TLBIONER: TRANSFER LEARNING BASED NAMED ENTITY RECOGNITION ON MEDICAL LITERATURE DOCUMENTS

Dr. K. Arutchelvan

Assistant Professor, Department of Computer and Information Science, Annamalai University, Chidambaram, Tamilnadu, India
Email: karutchelvan@yahoo.com

R. Ramachandran

Assistant Professor, Department of Computer and Information Science, Annamalai University, Chidambaram, Tamilnadu, India
Email: ramachandranr.au@gmail.com

Abstract

Nowadays, Natural Language Processing (NLP) plays a significant role in extracting the concealed information from the unstructured data which is being loaded with voluminous data over the web. Various tasks such as Tokenization, Stemming, Parts-of-Speech identification, Lemmatization, Named Entity Recognition (NER), etc are being popular in NLP research area. In recent years, NER is getting more attention among the researchers to extract the important entities from the huge set of documents. In life science domain NER is playing major role to identify the medical-term entities from the medical related documents such as literature documents, clinical trials, Electronic Medical Record (EMR), etc. This research work aims to provide a new NER approach to get the named entities from the medical literature documents. Instead of build and trained a new model, the proposed model works based on the Transfer Learning. In order to reduce the training time, the pre-trained model is re-trained with the newly annotated entities. The proposed NER produces better accuracy and able to identify a greater number of entities. The NER model is experimented with PubMed articles.

***Keywords*: Transfer learning; Spacy; BioNER; Natural Language Processing.**

1. Introduction

In the medical domain, extracting the hidden information from the medical literature written document is much difficult for both researchers and industrialists. Splendid information in the medical literature document is unstructured in nature. Electronic Medical Records (EMR) have lot of insightful information which helps the healthcare management to take meaningful decisions. The hidden information from the unstructured data is being extracted by the developers with help of advancements in the Artificial Intelligence. The medical domain contains huge corpus with the collection of literature documents, websites, forums, EMR, etc. These documents are having high rich text contents. Text mining is the research area which aims to analyze the huge text corpus and helps to take actionable insights. Natural Language Processing (NLP) is one of the unavoidable concepts in text mining. The role of AI in NLP makes the developers to extract the structure information from the big data. Several libraries are being developed in NLP for identifying the important keywords in the EMR [Ramachandran R and Arutchelvan K (2021)].

Due to the advancement of research and discoveries in the life science domain, huge number of applications are being introduced for various analysis. For example, the diseases persons who consumes drugs are being faced with some side effects and it is known as adverse reactions (ADR) [Yao Chen et. al. (2019)]. ADR is the unhappy reactions of medications that happen on consuming ordinary portion of endorsed drugs. These may differ after slight medical problems, for example, skin rashes to dangerous issues, for example, renal issues, cardiovascular breakdown and even demise now and again. The event of ADR, if there should be an occurrence of a medication, impacts an enormous populace so their ideal identification is critical to spare individuals from unsafe results of medications [Sara Santiso et. al. (2021), Seonghyeon Moon et. al. (2021), Kumar et.al. (2021)].

In this research work, the major issue of NER extraction from the EMR has been focused to identify the important entities during ADR processing. The arrangement expected this research work to separate the important elements, for example, endorsed chemicals with dose. Also, it extracts the indices and illnesses referenced in the EMRs. The separated elements will be prepared for further downstream to connect the elements. Then, it influences the

word reference-based strategies to hail any side effects which might actually be inauspicious medication responses of the recommended meds.
The remaining section of this paper is organised as follows. Section 2 describes the background study of the proposed work. The data preparation and proposed research flow is discussed in the section 3. The experimented results are given in the section 4. Conclusion of this research paper is given in the section 5.

