How to Cite:
Dash, A., Pandey, M., & Rautaray, S. S. (2022). A novel deep neural network framework for biomedical named entity recognition. *International Journal of Health Sciences*, *6*(S5), 4310–4324. https://doi.org/10.53730/ijhs.v6nS5.9557

A novel deep neural network framework for biomedical named entity recognition

Adyasha Dash
School of Computer Engineering, KIIT Deemed to be University, India, 751024

Dr. Manjusha Pandey
School of Computer Engineering, KIIT Deemed to be University, India, 751024

Dr. Siddharth Swarup Rautaray
School of Computer Engineering, KIIT Deemed to be University, India, 751024

***Abstract*---**With the dramatic improvements in the field of bio-informatics, extracting information from text and analyzing the association between the entities has received more attention in the past few years. Entity Recognition (ER) meant to extract and recognize the entities from any text. Biomedical Named Entity Recognition (BNER) gets more and more attention from the researchers since it is a fundamental task in biomedical information extraction. Various methods has been proposed to perform the task of BioNER. Different kind of approaches are dictionary based, rule based approaches, traditional machine learning approaches that combines supervised and unsupervised methods and neural network based approach. The state-of-art systems previously adopted various supervised machine learning methods Hidden Markov Models (HMMs), Maximum Entropy Markov Models (MEMMs), Support vector machines(SVM),Structural Support Vector Machines(SSVMS),Conditional Random Fields(CRF) to derive semantic and syntatic features from annotated datasets. However, CRF is one of the most successful method used for NER and it has obtained finest result because of the robustness and ability for sequence lebelling task. Recently, studies have demonstrated the application of deep learning based approaches for biomedical named entity recognition (BioNER) and shown promising results. Deep learning (DL) has become conventional across various ML based NLP tasks and helps in identifying prominent features by eliminating the need of task-specific feature engineering based on high-level domain knowledge. Long Short-Term Memory (LSTM) is most widely used neural network. In this paper, we have proposed a novel deep neural networks combined with conditional random field to extract biomedical named entities in biomedical texts. Our model gets state-of-art results outperforms the existing system on publicly available

datasets. We have analyzed the impact and effectiveness of different embedding schemes (word and character-level based embedding) in order to improve the performance of our system.

***Keywords*---**Information Extraction, Biomedical Named Entity Recognition, Deep Neural Networks, Natural language Processing, Word Embedding, LSTM, CRF.

Introduction

A great diversity of massive medical documents are getting generated from biomedical domain these days with billions of essential information hidden inside it. The voluminous data are predominantly in textual form.Some significant techniques have been proposed by researchers for extracting the essence and semantic meaning of information with the application of text mining. Information extraction, clustering, sentiment analysis, opinion mining, text summarization, clustering, visualization, deep learning are different approaches for producing precise and meaningful content of text. Information extraction(IE) and Natural language processing(NLP) has become most popular research field in deriving patterns from non structured resources and can bring tremendous benefits in the biomedical field.IE converts an unstructured medical report to structured format so that the information can be better analyzed,aggregated and mined for insightful patterns.Information Extraction (IE) is the most widely used text mining technique to extract specific, structured information from unstructured and semi structured text. With the substantial improvements in the field of bio informatics,finding information from text and analyzing the association among the entities has received much more attention in the last few decades.IE helps in deriving meaningful pieces of knowledge from natural language text.Information extraction is a crucial method for analysis of data.Information extraction in medical domain performs the task of identifying medical related terms,identifying the relationships among the medical entities and mapping the entities in the medical document to accurate domain specific ontologies.There is an explosion of information in medical records which is difficult for reasoning and interpretation.Natural language processing(NLP) makes use of computational methods to process the free unstructured text.NLP has a significant role in bio medical information extraction.The basic NLP task involves the processing of medical data in syntactic level as well as semantic level.syntactic level focuses on word segmentation,tokenization and POS tagging while semantic level is done to get the meaningful representation of the entity in context level.There are various sub tasks involved in IE namely:
Named Entity Recognition(NER), Named entity Linking(NEL), Co-reference Resolution(CR)Temporal Information Extraction, Relation Extraction(RE), Knowledge base construction and reasoning.

