International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT23903102

A Survey on Medical Health Records and AI

Shubham Shinde, Mitesh Shetkar, Mayuri Shigwan, Abhishek Shinde, Sai Shinde
Department of Computer Engineering, Zeal College of Engineering and Research, Pune, Maharashtra, India

ARTICLEINFO

Article History:

Accepted: 01 June 2023
Published: 05 June 2023

Publication Issue
Volume 9, Issue 3
May-June-2023

Page Number
368-372

ABSTRACT

As the population increases, so do the number of patients getting admitted in hospitals. This generates an overwhelming amount of data within the electronic health records (EHRs) that is impossible to manage manually. This is where machine learning concepts come in handy. The ML algorithms for regression-classification have become increasingly popular within the healthcare sector. Especially for cardiovascular diseases (CVD). Estimation states that by 2030, over 23 million people will die from CVD each year. But, it is estimated that 90% of CVD is preventable. The on time recognition and diagnosis of heart failure from the pre-existing medical records is a way to do so. However, the EHRs are not particularly reliable when it comes to the comparison of structured and unstructured data. This is due to the use of colloquial language and possible existence of sparse content. To tackle this issue, the proposed system uses the KTI-RNN model for recognition of unstructured data and removal of sparse content, TF-IWF model to extract the keyword set, the LDA model to extract the topic word set, and finally, the GA-BiRNN model to identify heart failure from extensive medical texts. The GA-BiRNN model is made up of a bidirectional recurrent neural network model and its output layer embedded with global attention mechanism and gating mechanism.

Keywords : Recurrent Neural Network, Electric Health Record (EHR), Cardiovascular Diseases Natural Language Processing (NLP), Bi-directional Recurrent Neural Network (BiRNN), Unstructured data.

I. INTRODUCTION

The diseases are the part of human life and are drastically increasing its presence our daily life, some of them are acute diseases which develop suddenly and last for a very short period whereas some are Chronic diseases which develop slowly and may worsen over period of time and are often life threatening. The increase of diseases in an average life of human being is due to the fast life it is living, causing improper timings of meals, poor quality of food consumption, poor quality of hospitalization etc. The life-threatening diseases such as cancer, diabetes, heart attack are the nightmares for every person suffering from it. Talking particularly about heart related diseases and it's diagnosis are the complex things and do require timely

and reasonable treatment. This treatment can be provided by early detection of symptoms of Heart failure such as shortness of breath, ankle swelling, ling crackles etc [2]. Thus, early detection of heart failure can reduce burden of this disease on individuals.

Medical text plays a very vital role in medical profession which provides significant data for diagnosis and pathological study. The digital medical records of patients are maintained and used by the hospitals. The Electronic Health Records (EHR) are the abundant source of clinical information of patient's disease [3]. Hospitals standardize and integrate these records written by doctors. All these records are mostly unstructured, diverse, and heterogeneous. At time, researchers only use structured EHR records, that contains previous diagnosis and treatment process and does not include patient's clinical information as his/her previous disease and current disease information. Anyhow if this EHR data is combined with the clinical records of patient, prediction can be improved further. Thus, this data can be used and studied by doctors which will help them better understand the patient's medical condition and provide reasonable treatment accordingly. Thus, the use of this unstructured data has a good significance in Clinical treatments.

The EHR data is unstructured and do contains diverse and sparse contain, moreover proper classification and processing of this data will improve the readability of the data. For example, just removal of corpus and sparse texts from itself can increase the recognition efficiency. This data can be processed by using some Natural Language Processing Algorithms. The processing can include removal of sparse and corpus texts, expanding medical texts, keyword extraction and text recognition [6]. Along with medical records classification, its use in diagnosis process can play a vital role in providing accurate suggestion for the timely diagnosis of the patient. Thus, this classified and processed data needs to be presented in a manner that can be easily accessed by doctors.

The proposed system named KTI-RNN model helps the processing of data and its representation to doctors and medical works [1]. The model uses recurrent neural networks, keyword extraction techniques etc.

II. MODEL ARCHITECTURE

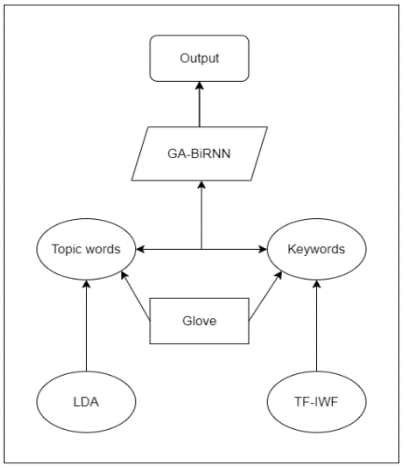

Fig.1 : Basic KTI-RNN Architecture

A. LDA Model

Latent Dirichlet allocation [4] is one of the most popular methods for performing topic modelling. Each document consists of various words and each topic can be associated with some words. The aim behind the LDA to find topics that the document belongs to, based on words contains in it. It assumes that documents with similar topics will use a similar group of words. This enables the documents to map the probability distribution over latent topics and topics are probability distribution.

