in between the various regiments all over the country, along with helping the researchers save time and resources in traveling to different locations around the world for their research.
In this paper, section 2 is dedicated for literature review of past work and Finally Section 3 concludes this paper.

Literature Review

This section of the literature survey eventually reveals some facts based on thoughtful analysis of many authors work as follows.

N. Limsopatham states that due to the popularity of social media websites nowadays a lot of individuals are online and interacting with the other users every day. Some of the interaction also contains some medical terms that are necessary to be recognized as it would allow for the monitoring of an epidemic disease. The processing of the language and determining medical ontologies is a complex task. Therefore, the authors have proposed a technique for the normalization of the medical text through the use of Natural Language Processing and Neural networks. The experimental results indicate that the proposed technique outperforms traditional approaches by a large margin. The major limitation of the approach is that the

authors have only achieved 44% normalization of the medical text.

V. Plachouras explains that various different drugs have different effects on an individual's body. There are a lot of different side effects depending upon the drug and various organizations and the pharmaceutical companies monitor the adverse effects of a drug on the populous. Therefore, the authors have devised a plan for the extraction of self-reported adverse effects of the drugs from social media websites such as Twitter [2]. The authors have utilized the Support Vector Machine or SVM for classification purposes along with efficient Natural Language Processing. The experimental results indicate that the system achieves a high level of accuracy from the NLP paradigm. The limitation of the proposed methodology is that the authors have not devised a system for the systematic collection of adverse effects from the verified users.

K. Denecke elaborates on the availability of medical data on various different experiences, treatments, and diagnoses on the social media platform. This is due to the fact that a large number of individuals utilize social media to express their thoughts and communicate and socialize with their loved ones. Due to the fact that an average individual is not well versed in medical terminology, they use standard phrases and common terminology, which is difficult to identify. Therefore, the authors present a technique for the extraction of medical concepts from social media using clinical Natural Language Processing tools [3]. The proposed methodology has been experimented extensively to verify the performance gains. The major drawback of this technique is that the authors have not combined the mapping tools with entity recognition to achieve higher accuracy.

E. Tutubalina introduces the impact of social media on the health of the users and the entire scientific community. The mining of various different medical texts from social media is highly reliable and can help analyze the spread of an epidemic of a deadly disease with accuracy [4]. The challenge in achieving this is due to the fact the users do not utilize the medical terms used by professionals when describing their symptoms. Therefore, the authors devise a technique based on Natural Language Processing along with the addition of the Recurrent Neural Networks that increase the accuracy of the prediction by a large margin. The major drawback of this technique is that the authors have not utilized the UMLS framework in their methodology.

S. Han states that there are various different systems that are designed for the purpose of text segmentation and classification. The author proposes that twitter posts can be used to identify Adverse Drug Reactions that can help the pharmaceutical companies identify the flaw in their drug and fix it. But due to the lack of medical terminology in the user's post makes this task a lot challenging. Therefore, the authors develop an innovative technique called Team UNKLP that detects and classifies the various different medical texts through the use of CNN, LSTM and the TLM approaches [5]. The proposed technique achieves an accuracy of 87%. The major drawback of this technique is the increased computational complexity that is observed.

N. Pattisapu explains that there is a need for a medical concept normalization technique to understand the symptoms of a patient accurately. This is a challenge due to the fact that an average individual is not aware of the medical terminology which leads to a very vague medical text that needs to be normalized automatically [6]. Therefore, the authors present a novel technique for the purpose of medical text normalization through encoding the target knowledge. The experimental results demonstrated the superiority of the proposed technique in comparison to the traditional normalization techniques. The main limitation of this technique is that the authors have not utilized the Knowledgebase embeddings for the knowledge encoding process.

B. Parlak elaborates on the various advances that have been made in the field of technology with fields like automated data processing, Natural Language Processing, and Machine learning has made great strides and has conveyed a lot of potential in the medical field [7]. The authors also state that these have not been utilized fully to their limits in the field of medical sciences, therefore, an innovative approach towards integrating customer medical terminologies and utilizing it for the purpose of classifying medical documents containing health terminologies. This is achieved through the use of LSTM, CNN, Extreme Gradient Boosting and Naïve Bayesian Model. The major limitation of the proposed methodology is the increased time complexity that is observed.

C. Carbery states that there is a growing need for the development of various different learning models as these learning models are highly useful and helpful for performing deep learning on various different types of data and generate insight which can be highly valuable. The authors have analyzed the popular deep learning techniques in detail and elaborated on their advantages and disadvantages [8]. The researchers have also proposed a dynamic probabilistic system that is fully integrated with deep learning. The presented deep learning technique has been generalized to be applicable to a wide variety of applications such as medical imaging and predictive analysis. The major limitation of this technique is that it has not been analyzed for its performance on medical texts.