Table1.1 Sub-tasks of IE and functionality

SUB-TASKS	FUNCTIONALITY
Named Entity Recognition(NER)	Systems are required to recognize the

	Named Entities occurring in the text. More specifically, the task is to find Person (PER), Organization (ORG), Location (LOC) and Geo-Political Entities (GPE)
Named entity Linking(NEL)	Named Entity Linking (NEL) also known as Named Entity Disambiguation (NED) or Named Entity Normalization (NEN) is the task of identifying the entity that corresponds to particular occurrence of a noun in a text document
Co-reference Resolution(CR)	Co-reference Resolution is the task which determines which noun phrases (including pronouns, proper names and common names) refer to the same entities in documents
Temporal Information Extraction	Temporal information extraction or event extraction refers to the task of identifying events (i.e information which can be ordered in a temporal order)in free text and deriving detailed and structured information about them, ideally identifying who did what to whom, where, when and why
Relation Extraction(RE)	Relation Extraction is the task of detecting and classifying predefined relationships between entities identified in the text

2. Application of Named Entity Recognition in Biomedical domain

Named entity recognition(NER) is a sub-task of NLP,the purpose of identifying named entities based in articles into predefined classes.A typical Bio-NER system is shown in the figure 1.3 and taken as example of a sentence from a research paper and has shown different phases of Bio-NER system.BioNER system contains two fundamental stages:Named entity boundary detection and NE type classification.The named entity boundary detection phase determines whether a single token or several adjacent tokens represent an NE without considering the type of NE, thus distinguishing NEs from non-NEs.The named entity type classification is a special type of muti-class classification problem of assigning a specific class label from the predefined set of labels to each named entity of labels to each NE identified in the boundary detection phase.

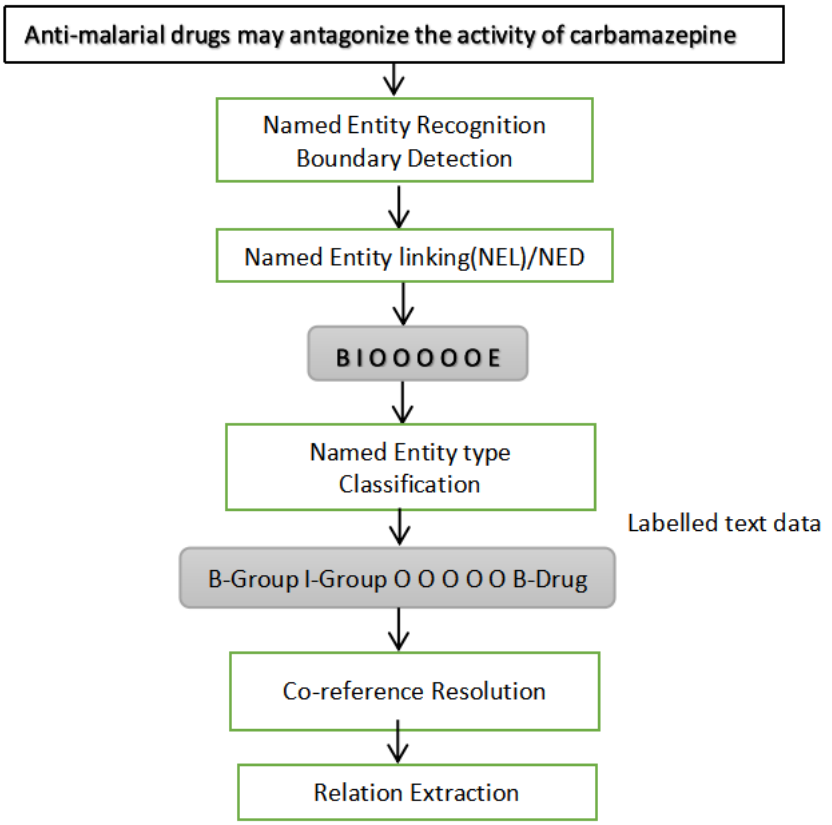


Figure 1.1 A typical Bio-NER workflow

2.1 Techniques and Algorithms to solve NER Problem

Over the year many approaches have been explored by researchers inorder to develop NER system.The approaches to NER fall into four different categories.

