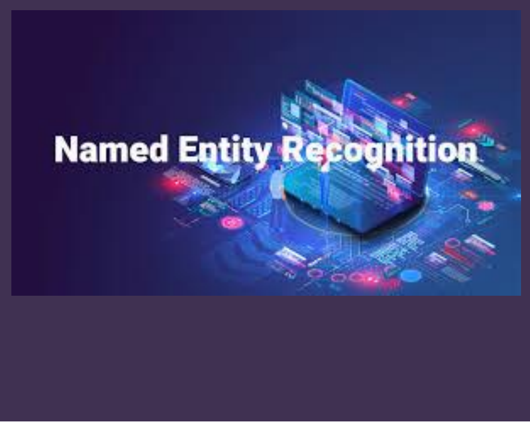
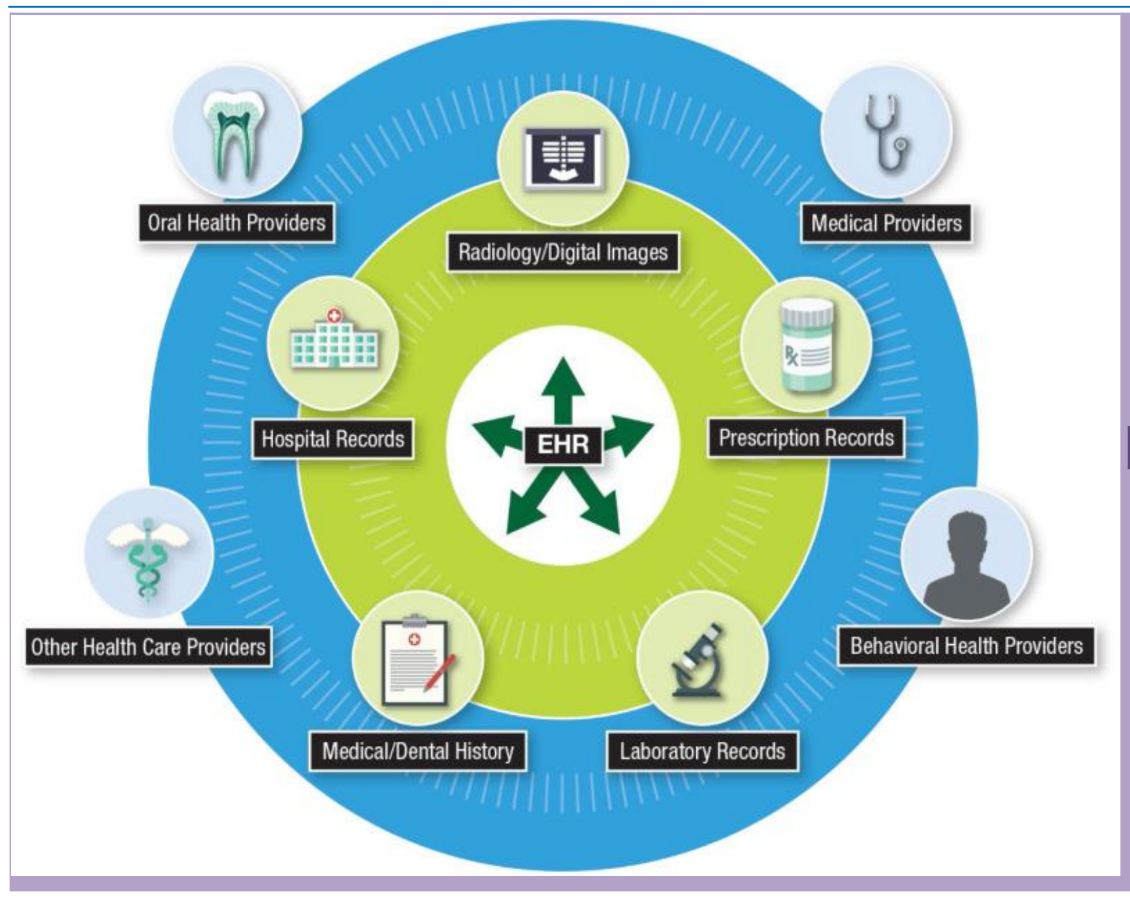
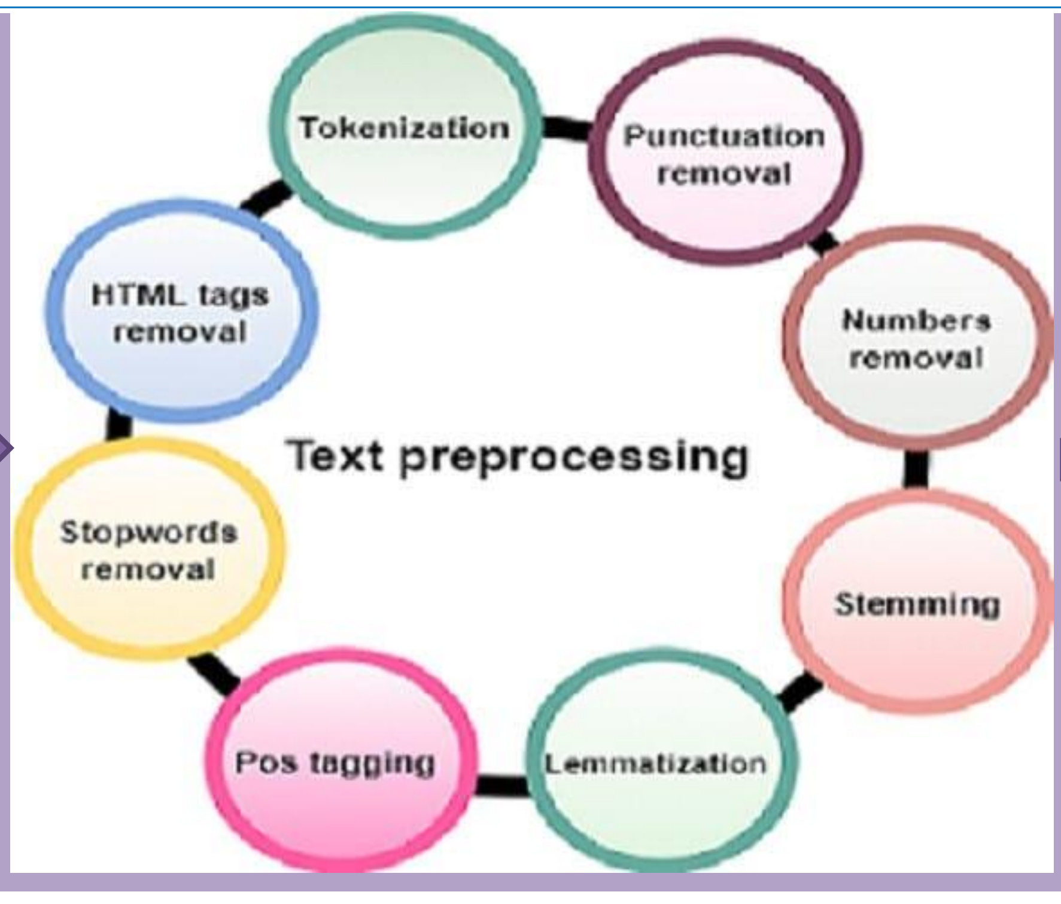