S. Gupta explains that there is a lot of information available in medical documents pertaining to the patient as well as the illness and the treatment that is

being offered. This is valuable information that is highly useful in treating other patients with similar ailments. Due to the unstructured nature of the information that is stored on the documents, the information has to be extracted manually which is highly timeconsuming. Therefore, the authors have proposed a data mining approach that extracts the relevant features through the medical terminologies automatically, saving valuable time for medical professionals. The experimental results indicate that the technique has performed efficiently. The main drawback of this technique is that the accuracy of the methodology is below par and can be further improved.

Z. Pella introduces the field of analysis of medical data and patient records as it is a highly useful procedure that can help alleviate the patient's stress and pain by a large margin. The author's comment that the medical data has been utilized for this purpose needs to be anonymized to prevent the leakage of personal data of the patients [10]. The authors have utilized the Natural Language Processing paradigm for the assessment of the text and for classification purposes. The experimental results indicate that the proposed methodology produces promising results. The main limitation of the proposed technique is that the authors have only utilized the data from the cardiovascular for the purpose of analysis.

H. Yoon elaborates on the topic of cancer which is one of the most debilitating ailments that have been plaguing humans physically and mentally. Cancer is highly dangerous and has a very low survival rate in humans suffering from this disease. The authors have proposed an efficient technique for the optimization of the information extraction technique through the use of Conventional Neural Networks [11]. The authors have also utilized the Bayesian optimization through the use of the CANDLE framework which achieves a high-performance computing environment. the experimental results indicate the superiority of the proposed technique. The main drawback of the presented approach is that the technique increases the computational load on the system.

A. Zalewski states that there is a large amount of data that is being generated by the intensive care units that have a lot of critically ill patients that are being monitored. There isn't a technique that is designed to analyze and interpret the data that is being generated in such a large volume. Therefore, the authors in this paper have proposed a Bayesian non-parametric framework named HDP or Hierarchical Dirichlet Process that assists the medical professionals to analyze the data efficiently and faster [12]. The experimental results indicate that the proposed technique has been successfully implemented. The main limitation of this technique is that the authors have not automated the system for the feature learning process.

H. Al-Mubaid explains the process of classification of different disease documents automatically, as the manual segregation is highly time-consuming and leads to a waste of the medical professional's valuable time. Therefore, there is a growing need for a system that is capable of achieving automatic classification [13]. The authors in this paper propose an innovative technique that utilizes an improved Bayesian algorithm for the purpose of automatic disease document classification. The presented technique was demonstrated through a series of experiments that produced promising results. The major limitation of this research is that the authors have not normalized the weight of the attribute based on its probabilities.

B. Parlak introduces the concept of medical document classification and its various different techniques. The field of automatic medical document classification is one of the most popular fields that are used for the purpose of accurate classification of the documents based on the disease, which can be used in the future for an individual with a similar set of symptoms [14]. Therefore, to ameliorate these effects the authors have proposed a technique for automatic classification of medical documents through the use of a Bayesian Network Classifier and Distinguishing Feature selector. The experimental results indicate that the proposed methodology executes with acceptable accuracy. The major limitation of this research is that the authors have not utilized different languages for classification purposes.

K. Liu explains that there have been significant advancements in the area of Natural Language Processing, Machine Learning, etc. these applications require the use of a very large amount of data for their processing. At this juncture, social media comes into play where massive amounts of data are being generated every day. therefore, the authors in this paper have proposed an efficient technique for the Medical Social Media Text Classification (MSMTC) [15]. The researchers utilize adversarial networks for the extraction of health terminology. The experimental results indicate that the proposed methodology produces efficient results. The main limitation of this paper is that the authors have not covered the drug adverse reaction and disease severity classification in their methodology.

Proposed Methodology

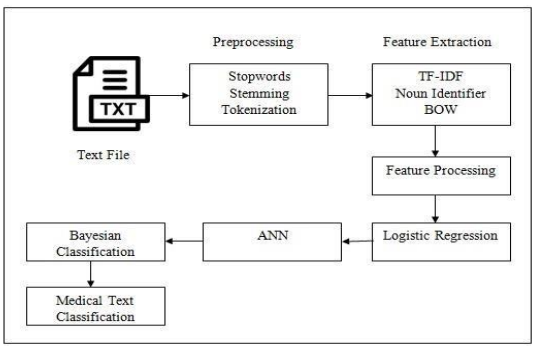

Figure 1: Proposed model System Overview

The proposed methodology for medical text classification's system overview diagram is depicted in the above figure 1.And the steps that are involved in this process are described below.