2. Background Study

Taking in word depictions from a ton of unannotated text is a since a long time prior settled system. While past models for instance Word2Vec [Payal Biswas and Aditi Sharan (2021)], GloVe [Ning, Gelin and Bai, Yunli (2021)] focused in on getting the hang of setting self-sufficient word depictions, late works have focused in on picking up setting subordinate word depictions. For instance, ELMo uses a bidirectional language model, while Inlet [17] uses machine understanding to embed setting information into word depictions.
NER is a sort of information recuperation which revolves around perceiving models i.e., names of various kinds of components. For example, threat would be an instance of disorder; becoming would be an illustration of signs, and so on Latest NER models are relied upon decision tree [X. Dong et. al. (2016)]. Sekine developed a system was created for Japanese. The maker used feature through grammatical feature (linguistic component) marks isolated by a morphological analyzer, information reliant upon character and thought word reference. The expert [Ravikumar J. and, Ramakanth Kumar P (2021)] presented an estimation strategy which included two phases are as per the following: one for decision tree creation from planning data and other for delivering named data that relies upon decision tree.
In another work the authors [Kumar et.al. (2021)] talked about different methodologies proposed in NER to determine the issues introduced previously. The authors have examined about the standard put together NER approaches with respect to the text-based data. Before the advancement of AI in the existence science area, word reference-based methodology has been utilized. The word references are huge assortment of predefined elements. The word reference-based methodology is controlled by the standard based NER. The greater part of the principles is physically produced. The word references are kept up with independently for infections, synthetic compounds, qualities, and so forth [Ramachandran R and Arutchelvan K (2021)].

3. TLBioNER

Considering ADR during medication on the disease affected person is a key to tranquilize improvement in medical care. Pharmacology-Vigilance (PV) [Sajid Hussain et.al. (2021)] as portrayed by World Health Organization (WHO), it defined PV as “the science and exercises identifying with the recognition, appraisal, comprehension and anticipation of antagonistic impacts or some other medication related issue.” Drug organizations on a regular basis, it needs to assimilate the conditions and pre-conditions of the ADR after intaking the medicine. This will help the PV department in the examination and analysis of the medications. Furthermore, it will diminish or anticipating insecurity of any misconduct to the diseased person. Co-event of illness and synthetic compounds in an EMR of a diseased someone is valuable during the examinations and analysis for many chemical and drug-related manufacturers.
However, EMR is an unstructured document which is having valuable data. It requires self-attention to extract the NER from the document. Above-mentioned extraction could make the analysis easier and reduce the human interception in producing the reports. The NER extraction application will reduce more time for the organizations. It created custom medical services NER models to remove phrases identified with (drug) synthetics with measurements, infections and indications from EMRs. A newly dictionary is built to identify more entities. The proposed concept is built based on the human-in-the-loop concept. Various examinations were planned and conducted to prepare customized NER models. The proposed model is built from base models [Spacy (2021)] by applying TL based models. The base model that was used for this work are as follows: Spacy and SciSpacy.

3.1. *TLBioNER*

TLBioNER (Transfer Learning based Bio NER) model is used to identify the medical entities such as drugs, dosages, date and diseases. The proposed research work is represented in the figure 1. The principal segment includes Python contents to get and order clinical notes text from the PubMed website. The proposed research work has developed the top of the newly built dictionaries. The newly developed dictionaries are processed by the spacy string matcher and newly annotated sentences are made.
The PSL (Python-Spacy-Language) annotation tool is built using the “Python-Flask” based REST API as the back end [Flask (2021)]. Explanation apparatus measures archive comments and yields clarified information in the configuration needed to prepare transfer learning models using spacy. Python module that uses the spacy pre-trained model which highlighted for language model. Another Python module which fabricates the existing blank

spacy model and execute over the clarified information. The model which is built on the top of the blank model utilized the TL concept and produced a new model. The algorithm for the proposed research work is given below.