2.1.1 Rule based and dictionary based Methods

Rule based and dictionary based methods are the traditional methods.Rule based method rely on handcrafted rules,use named entity libraries and assigns weights to rules where as in dictionary based approach BioNEs in a text are identified by looking up a provided list of known BioNEs in the form of a dictionary and gazetteer,typically compiled from existing lexicons or databases.Some well manged available databases and lexicons for Biomedical named entities are *ChemIDplus*5, *GenBank*6 for genes and *UniProt*7 for proteins.Rule based and dictionary based approach mostly depended on specific language.Most lexicons used in dictionary based approach are constructed by experts in a specific domain and will usually have restricted coverage in terms of size, due to the growth of new BioNEs. As the dictionaries are domain specific, dictionary based approaches are not portable to other domains.Dictionary based method followed string matching algorithm to extract biomedical named entities.Major limitations of rule base techniques is it demands domain speciality,linguistic experts to

rewrite the rules for specific language,domain and text.The compilation of rules are time consuming.

2. 1.2 Machine learning based Methods

Machine learning based approach makes use of a function or classification rule or classifier to detect the biomedical named entity boundaries.It classifies the named entities into one of the predefined classes. A classification rule can be defined as f: p→Y(task of evaluating the label Y of k-dimensional input vector p)where P∈R^k(input always taken as real valued vectors. y∈Y={c1,c2,c3….c_n}.The classification will be called as binary classification for n=2 and for n>2 it is called as multi-class classification.Several classifiers such as Naive Bayes, *k*-Nearsest Neighbors, SVM, Hidden Markov Models (HMM) ,CRF have been used for different applications.

Mathematically, Bio-NER system can be defined as a classifier that takes a sentence S as input having n words in a sequence expressed as $S=\{w1,w2,w3....w_n\}$ and can be assigned to one of the predefined labels $\{c1,c2,c3....c_n\}$to each word or a sequence of words based on the characteristics of the words captured during the training phase. The classifier model is built by features extracted from the training data which is expressed as $T = \{<w_1, y_1>,<w_2, y_2>,... < w_m, y_m>\}$.where w_i is the word and y_i is the corresponding tag of that word.

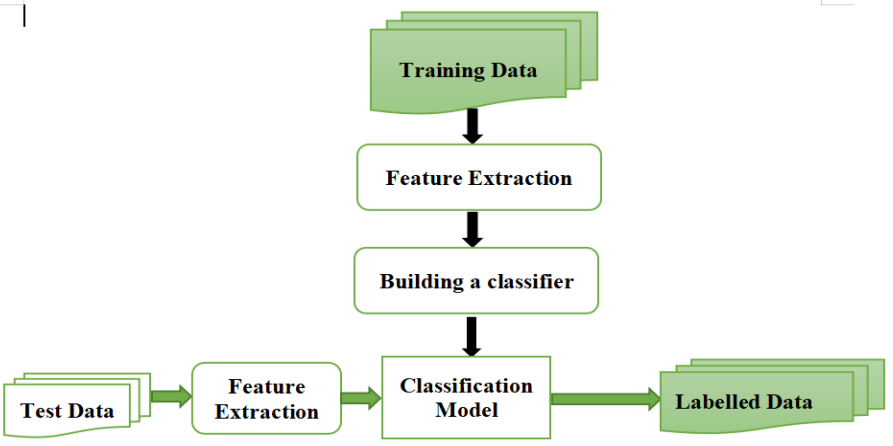


Figure 2.1 Machine learning framework for classification

3. Proposed Deep Learning based NER Model

In the last few years, Deep learning(DL) based model has gained popularity due to their unexceptional success at overcoming complex learning problem. Deep learning based model has been applied in biomedical domain for various Natural Language Processing task like POS tagging and ML.In this chapter a novel Deep learning model for biomedical NER has been proposed using dictionary information.Our proposed model is evaluated on BC2GM,NCBI and JNLPBA data

set.The input to our model is word embedding trained over general domain corpus along with biomedical corpus.Our study proposed IOBES tagging scheme for biomedical disease NER.Our NER system use the dictionary information along with word embedding and character embedding as the features along with Conditional Random field(CRF) to improve the performance of Bio-NER system.