Step 1: *Data preprocessing* – This is the initial step of the proposed model where a Dataset of the social media medical text is being fed to the system. This dataset eventually consists of an unstructured text that is being used for the conversational process in the social media paradigm.
For the experimental use the proposed model uses the *DingXiangyisheng*'s question and answer module. This contains the patient's description texts as the input. This text is in the form of the unstructured data so it needs to preprocess. So this step of preprocessing eventually contains 4 major steps through which the data have gotten rid of the unwanted text. The steps that are involved in the preprocessing are listed and described below.

Special Symbol Removal – This is the initial step of the preprocessing, where the unwanted special symbols are shredded off by replacing them with an empty character. By doing this the input text is free from the special symbols. The special symbols can be like?,!,, Etc.

Tokenization- Once the special symbols are removed then the words of the text need to be stored in a list, because by storing the words in a list that makes easy for the processing of the text in the coming steps. For this purpose the proposed model uses the split( ) function of the Java to spit words on empty space character " " , and then store them in a list for the efficient processing.

Stopword Removal – As we know each and every language are containing the conjunctions. The removal of the conjunctions doesn't affect the core meaning of the sentences. By keeping this thing in mind proposed model shreds off the stopwords from the input text by replacing them with an empty character.

Stemming- Here in stemming process the post fix words are trimmed to get the base words. By doing this, words become small and space complexity is also can be easily manageable.

Step 2 : Feature Extraction and Feature processing – This is the decisive step of the proposed model where the features are extracted and stored in a well-managed list. Here this step handles mainly 3 features as described below

TF-IDF : This refers to identify the most important words from the given string. For this purpose initially a term frequency of the each of the words in the text is being identified that is referred by the term TF. And then Inverse document frequency is estimated by the logarithm of the ratio of the number of documents to the number of documents that contain the word. So by doing this each and every word of the fed text will get the value of the TF-IDF which is evaluated using the equation1.

$$\mathrm{TF-IDF} = TF\ X\ Log \frac{Number\ of\ Documents}{Number\ of\ Documents\ Containing\ Word\ W}$$

Noun Identification: Here in this step the nouns are being identified from the given textual dataset using the dictionary words. Here this dictionary contains around 1,0700 words that represents all most all possible words using in the English language collected from the Oxford Dictionary site. The process of Noun Identification can be represented using the following algorithm 1.

Algorithm 1: Noun Estimation

//input: Set of textual word T_{SET} , Dictionary List D_L
// Output: Noun List N_L

Step 0: Start
Step 1: Initialize N_L= NULL
Step 2: **FOR** i=1 to size of T_{SET}
Step 3: WRD= $T_{SET[i]}$
Step 4: Set FLAG=TRUE
Step 5: **FOR** j=1 to size of D_L
Step 6: **IF** (D_L[j] == WRD)
Step 7: Set FLAG=FALSE
Step 8: Break
Step 9: **END IF**
Step 10: **END FOR**
Step 11: IF (FLAG==TRUE)
Step 12: ADD WRD to N_L
Step 13: **END FOR**
Step 14: return N_L
Step 15: Stop

Bag of Words: Here in this step the tokenized input words are matched with the stored medical terms and then they are collected in separate list to call as the BOW features.
Once all these features are extracted they are indexed properly to store in a double dimension list to process the extracted features in a static object.

Step 3: Logistic Regression – Here in this step a preliminary prediction for the natural words are carried out for being matched with the medical terms. This can be done the using logistic regression analysis process based on the equation 2.

Y=mx+b _____(2)

Where:

- y = how far up

- x = Current word features score
- m = Gradient of the input words towards the medical text
- b = Intercept distance of the word towards the medical text

Step 3: Artificial Neural Network and Bayesian Classification model – All the collected features from the past steps and regression values are considered as the input layers for the ANN model. These input layer data are being used by the hidden layer using the sigmoid activation function to predict the output layer. Where output layer data indicate the term classification score for the medical terminology.

These scores are used by the Bayesian classification probability function to ultimately estimate and take the decision that whether a term from the input text is deserved to be a medical text or not.

The whole proposed system is expressed mathematically with the below model.

Mathematical Model

1. S= { } be as system for Medical Text Classification
2. Identify Input as T_F={ T_{F1}, T_{F2}, T_{F3}..... T_{Fn}}

Where T_{Fn} = Text File

3. S={ T_F }

4.Identify M_{TC} as Output i.e. Medical Text Classification S= { T_F , M_{TC} }

5. Identify Process P S= { T_F , P ,M_{TC} }
6. P= {P,F_E,L_R,A_{NN},B_C}

Where

P=Preprocessing
F_E =Feature Extraction
L_R= Logistic Regression
A_{NN}= Artificial Neural Network
B_C= Bayesian Classification

So the Complete system for heart failure prediction can be given as

7. S = { T_F ,P,F_E,L_R,A_{NN},B_C, M_{TC}}

Results and Discussions

The presented technique for the purpose of medical text classification has been developed using the Java Programming language on the NetBeans development environment. The implementation machine is of a standard configuration consisting of an Intel Core i5 processor paired with 4GB of primary memory and 500 GB of storage. For storage purposes, the MySQL database server is being utilized.