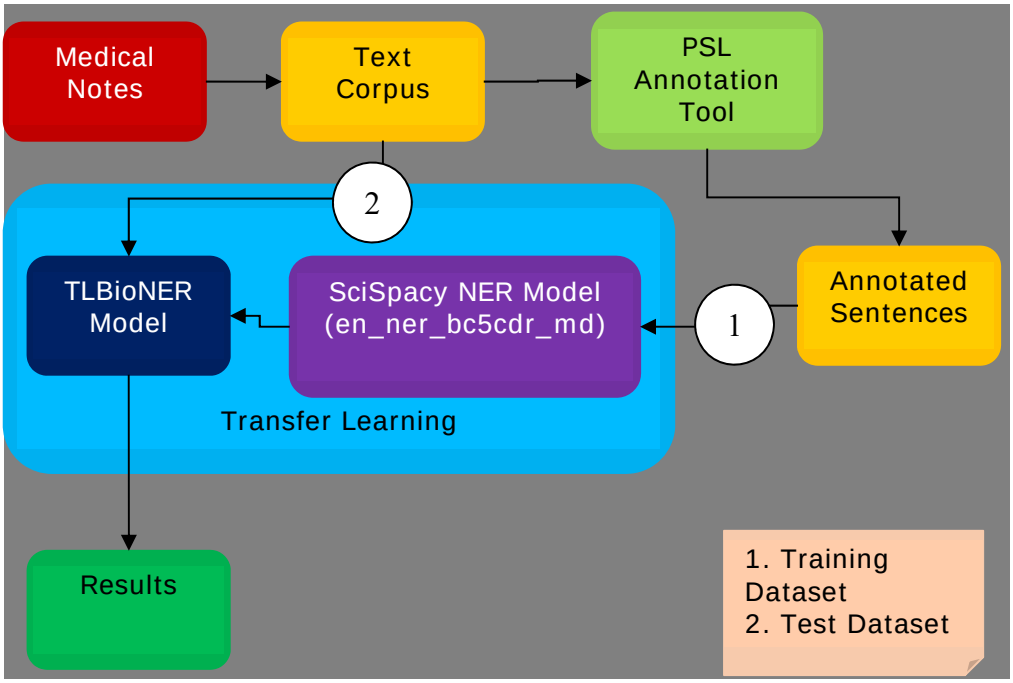

Fig. 1. Proposed Research Flow

Algorithm 1: TLBioNER
Input: Annotated Sentences
Output: Medical entities

1: Load en_ner_bc5cdr_md model
2: Set the named entities NE
3: Resume the NLP training
4: Load the annotated sentences
5: Set epoch ← 95
6: for sentences, annotations in training data:
7: set batch = [ ]
8: append new sentences into batch
9: calculate loss using the gradient decent function
10: update the model
11: save model to disk

3.2. *Preparation of Datasets*

For this research work, it made a highly effective computational corpus by examining the openly accessible example clinical information. The publicly available medication evaluation reports from "PubMed website" [Ramachandran R and Arutchelvan K (2021)] are used for preparing the datasets. The dictionaries are collected for the following entities:

- Drugs
- Diseases
- Dosages
- Date

The crawled documents are then preprocessed and converted into plain texts. The pypdminer library is used to convert the PDF documents into plain text documents. The elaborate numerical description of the taken datasets is given in the table 1.

This research work had used the "spacy phrase matcher" libraries to annotate the sentences. Initially it annotated 29774 sentences for entities individually. The detailed numerals of each entity that have been used is given in the table 1. The annotated sentences are then used to train the model using the spacy blank model.
The dataset taken for this research work was randomly split into 80% (training) and 20 % (testing). Basically, spacy follows a dataset model which contains sentences followed by the position of entities. It won't allow the sentences to be overlapped. The redundant sentences are removed from the data annotation process in order to reduce the computational time.

Table 1. Training and Testing Dataset

Dataset	Total Sentences	Entity Counts			
		Drugs	Disease	Dosage	Date
Annotated Sentences	29774	6157	8391	1580	4412
Train Data	23620	4726	6513	1064	3330
Test Data	5154	431	878	516	82

4. Results and Discussions

For working with scientific NLP based problems, spacy has provided scispacy models. "en_ner_bc5cdr_md" is a model which has been trained with medium number of datasets. This model identifies the named entities that are related to the medical terms. The prepared annotated datasets are trained over this model by retraining the pretrained model. The NLP model is resumed from the beginning. This makes to add the new entities with the existing entities. For this research work, the model is trained with 95 epochs. The dropout rate considered for this model is 2%. based on the dropout rate the model loss is calculated. The following are the system configuration that has been used for this research work. Ubuntu 20.04 LTS Operating System; Intel Core i5 Processor 10th generation; 16GB RAM; 16GB Optane Memory; Python 3.8; Spacy 2.0 for NLP processing; Beautiful Soup 4 for web crawling

4.1. *Evaluation Metrics*

The most used evaluation metric like *"precision, recall and F-score"* are applied to analyze the proposed model.
Precision: It is the fraction of correctly classified positive label (i.e., True Positive (TP)) with the all-positive labels (TP, False Positive (FP)) obtained from the datasets. The equation is given in the equation (1).

$$P = \frac{tp}{tp+fp} \quad (1)$$

Recall: It is the fraction of TP with the correctly classified labels and incorrectly classified labels (False Negative (FN)). The equation is given in the equation 2.

$$R = \frac{tp}{tp+fn} \quad (2)$$

F-score: The harmonic means of equation (1) and (2) is given in the equation 3.

$$F = 2 * \left(\frac{precision*recall}{precision+recall}\right) \quad (3)$$

The overall performance for the proposed work is calculated and the results are provided in the table 2.
The scispacy model which has been pre-trained with the "en_ner_bc5cdr_md" model that is retrained with the newly built dataset produced better results than the existing baseline model. The proposed model is added on the existing scispacy model and resulted decent F1 score. It is observed that overall percentage for all the entities have around 4 percentage of increase than the baseline model. The custom model which has been built for this research work produced 89.28 percentage as F1 score for the drug entities and leaded as better model than the pre-trained model. Figure 2,3,4 and 5 represents the bar chart of the comparison for entities drug, disease, dosage and date respectively.