3.1 NER Modeling: Sentence Level

To improve the accuracy of entity detection, we aim to combine the TF-IDF features with word vectors obtained from Word2vec,TF-IDF works on bag-of-words (BOW) model and gives most relevant entities by capturing the semantic relationship. 2) Word2vec model helps to maintain the syntactic and semantic relationship among word

$$TF–IDF=c(\Gamma,d)/|d|\times\log|D|/|\{d\in D:\Gamma\in d\}| \qquad (1)$$

where, $c(\Gamma,d)$ is the count that the term Γ occurs in document d, $|d|$ is the length of document d, $|D|$ is total count of documents in document collection and $|\{d\in D:\Gamma\in d\}|$ denotes total count of documents in which term Γ occurs. TF-IDF weighting is applied to each word in a document and weights are averaged across the document. Therefore, as an output of the pre-processing step, we get the joined TF-IDF word vectors.

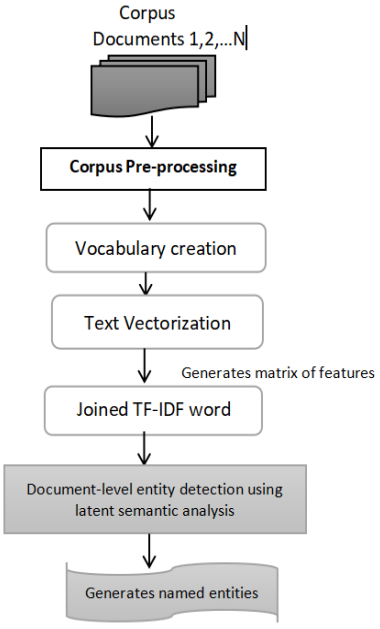


Figure 3.1 Machine learning framework for classification

3.1.1 TF-IDF Model for weight calculation for textual data
Bow Algorithm for word occurrence calculation

Input:-Corpus of documents or collection of documents. {docA, docB…}

Step 1 **Tokenized BOW**:Tokenized Bag of words model to represent a document in tokenized form.

BowA= docA.split (“ ”)

BowB= docB.split (“ ”)

BowA= {‘t1’, ‘t2’,’t3’,’t4’……..}

where t1, t2, t3=split terms

Step 2 Converting the tokenized BOW into numbers by creating a vector of all possible words for each document.

Step 3 Calculating the word count appears for each document.

Wordset=set(bowA).union(set(bowB))

Step 4 Generation of unique words in the corpus.

Step 5 Creation of dictionaries one for each bag-of-words.

{WordDictA, WordDictB}

Output:- Forming matrix for representation of occurrences

TF-IDF is a better strategy for counting words in weighing the counts appropriately through TF-IDF score to rank the importance of the word.

Step 1 Calculating the term frequencies for bowA and bowB i.e. tfbowA & tfbowB.

Step 2 Calculating the document frequency.

Step 3 Computing Inverse document frequency for number of occurrences of words across all the documents in the corpus.

N= len(document list)

Step 4 Counting the number of documents that contains word w.

Step 5 Generation of TF-IDF score of word w.

In the Bow model the common words do not provide much differentiation although they are used in different sense of meaning in the documents but in TF-IDF model, TF-IDF score for the common words that appears equally across number of documents is found to be zero which clearly states that words are not featured words.TF-IDF score are much accurate for the featured words that are semantically important.For relevant medical term generation this approach can be used.