B. TF-IWF Model

TF-IDF stands for Term Frequency Inverse Word Frequency of records. It can be defined as the calculation of how relevant a word in a series or corpus is to a text. The meaning increases proportionally to the number of times in the text a

word appears but is compensated by the word frequency in the corpus (dataset).

Terminologies:

- **Term Frequency:** In document d, the frequency represents the number of instances of a given word t. Therefore, we can see that it becomes more relevant when a word appears in the text, which is rational. Since the ordering of terms is not significant, we can use a vector to describe the text in the bag of term models. For each specific term in the paper, there is an entry with the value being the term frequency.
- **Document Frequency:** This tests the meaning of the text, which is very similar to TF, in the whole corpus collection. The only difference is that in document d, TF is the frequency counter for a term t, while df is the number of occurrences in the document set N of the term t. In other words, the number of papers in which the word is present is DF.
- **Inverse Document Frequency:** Mainly, it tests how relevant the word is. The key aim of the search is to locate the appropriate records that fit the demand. Since tf considers all terms equally significant, it is therefore not only possible to use the term frequencies to measure the weight of the term in the paper. First, find the document frequency of a term 't' by counting the number of documents containing the term:

Term frequency is the number of instances of a term in a single document only; although the frequency of the document is the number of separate documents in which the term appears, it depends on the entire corpus. Now let's look at the definition of the frequency of the inverse paper. The IDF of the word is the number of documents in the corpus separated by the frequency of the text.

C. Bidirectional Recurrent Neural Networks (Bi-RNN)

Bidirectional Recurrent Neural Networks (Bi-RNN) [5] connect two hidden layers of opposite directions to the same output. With this form of generative deep learning, the output layer can get information from past (backwards) and future (forward) states simultaneously. BRNNs were introduced to increase the amount of input information available to the network. For example, multilayer perceptron (MLPs) and time delay neural network (TDNNs) have limitations on the input data flexibility, as they require their input data to be fixed. Standard recurrent neural network (RNNs) also has restrictions as the future input information cannot be reached from the current state. On the contrary, BiRNNs do not require their input data to be fixed. Moreover, their future input information is reachable from the current state.

BiRNN are especially useful when the context of the input is needed. For example, in handwriting recognition, the performance can be enhanced by knowledge of the letters located before and after the current letter.

D. KTI-RNN

The study introduces a model called KTI-RNN [1], which aims to enhance the content of medical text by incorporating topic words and keywords. Additionally, an upgraded classifier is employed to accomplish the classification of medical texts. The KTI-RNN model consists of three modules: the input module for extracting TF-IWF keywords and LDA topic words, the GA-BiRNN module, and the classification output module [1]. The basic architecture diagram depicting these modules is illustrated in Figure 1. The steps can be outlined as follows:

1. Initial text processing and test set formation: The first step involves preprocessing the original text and creating sets of data for testing purposes.

2. LDA model processing on the test sets: Once the test sets are ready, the LDA model is applied to process the texts within each set. This process helps in extracting topic-specific words from each set of test data.
3. TF-IWF model for category keyword extraction: The TF-IWF model is utilized to extract a set of keywords associated with each category. This step helps in identifying relevant keywords for each set of data.
4. Construction of the GA-BiRNN model and integration of pretrained Glove [8] word vectors: The next step involves building the GA-BiRNN model and incorporating pretrained Glove[8] word vectors. These vectors are combined with the extracted keywords and topic words to train the neural network model.
5. Training and predictive classification using the GA-BiRNN model: The constructed GA-BiRNN model is trained using the prepared data. It learns patterns and relationships between the input text, keywords, and topic words.

III. APPLICATIONS

A. Clinical Decision Support:

The system can provide suggestions and medical references to doctors for timely treatment of patients with cardiac conditions, such as heart failure. By analyzing the extensive medical texts and extracting relevant information, the system can assist healthcare professionals in making informed decisions about patient care.