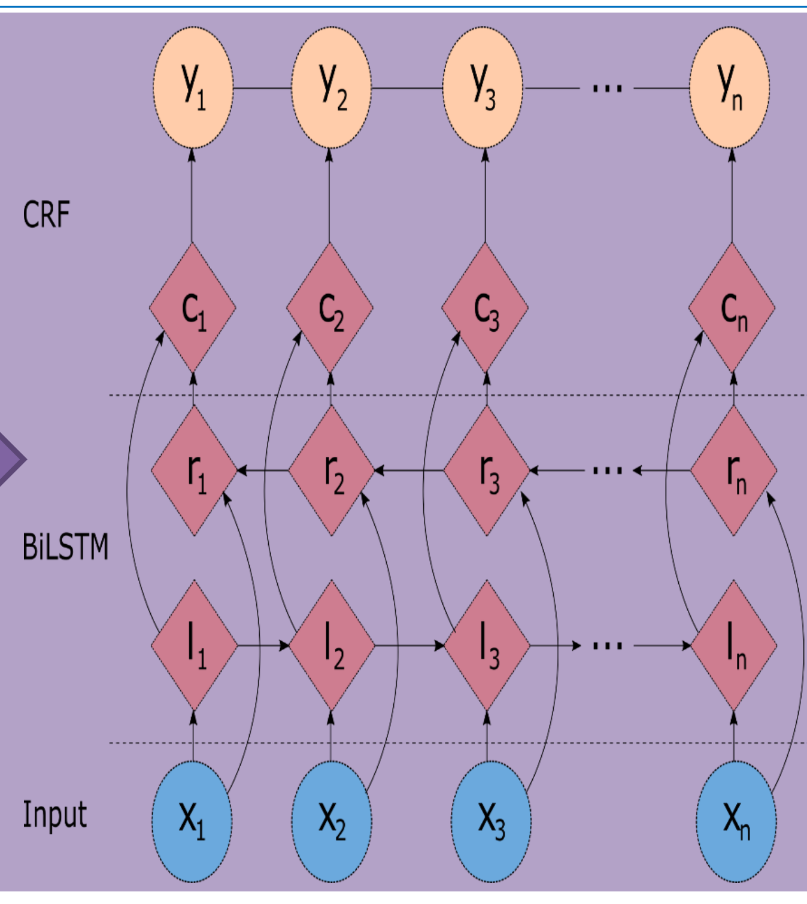
MSc in Data Science: Named Entity Recognition in Healthcare using Self-Supervised Learning.