3.2 Overview of Proposed Methodology

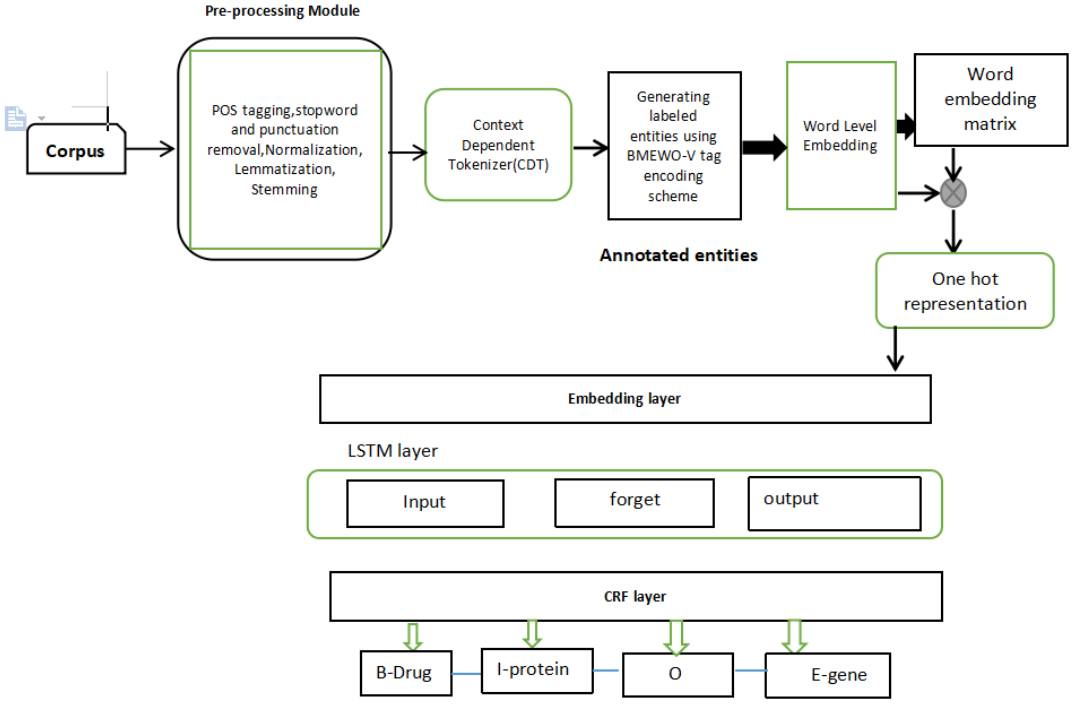


Figure 3.2 Overall Architecture of the Proposed Model

Algorithm for Context Dependent Tokenizer (CDT)

Algorithm 1: Algorithm for Context dependent Tokenizer(CDT)

```
S = Input corpus file
Read S
LVw[tags]list of tags                 //LVw contains list of words in vocabulary
LWi[tags]list of words                //LWi contains actual input file
for every word in LWi
    If tag is not equal to zero
        If   LWi are unique and available in LVi
                Assign token to the words;
        elseif LWi not available in LVi
            Split word into sub words and assign token T
        else
        set type property of token=NULL
return token
```

BMEWO-V extended tag encoding

Tags	Meaning
B	Beginning of an entity
M	Middle of any entity/continuity of entity
E	End of an entity
W	Single token entity
V	Two or more entities overlap in that token
O	Tokens does not represent anything

Embedding Techniques

Process of converting the text data into numerical data, which is called Vectorization or in the NLP world, it is called word embedding. Vectorization or word embedding helps in extracting features from text to build multiple natural language processing models.The Continuous Bag-of-Words model (CBOW), Glove are frequently used in NLP deep learning. It is a model that tries to predict words given the context of a few words before and a few words after the target word.