Table 2. Comparative Analysis of Models

Entity	Metrics	Baseline	TLBioNER
Drugs	Precision	78.86	81.33
	Recall	81.57	83.82
	F-Score	80.19	82.55
Diseases	Precision	89.21	88.10
	Recall	84.45	90.49
	F-Score	86.76	89.28
Dosage	Precision	85.10	86.41
	Recall	80.80	88.31
	F-Score	82.90	87.34
Date	Precision	69.42	71.39
	Recall	75.99	79.06
	F-Score	72.55	75.03

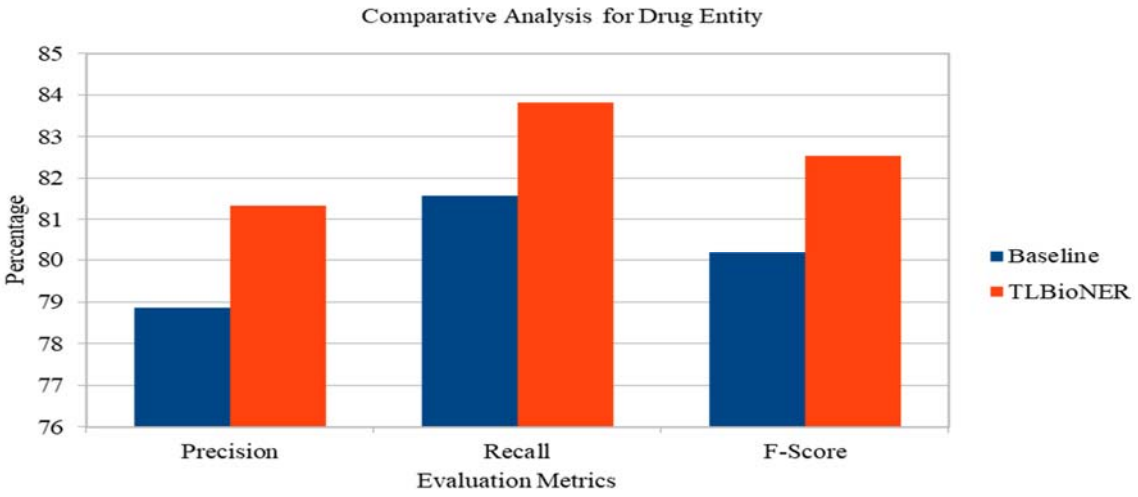

Fig 2. Comparative Analysis of Drug Entity

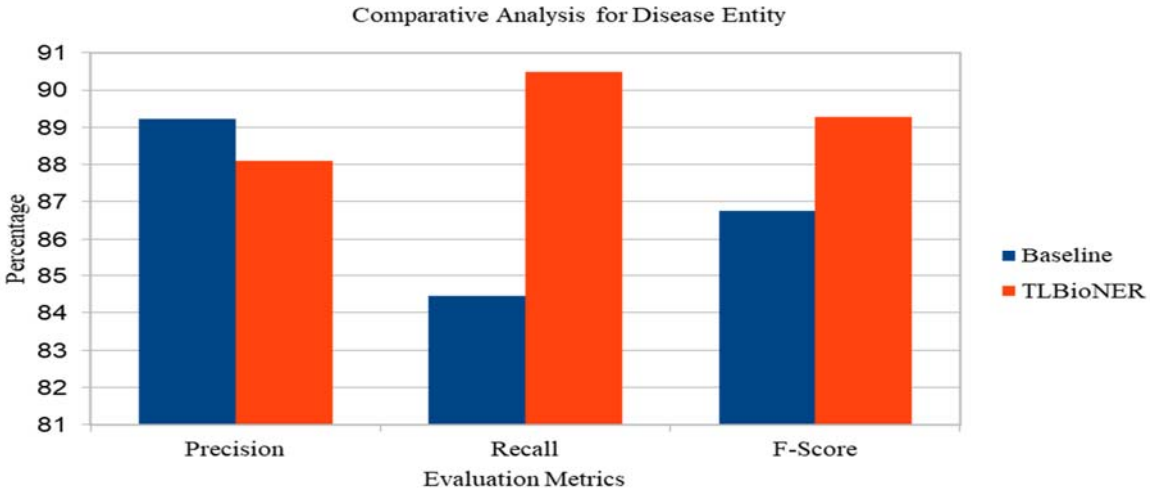

Fig 3. Comparative Analysis of Disease Entity

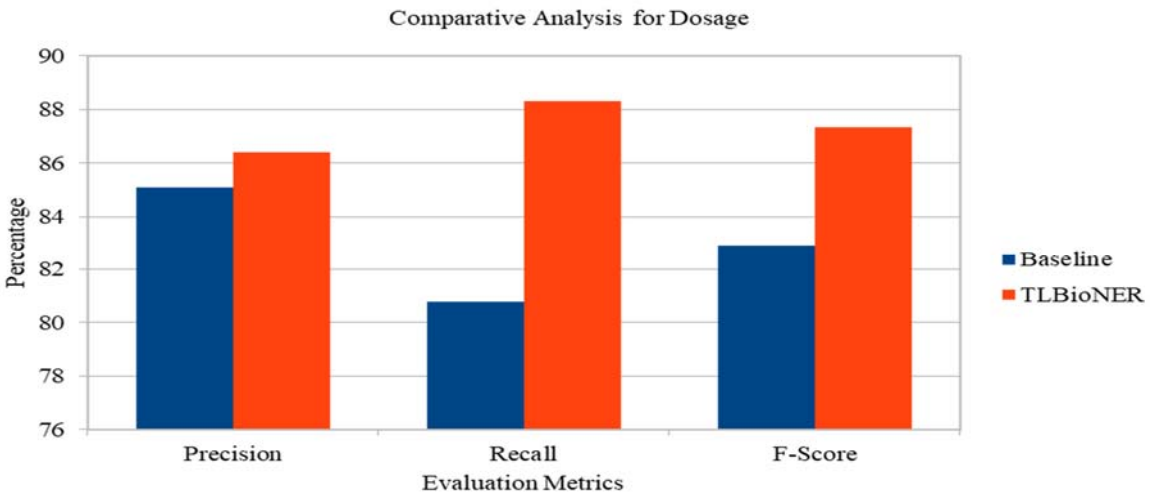

Fig 4. Comparative Analysis of Dosage Entity

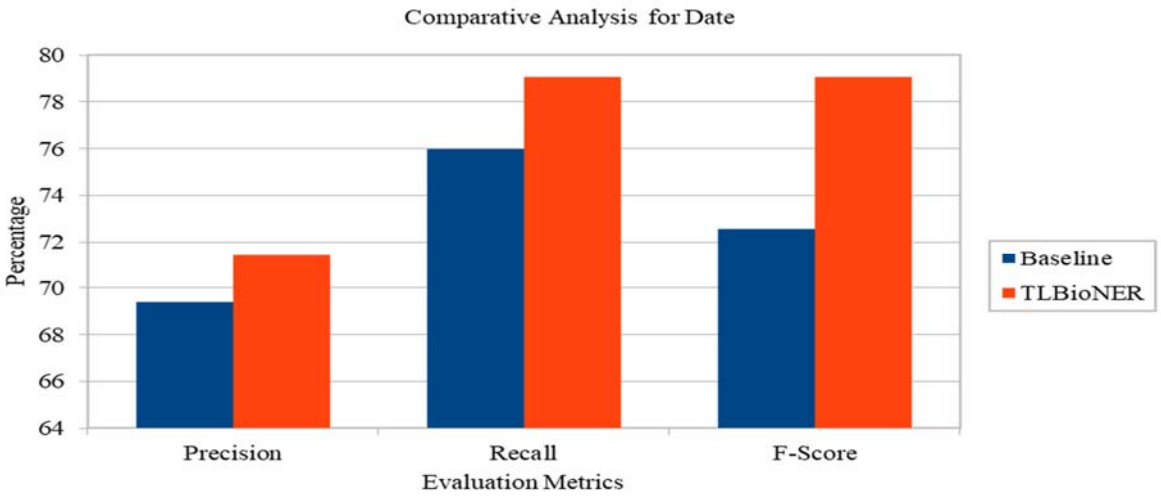

Fig 5. Comparative Analysis of Date Entity

5. Conclusion

Named entity recognition in the medical domain is playing a major role in identifying the important terms in the medical related textual documents. The proposed research work aims to identify the four important entities such as diseases, symptoms, drugs and dosages. NLP play the major role in extracting these entities. The proposed model is tested on the spacy NER. This research work discussed the importance of transfer learning (TL). TL is the process which makes the learning model smoothly with the limited amount of training data. It is observed that the pre-trained model is retrained with the new annotated data and produced better accuracy for the new added entities. The proposed model outperforms well and produced 83.55 percentage of overall F-Score. The data has been extracted from the medical documents which are available in PubMed websites. The proposed model provided significant benefits even with the overlapping annotated entities. As the future enhancements, it is aimed to increase the number entities such route of administration, dosage levels, species, organs, etc. In future, the proposed model can be enhanced and to be used in applications such as dynamic chatbot, question-answering, toxicology report generation, etc.