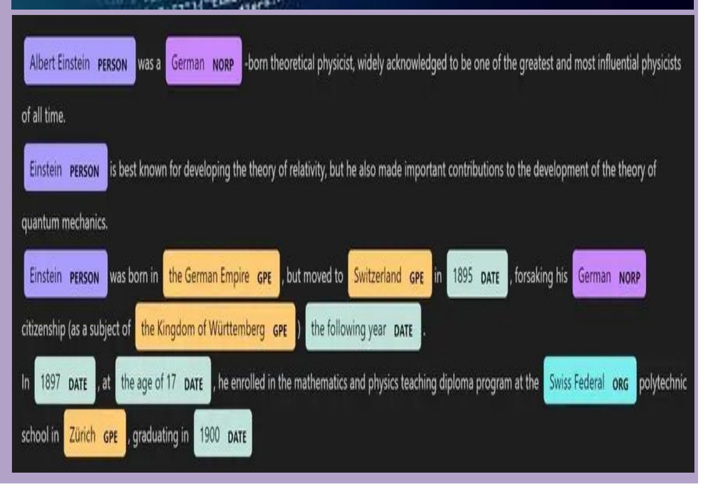


M.S.Shruthi ,Jason Barron | Department of Computing , South East Technological University Carlow, Ireland






CRF
BiLSTM
Input
y_1 y_2 y_3 … y_n
c_1 c_2 c_3 c_n
r_1 r_2 r_3 … r_n
l_1 l_2 l_3 … l_n
x_1 x_2 x_3 x_n



Albert Einstein PERSON was a German NORP -born theoretical physicist, widely acknowledged to be one of the greatest and most influential physicists of all time.

Einstein PERSON is best known for developing the theory of relativity, but he also made important contributions to the development of the theory of quantum mechanics.

Einstein PERSON was born in the German Empire GPE , but moved to Switzerland GPE in 1895 DATE , forsaking his German NORP citizenship (as a subject of the Kingdom of Württemberg GPE) the following year DATE .