Steps for Word embedding:one-hot representation

Steps for word representation:**one-hot representation for LSTM**

1. Input: Sentence S
2. Initialize parameter{|V|= vocabulary size,dimension_size}
3. Generate I_v = one-hotf(W_i,V)
4. Produce one-hot word representation [S1[I_{wi}, I_{wi+1},I_{wi+2}],.....S_n] //I_{wi} =index value in vocabulary
5. Pass vector into embedding layer
6. Generate embedding matrix
7. Evaluate max_len=Sentence_length
8. Adding padding to generate embedded_doc
9. Output-featurized vector

Dimension size=no of features

Mathematical representation

Input layer:sequence of vectors X{x1,x2,.....,xn}
return sequence of LSTM:(h1,h2,.....hn)
In general,to update an LSTM unit at time t,following formulas are used:

$i_t = \sigma(W_i h_{t-1} + U_i x_i + b_i)$ **(1)**

$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f)$ **(2)**

$c_t = \tanh(W_c h_{t-1} + U_c x_t + b_c)$ **(3)**

$c_t = f_t * c_{t-1} + i_t * c_t$ **(4)**

$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o)$ **(5)**

$h_t = o_t * \tanh(c_t)$ **(6)**

for a given sentence (x1,x2,.....xn) in the corpus consisting n words final output of the word computed

ht = [$ht_{backword}$, $ht_{forward}$]

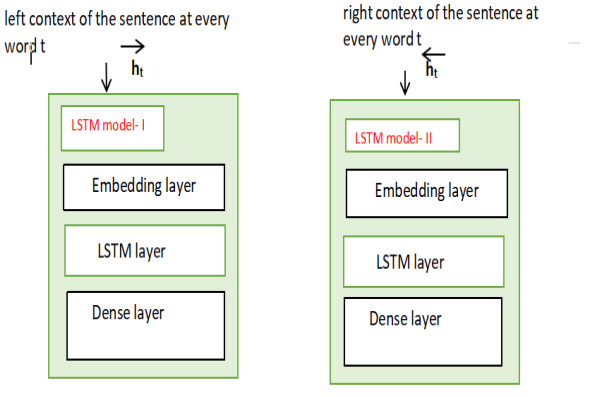


Figure 3.3 LSTM model with embedding layer

LSTM-CRF layer

For sequence labeling tasks, for a given sentence, it is useful to consider the correlations between labels in neighborhoods and jointly decode the best chain of them.Instead of decoding each label independently it can be modeled jointly using a conditional random field (CRF).The model consists of entity mapping and entity boundary classification.It produces state transition matrix to predict the current tag.

Transition matrix by Ti,j representing the transition score from the i-th tag to the j-th tag.For a given sentence X = {x1,x2,x3....xn}, we compute the matrix of score output by LSTM network for sentence [X].The sum of the scores from LSTM network along with the transition scores gives the final score for a sentence X

$$s([X]_1^T, [i]_1^T) = \sum_{t=1}^{T} (T_{[i]_{t-1},[i]_t} + M([S]_1^T)_{[i]_t,t})$$

Statistics of Dataset

Biomedical NER corpus statistics for entity detection

Name of data set	Entity type	Dataset size
BC2GM	Gene/Protein	2000 sentences
JNLPBA	Gene/Protein, DNA, Cell Type, Cell Line, RNA	2,404 abstracts
NCBI-disease	Disease	793 abstracts
BC5CDR	Chemical, Disease	1,500 articles

Phase	Dataset
Training	(BC2GM)18,546 sentences
Testing	NCBI-disease
	JNLPBA

Evaluation metrics

The performance of the model is evaluated using standard parameters precision(P),recall(R), and F1 score(F)

$$P = \frac{TP}{TP+FP}$$

$$R = \frac{TP}{TP+FN}$$

$$F1 = \frac{2*P*R}{P+R}$$

TP=number of correct entity identified
FP=number of incorrect entities returned
FN=number of missing entities