References

[1] Ramachandran, R., Arutchelvan, K. Named entity recognition on bio-medical literature documents using hybrid-based approach. *J Ambient Intell Human Comput* (2021). https://doi.org/10.1007/s12652-021-03078-z
[2] Yao Chen, Changjiang Zhou, Tianxin Li, Hong Wu, Xia Zhao, Kai Ye, Jun Liao, Named entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training, Journal of Biomedical Informatics, Volume 96, 2019, 103252, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2019.103252
[3] X. Dong, L. Qian, Y. Guan, L. Huang, Q. Yu, J. Yang, A multiclass classification method based on deep learning for named entity recognition in electronic medical records Sci. Data Summit (2016), pp. 1-10, 10.1109/NYSDS.2016.7747810
[4] Sara Santiso, Alicia Pérez, Arantza Casillas, Adverse Drug Reaction extraction: Tolerance to entity recognition errors and sub-domain variants, Computer Methods and Programs in Biomedicine, Volume 199, 2021, 105891, ISSN 0169-2607, https://doi.org/10.1016/j.cmpb.2020.105891.
[5] Seonghyeon Moon, A Gitaek Lee, A Seokho Chi, A Hyunchul Oh, Automated Construction Specification Review with Named Entity Recognition Using Natural Language Processing, 2021, J Journal of Construction Engineering and Management, V 147, doi:10.1061/(ASCE)CO.1943-7862.0001953
[6] Kumar, A., Starly, B. "FabNER": information extraction from manufacturing process science domain literature using named entity recognition. *J Intell Manuf* (2021). https://doi.org/10.1007/s10845-021-01807-x
[7] Payal Biswas, Aditi Sharan, A Noble Approach for Recognition and Classification of Agricultural Named Entities using Word2Vec, International Journal of Advanced Studies In Computer Science & Engineering IJASCSE Volume 9 ISSUE 12, 2021