B. Early Detection and Prevention Programs:

The system can be integrated into healthcare systems to identify individuals at high risk of cardiovascular diseases, such as heart failure. By analyzing the electronic health records of patients, the system can identify patterns, risk factors, and early warning signs, enabling healthcare providers to implement preventive measures and interventions.

C. Disease Surveillance and Outbreak Detection:

The system can contribute to disease surveillance efforts by analyzing large volumes of medical texts and identifying patterns related to cardiovascular diseases. It can assist public health agencies in detecting outbreaks, monitoring disease trends, and implementing targeted interventions to prevent and manage cardiovascular conditions at a population level.

D. Personalized Medicine:

By analyzing the extensive medical records and extracting relevant information, the system can assist in the development of personalized treatment plans for patients with cardiovascular diseases. It can help healthcare providers tailor interventions, medications, and lifestyle recommendations based on individual patient characteristics, previous medical history, and response to treatment.

E. Clinical Trials and Drug Development:

The system can aid in the identification of potential candidates for clinical trials related to cardiovascular diseases. By analyzing the structured and unstructured data from electronic health records, the system can identify eligible patients, assess their suitability for specific trials, and support the recruitment process. It can also contribute to the evaluation of drug efficacy and safety by analyzing real-world patient data.

F. Health Insurance and Risk Assessment:

The system can be utilized by health insurance companies to assess the risk profiles of individuals and determine appropriate insurance coverage. By analyzing medical records and identifying risk factors associated with cardiovascular diseases, the system can support insurance underwriting processes and help in setting premiums and coverage options.

IV. CONCLUSION

In conclusion, proposed a system for identification of heart failure from massive hospital databases using the KTI-RNNmodel, which includes the TF-IWF and LDA models for extraction of keyword and topic word sets from medical notes. It also involves the enhanced version of the BiRNN model, the GA-BiRNN model, which is usedto train and classify the medical texts.This model can help reduce the suffering anddeath of many and reduce the pressure on people working in the medical field thathave to look through large amounts of data for identifying heart failure. Further researchcan be done to diagnose heart failure by combining structured and unstructured data.Testing of different word embedding methods for clinical notes using KTI-RNN is also possible.

IV. REFERENCES

[1]. Dengao Li , Huiting Ma, Wenjing Li, Baofeng Zhao, Jumin Zhao, Yi Liu, and Jian Fu, KTI-RNN: Recognition of Heart Failure from Clinical Notes, TSINGHUA SCIENCE AND TECHNOLOGY ISSNll1007-0214 11/18 pp117–130 DOI: 1 0. 2 6 5 9 9/T ST. 2 0 2 1. 9 0 1 0 0 9 3.

[2]. A.Triantafyllidis, C.Velardo, T. Chantler, S.A.Shah, C.Paton, R.Khorshidi, L. Tarassenko, K. Rahimi, and on behalf of the SUPPORT-HF Investigators, A personalized mobile-based home monitoring system for heart failure:The support-HF study, International Journal of Medical Informatics, vol. 84, no. 10, pp. 743–753, 2015.

[3]. M. Z. Nezhad, D. Zhu, N. Sadati, K. Yang, and P. Levi,SUBIC: A supervised bi-clustering approach for precision medicine, arXiv preprint arXiv: 1709.09929, 2017.

[4]. D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent dirichlet allocation, The Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003

[5]. J. L. Leevy, T. M. Khoshgoftaar, and F. Villanustre, Survey on RNN and CRF models for de-identification of medical free text, Journal of Big Data, vol. 7, no. 1, pp. 1–22,2020.

[6]. H. Liang, B. Y. Tsui, H. Ni, C. C. S. Valentim, S. L. Baxter, G. Liu, W. Cai, D. S. Kermany, X. Sun, J. Chen, et al., Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence, Nature Medicine, vol. 25, no. 3, pp. 433–438, 2019.

[7]. V. Carubelli, G. Cotter, B. Davison, J. Gishe, S. Senger, I. Bonadei, E. Gorga, V. Lazzarini, C. Lombardi, and M. Metra, In-hospital worsening heart failure in patients admitted for acute heart failure, International Journal of Cardiology, (vol. 225, pp. 353–361, 2016.)

[8]. J. Pennington, R. Socher, and C. D. Manning, GloVE: Global vectors for word representation, in Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, pp. 1532–1543.

Cite this article as :

Shubham Shinde, Mitesh Shetkar, Mayuri Shigwan, Abhishek Shinde, Sai Shinde, "A Survey on Medical Health Records and AI", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 3, pp.368-372, May-June-2023. Available at doi : https://doi.org/10.32628/CSEIT23903102
Journal URL : https://ijsrcseit.com/CSEIT23903102