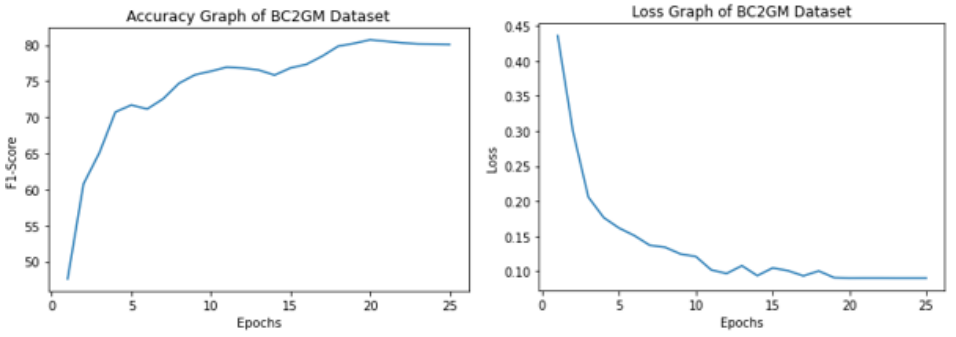
In context of biomedical NER,precision is the percentage of biomedical named entities identified by system that is correct recall=percentage of relevant entities that have been retrieved over total amount of relevant entitiesF1 score=measure of tests's accuracy.

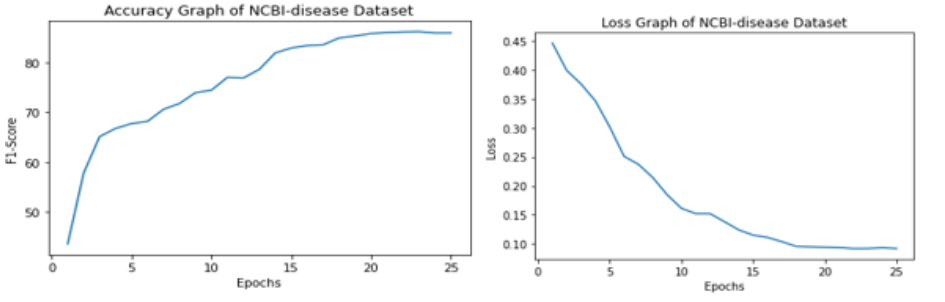
Performance evaluation

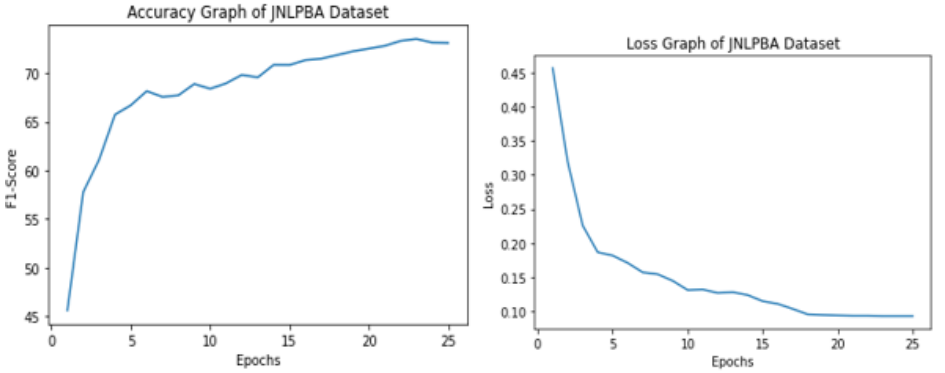
Results obtained using proposed model for NER detection and classification on BC2GM data set

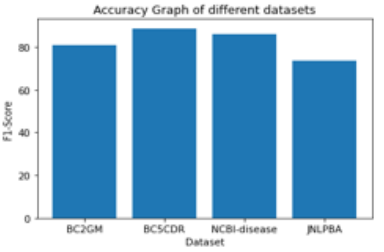
Name of data set	P_n	R_n	$F1_n$
BC2GM	87.24	87.15	87.19
NCBI	86.2	88.2	87.2
JNLPBA	84.7	84.2	85.3

Entity type	P_n	R_n	$F1_n$
gene	80.73	89.22	84.7
protein	84.6	72.8	78.3









Conclusion and Future Work

In this paper, we have proposed deep learning based NER model by integrating word embedding representation. we evaluated our model worked on 3 benchmark datasets ,BC2GM data set achieved a maximum F1 score of 87.19. .The proposed approach is suitable for named entity identification and classification and achieved state-of-art performance in two datasets namely NCBI and JNLPBA.Our model is based on deep neural networks that gave us ability to build a Biomedical

NER system without using any dictionary or gazetteers.e eliminate the need for most hand feature engineering task still suffer out-of-vocabulary(OOV) problem.As a future work,we will combine few additional module: Character level embedding and relation classification to improve the performance of the system.