To ascertain the performance and efficiency of the proposed medical text classification technique there is a need to perform extensive experimentation. The presented technique has been evaluated for the incidence of various different errors in the classification purposes through the use of the Mean Absolute Error (MAE).

Performance Evaluation based on Mean Absolute Error

The performance evaluation through the use of Mean Absolute Error is executed in terms of percentage. This is due to the fact that a percentage is one of the easiest forms to understand and interpret. The continuous parameters are selected for use in this technique, for the evaluation of the TFIDF, Term Frequency and Inverse Document Frequency. The error percentage for the classification of words through term and inverse document frequency is evaluated through the Mean Absolute Error. The mathematical formulation of the process is detailed below.

$$\text{MAE} = \frac{(\sum_{i=1}^{n} | xi - yi |)}{n}$$

Where, xi - Number of obtained classified words through TF-IDF. yi - Number of actual classified words through TF-IDF.
n - Number of Experiments Conducted.

Table 1: Experimentation and Calculation of MAE.

Experiment Number (n)	Number of actual classified words (xi)	Number of Obtained classified words (yi)	Difference (xi-yi)
1	10	8	2
2	15	12	3
3	13	10	3
4	17	16	1
5	9	8	1
6	14	12	2
7	11	8	3
8	19	18	1
9	12	9	3
10	10	8	2
		MAE	2.1

The table above tabulates the values of classification results along with the expected results to calculate the Mean Absolute Error of the proposed methodology. The results obtained are plotted onto a bar graph for easier visualization and interpretation in figure 2 given below.

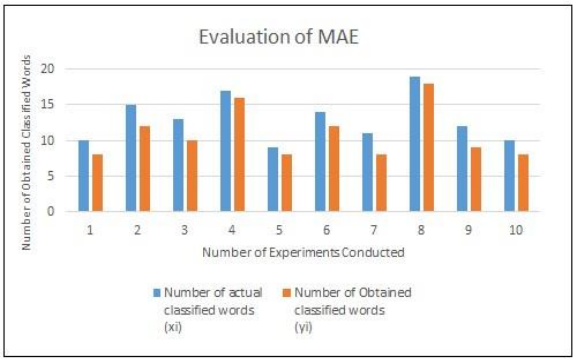

Figure 2: Evaluation of MAE.

The experimentation results indicate the important aspects of the text classification through the use of Term Frequency and The Inverse Document Frequency. The experiments conducted depicted that the technique for the classification yielded a Mean Absolute Error of 2.1. This is an indication that the proposed methodology is working as intended and has reached the halfway mark.

Conclusion and Future Scope

Due to the modernization of every aspect of human life, technology has infiltrated almost every part of human life. The largescale growth in the number of online medical advice and medical expert diagnosis is a testament to this argument. This also indicates the shift towards utilizing technology for convenience and medical problems. But there are certain limitations that are specially observed when a patient conveys their symptoms such as, the lack of vocabulary and not being adept in medical terminology. The medical terminologies used by the doctor are very different and is usually out of the expertise of a normal human being. This significant disparity in the though process and understanding of the disease between the doctor and the patient is increased due to the physical limitations of the online chat process. The doctor or the chatbot might unintentionally miss out on some symptoms or misidentify the symptoms due to a wrong word used by the patient, which could lead to disastrous results. Therefore, a medical text classification technique has been elaborated in this publication which has been deployed with the assistance of the Bayesian Classification Model along with the inclusion of Natural Language Processing (NLP) and Artificial Neural Networks (ANN). For the purpose of future work, the proposed technique can be implemented in a real-time application that can be implemented on the web page of a chat-bot.

References

[1]. N. Limospatham and N. Collier, "Normalizing Medical Concepts in Social Media Texts by Learning Semantic Representation", Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016.

[2]. 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.

[3]. 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.

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

[5]. 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.

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

[7]. B. Parlak and A. Uysal, "Classification of Medical Documents According to Diseases", 23nd Signal Processing and Communications Applications Conference (SIU), 2015. [8] 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.

[8]. 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.

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

[10]. 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.

[11]. 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.

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

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

[14]. K. Liu and L. Chen, "Medical Social Media Text Classification Integrating Consumer Health Terminology", IEEE Access, 2019