[8] Ning, Gelin and Bai, Yunli. 'Biomedical Named Entity Recognition Based on Glove-BLSTM-CRF Model'. Journal of Computational Methods in Sciences and Engineering, vol. 21, no. 1, 1 Jan. 2021: 125 – 133
[9] Ravikumar J. and, Ramakanth Kumar P, Machine learning model for clinical named entity recognition, International Journal of Electrical and Computer Engineering (IJECE) Vol. 11, No. 2, April 2021, pp. 1689~1696 ISSN: 2088-8708, DOI: 10.11591/ijece. v11i2.pp1689-1696
[10] Sajid Hussain, Hammad Afzal, Ramsha Saeed, Naima Iltaf, Mir Yasir Umair, "Pharmacovigilance with Transformers: A Framework to Detect Adverse Drug Reactions Using BERT Fine-Tuned with FARM", *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 5589829, 12 pages, 2021. https://doi.org/10.1155/2021/5589829
[11] spaCy https://spacy.io Accessed on September-07-2021
[12] Flask, The Python micro framework for building web applications https://palletsprojects.com/p/flask/ Accessed on September-07-2021

Authors Profile

Dr.K.Arutchelvan is working as Assistant Professor in the Department of Computer and Information Science at Annamalai University since the year of 2002. He has received his Master's degree from Bharathidasan University, Tiruchirapalli and Ph.D degree from Bharathidasan University, Tiruchirapalli in the year of 2018. His research areas are Data Mining, Wireless Sensor Network, Named Entity Recognition, Natural Language Processing, Big Data Analytics etc. He has published ten research papers on the above topics in various International Journals and Conferences.

R.Ramachandran is working as Assistant Professor in the Department of Computer and Information Science at Annamalai University, Chidambaram. He is pursuing his Ph.D degree in Computer Science at Annamalai University, Chidambaram. His research areas are Big Data Analytics, Natural Language Processing, Named Entity Recognition, etc.