References

[1] Moreno, Antonio, and Teófilo Redondo. "Text analytics: the convergence of big data and artificial intelligence." IJIMAI 3.6 (2016): 57-64.

[2] Gomaa, Wael H., and Aly A. Fahmy. "A survey of text similarity approaches." International Journal of Computer Applications68.13 (2013): 13-18.

[3] Sumathy, K. L., and M. Chidambaram. "Text mining: concepts, applications, tools and issues-an overview." International Journal of Computer Applications 80.4 (2013).

[4] Hobbs, Jerry R. "Information extraction from biomedical text." Journal of biomedical informatics 35.4 (2002): 260-264.

[5] Ju, Zhenfei, Jian Wang, and Fei Zhu. "Named entity recognition from biomedical text using SVM." 2011 5th international conference on bioinformatics and biomedical engineering. IEEE, 2011.

[6] Alshaikhdeeb, Basel, and Kamsuriah Ahmad. "Biomedical Named Entity Recognition: A Review." International Journal on Advanced Science, Engineering and Information Technology6.6 (2016): 889-895.

[7] Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification." Lingvisticae Investigationes30.1 (2007): 3-26.

[8] Mansouri, Alireza, Lilly Suriani Affendey, and Ali Mamat. "Named entity recognition approaches." International Journal of Computer Science and Network Security 8.2 (2008): 339-344.

[9] Tang, Buzhou, et al. "Evaluating word representation features in biomedical named entity recognition tasks." BioMed research international 2014 (2014).

[10] Huang, Cheng-Hui, Jian Yin, and Fang Hou. "A text similarity measurement combining word semantic information with TF-IDF method." Jisuanji Xuebao(Chinese Journal of Computers)34.5 (2011): 856-864.

[11] Sasaki, Yutaka, et al. "How to make the most of NE dictionaries in statistical NER." BMC bioinformatics 9.11 (2008): S5

[12] Harkema, Henk, et al. "Information extraction from clinical records." Proceedings of the 4th UK e-science all hands meeting. 2005.

[13] Suryasa, I. W., Rodríguez-Gámez, M., & Koldoris, T. (2022). Post-pandemic health and its sustainability: Educational situation. International Journal of Health Sciences, 6(1), i-v. https://doi.org/10.53730/ijhs.v6n1.5949

[14] Boulis, Constantinos, and Mari Ostendorf. "Text classification by augmenting the bag-of-words representation with redundancy-compensated bigrams." Proc. of the International Workshop in Feature Selection in Data Mining. Citeseer, 2005.

[15] Djuraev, A. M., Alpisbaev, K. S., & Tapilov, E. A. (2021). The choice of surgical tactics for the treatment of children with destructive pathological dislocation of the hip after hematogenous osteomyelitis. International

Journal of Health & Medical Sciences, 5(1), 15-20. https://doi.org/10.21744/ijhms.v5n1.1813
[16] Kumar, S. (2022). A quest for sustainium (sustainability Premium): review of sustainable bonds. Academy of Accounting and Financial Studies Journal, Vol. 26, no.2, pp. 1-18
[17] Allugunti V.R (2022). A machine learning model for skin disease classification using convolution neural network. International Journal of Computing, Programming and Database Management 3(1), 141-147
[18] Allugunti V.R (2022). Breast cancer detection based on thermographic images using machine learning and deep learning algorithms. International Journal of Engineering in Computer Science 4(1), 49-56
[19] Geetha, S., GS Anandha Mala, and N. Kanya. "A survey on information extraction using entity relation based methods." (2011): 882-885.