

Feature Extraction from Medical Journals

Author: Rishi Kumar

Date: 20 Dec 2021

Abstract:

The Machine learning and the natural language processing techniques are used in various medical domains. These methods are used for extracting relevant information in the medical domains. In this paper we describe an efficient health care system for identification of relations between diseases and treatments. In addition to this it also identifies the symptoms associated with a particular disease.

The Multi-nominal Naive Bayes algorithm is used for this extraction method. Other than this the stop word removal process and stemming algorithm is used. With the help of this proposed system we can make efficient medical decisions.

Therefore, efficient and accurate natural language processing (NLP) techniques are becoming increasingly important for use in computational data analysis, and advanced text mining techniques are necessary to automatically analyses.

The biomedical literature and extract useful information from texts. To bridge the gap between academic development and industrial requirements.

We are developing a web application that recognizes and extracts the entities from the content which helps to minimize time commitments from domain experts and the manual efforts on researching content.

Keywords: Health Care System, Machine Learning, Medline, Natural Language Processing, Stop Words Removal.

1. Problem Statement:

The increasing number of biomedical articles and resources, searching for and extracting valuable information has become challenging.

Business Problem: To Extract maximum number of named entities from the medical journals

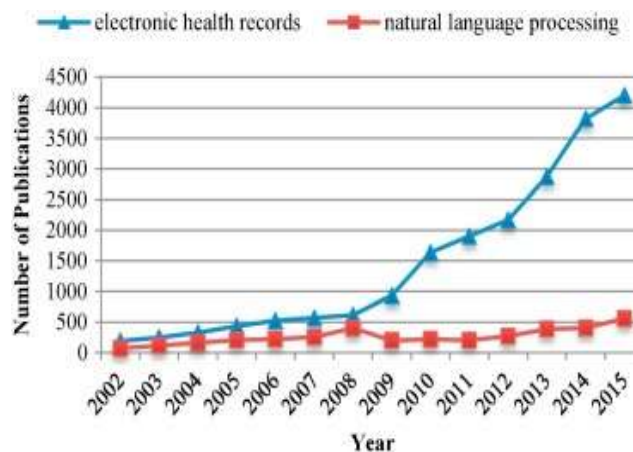
Maximize - The accuracy of the extracted entities.

Minimize - The error in selecting the named entity and time required.

1. Market/Customer/Business Need Assessment:

Analysis of narrative (text) data from electronic health records (EHRs) can improve population-scale phenotyping for clinical and genetic research. Currently, selection of text features for phenotyping algorithms is slow and laborious, requiring extensive and iterative involvement by domain experts.

This paper introduces a method to develop phenotyping algorithms in an unbiased manner by automatically extracting and selecting informative features, which can be comparable to expert-curated ones in classification accuracy.



The number of natural language processing (NLP)-related articles compared to the number of electronic health record (EHR) articles from 2002 through 2015.

2. Target Specifications and Characterization:

The medical databases such as Medline contain lots of medical related articles. The main problem with these databases is that all research discoveries come and enter the repository at high rate making the process of identifying and disseminating the healthcare information a difficult task.

Here the proposed approach eliminates such problem and introduces a new method for easily extracting relevant medical information from these articles. The proposed system displays to the user information regarding diseases, treatments and symptoms and in addition it also identifies three semantic relations between diseases and treatments.

Always an article contains both relevant and irrelevant information.

Windows, version 17.0 SPSS Inc.). 3. Results Fifteen patients (73.6% female) were studied. Arterial hyper-tension and hypothyroidism (under treatment and well controlled) were the most common comorbidities with patients in functional classes (NYHA) II and III (Table 1). It is important to note that COPD and ESKD patients were not excluded from the study. Concomitant medication was as follows: diuretic DRUG (73%) and adrenergic beta blocker receptor (BB, DRUG 73%) agents were the most commonly employed; 46.6% also received mineralocorticoid receptor antagonists (MRA medications) DRUG and angiotensin-converting-enzyme inhibitors (ACEIs)/angiotensin receptors blockers (ARB). \uparrow Cardiology Research and Practice 3 Maximum amplitude time/total time (MAT/TT) index Preischemic DRUG period Postischemic DRUG period Before 29.1 ± 2.2 30.4 ± 2.1 Ivabradine After 24.3 ± 3.2 23.3 ± 2.9 $P = 0.05$ $P = 0.002$ Figure 1: Pre and after ischemic period, before and after 6 months of follow-up of oral ivabradine. Patients received an average of 12.5 (range 10–15) mg/ day FREQUENCY during the 6 months of followup. It was particularly interesting that heart rate did not decrease in any patients below the 10% recommended (88 versus 82 beats/min) in the literature in spite of the top doses received. Figure 1 shows the maximum