In 1897 DATE , at the age of 17 DATE , he enrolled in the mathematics and physics teaching diploma program at the Swiss Federal ORG polytechnic school in Zürich GPE , graduating in 1900 DATE

1. Introduction

- Self Supervised Learning(SSL) is a type of machine learning where a Model is trained on unlabeled data allowing it to learn patterns and relationships without the need for labeled data. It can be used to analyze and extract useful information from clinical text data in healthcare.
- SSL can be used to train models to recognize patterns and relationships in the data, allowing for more accurate analysis and interpretation.
- Natural Language Processing (NLP) techniques are used to extract pertinent information from clinical text data, such as identifying diagnoses, prescriptions, and procedures.

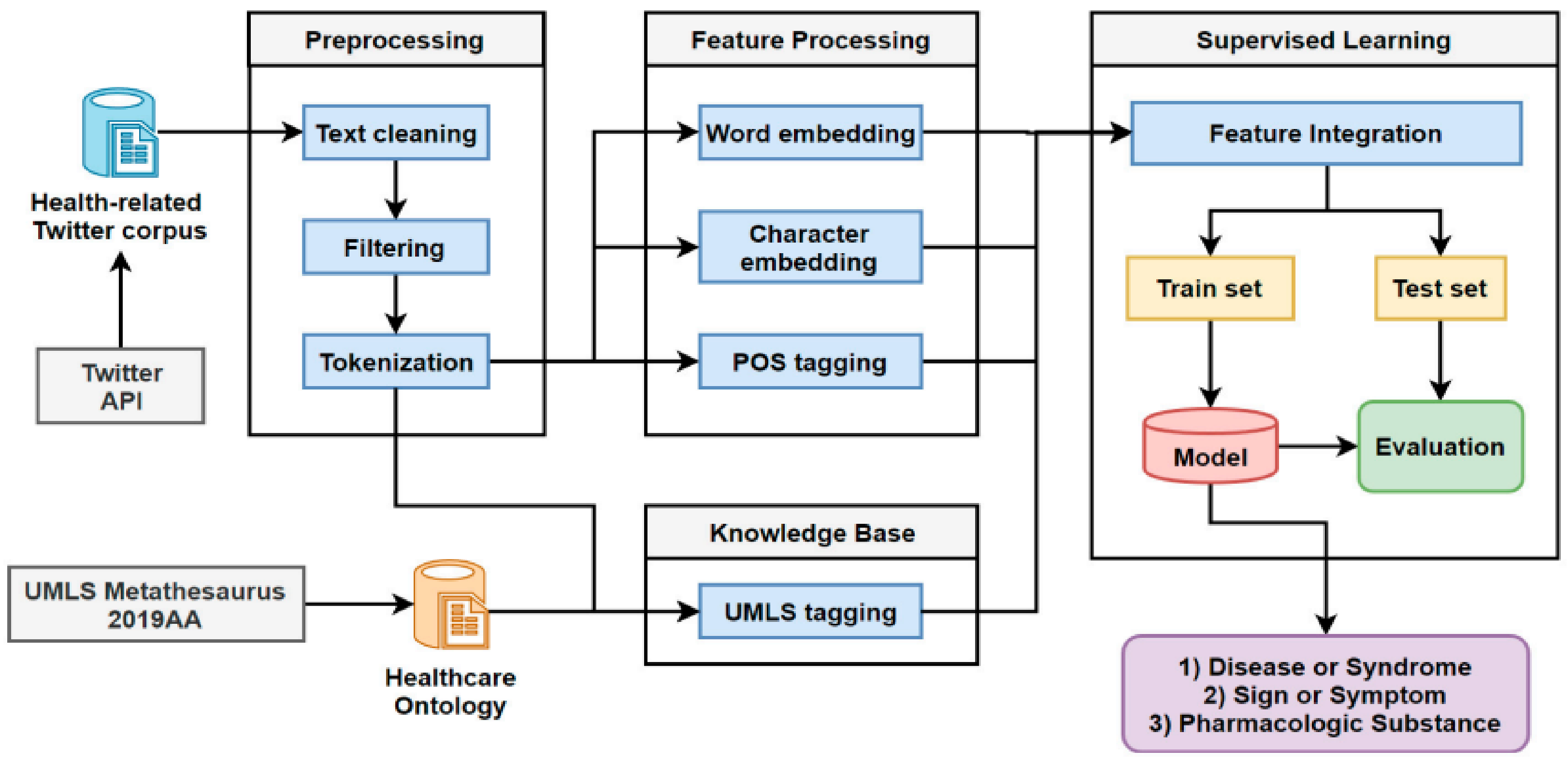


Figure 1. Overflow of HNER API: Application Programming Interface

2. Literature Review

- Traditional Rule-based Named Entity Recognition(NER) systems identify specific patterns and named entities however they have limited scalability and failed to handle ambiguity.
- To overcome these challenges, Deep learning models have been demonstrated to provide a significant improvement in predictive modeling by retaining the properties and activities of disease, symptoms, and drug discovery.
- The Bidirectional Long Short Term Memory Conditional Random Forest (BiLSTM-CRF) model has demonstrated its ability to attain a precision exceeding 90%, a recall of 73%, and an F1-score of 81% when identifying named entities related to diseases, and syndromes.

3. Research Objectives

This Research examines the effectiveness of using deep learning-based named entity recognition (NER) methods for extracting medical entities from unstructured healthcare text data and compares it with traditional rule-based NER methods in terms of accuracy, efficiency, and applicability to different healthcare domains.

4. The Data

Data is collected from UCI(University of California Irvine). The dataset contains 58000 health-related tweets from Twitter. It has been sourced from over 15 leading health news organizations, including BBC, CNN, and NYT.

Data Set Characteristics:	Text	Number of Instances:	58000	Area:	Computer
Attribute Characteristics:	Real	Number of Attributes:	25000	Date Donated	2018-02-19
Associated Tasks:	Clustering	Missing Values?	N/A	Number of Web Hits:	78492

5. Methodology

- The project utilizes a deep learning structure that relies primarily on a recurrent neural network (RNN).
- The data will be cleansed, Pre-processed which involves Tokenization through Jupyter Notebook using pandas, nltk.
- Word and character embeddings are performed using Convolutional Neural Network(CNN) and (Parts-of-speech)POS tagging methods.
- The training set was used to train both the Long Short Term Memory Conditional Random Forest (LSTM-CRF) and BiLSTM-CRF models, subsequently assessed on the testing set.
- Finally, Evaluate the performance using metrics such as precision, recall, and F1 score.

5. Early Indicators

- Pre-processing and cleaning of data involve tokenization, removing irrelevant tweets and stop words.

6. Next Steps

- Evaluating the precision, recall, and F1-score on the training and validation datasets.
- Fine-tuning the model can help improve the accuracy of the model for specific healthcare domains.

Contact: M.S.Shruthi
Tel: 899670082
Email: C00290801@itcarlow.ie

References:

1. Literature Review: Banville, H., Chehab, O., Hyvärinen, A., Engemann, D.A. and Gramfort, A., 2021. Uncovering the structure of clinical EEG signals with self-supervised learning. Journal of Neural Engineering, 18(4), p.046020.
2. Batbaatar, E. and Ryu, K.H., 2019. Ontology-based healthcare named entity recognition from twitter messages using a recurrent neural network approach. International journal of environmental research and public health, 16(19), p.3628.
3. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C., 2016. Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.
4. Jagannatha A, Liu F, Liu W, Yu H. Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0). Drug Saf. 2019 Jan;42(1):99-111. doi: 10.1007/s40264-018-0762-z. PMID: 30649735; PMCID: PMC6860017.