3. External Search (information sources/references):

- <https://pdfminersix.readthedocs.io/en/latest/>
- <https://pubmed.ncbi.nlm.nih.gov/34412834/>
- <https://github.com/kormilitzin/med7>
- <https://www.frontiersin.org/articles/10.3389/fcell.2020.00673/full>
- <https://flask.palletsprojects.com/en/2.0.x/tutorial/index.html>
- <https://www.w3schools.com/mysql/default.asp>
- <https://analyticsindiamag.com/guide-to-named-entity-recognition-with-spacy-and-nltk/>

4. Benchmarking:

In view of weakness inherent in manual searching of text, it has become imperative to seek other efficient ways to carry out text mining.

The massive volume of bio-medical information stored in soft documents copies form, which obviously could be due to a substantial increase in scientific research over the years has necessitated the use of text mining technology. Searching and processing information from documented data is time-consuming in many areas for example bio-medical literature and is becoming not practical and easy to achieve without computer support.

Thus, today the need for intelligent text handle applications that can replace or support human information exploration in bio-medical text documents is strong.

It has become extremely difficult for biologists to keep up with the relevant journals in their own discipline, let alone publications in other, related disciplines. Bio-medical literature considered as a source of authentic medical knowledge which is critical for e-health applications.

These kinds of e-health applications have a huge commercial prospect.



5. Applicable Patents:

Decision tree based NER model was built by Skein et al. that used features such as part-of speech tags extracted by morphological analyzer, specialized dictionary and character based information. This was developed for Japanese.

Bike et al. Used hidden Markova model (HMM) for identification of named entity. Features like bi-gram and orthographic features like word case, word shape etc. were used. In his Ph.D. thesis, Borthwick used maximum entropy (MaxEnt) algorithm.

McCallum ET al.extracted NER using algorithm based on conditional random fields. A semi Markov conditional random field algorithm was proposed by Sarawagi et al. for extraction of named entity.

The researches extended the semi Markov model with use of dictionary and notion of similarity function. An overall survey of NER research was provided by Naidu and Sekine.

Luu proposed a framework that is based on different text mining and machine learning algorithms for addressing the challenges of clinical named entity recognition. The framework proposed has multiple levels and builds complex NER tasks. Different data sets-the CLEF 2016 challenge and BIONLP/NLPBPA 2004 were used for evaluation of the proposed method and the results validated the framework.

6. Applicable Regulations:

Nowadays, in context of bio-medical domain, the bio medicinal work is going to increase rapidly because of the time, the developing measure of the content on World Wide Web (WWW). Internet, a viable and productive information recovery system, is required. So in bio-medical domain the bio medicinal work has been expanded; the measure of content in online sources i.e. MEDLINE, which is, as of now, the biggest archive for bio medical works.

In biomedical work, namely, elements signifies to word or grouping of the word which represent particular terms, such as; protein, DNA, RNA or ailment name. Because of the enormous development of content, effective information recovery and automation is required.

Bio-NER has been difficult when contrasted with normal NER (Area, Names, Time, Date and so on).

7. Applicable Constraints:

- First, the elements of biomedical filed unavailability of a tenacious morphology and consequently, they are not a formal noun (people), places or things comprising letters, numbers and so on which, additionally, expanding disambiguate of grouping.
- Second, highest critical arrangement problem is the united conveyance of the content, for instance; Cancer can be delegated a modifier; it can be additionally named a particular ailment and malady class and so on.

8. Business Opportunity:

Let's suppose you are designing an internal search algorithm for an online publisher that has millions of Medical articles. If for every search query the algorithm ends up searching all the Biological words in millions of articles, the process will take a lot of time.

Instead, if Named Entity Recognition can be run once on all the articles and the relevant entities (tags) i.e.: (Drug, strength, Dose, Doses, Disease) associated with each of those articles are stored separately, this could speed up the search process considerably. With this approach, a search term will be matched with only the small list of entities discussed in each article leading to faster search execution.

9. Concept Generation:

The final milestone that's to be achieved is simple – making things as simple as possible,

The greatest artist Albert Einstein has once quoted – “Everything should be made as simple as possible, but not simpler.” Speaking of simplicity

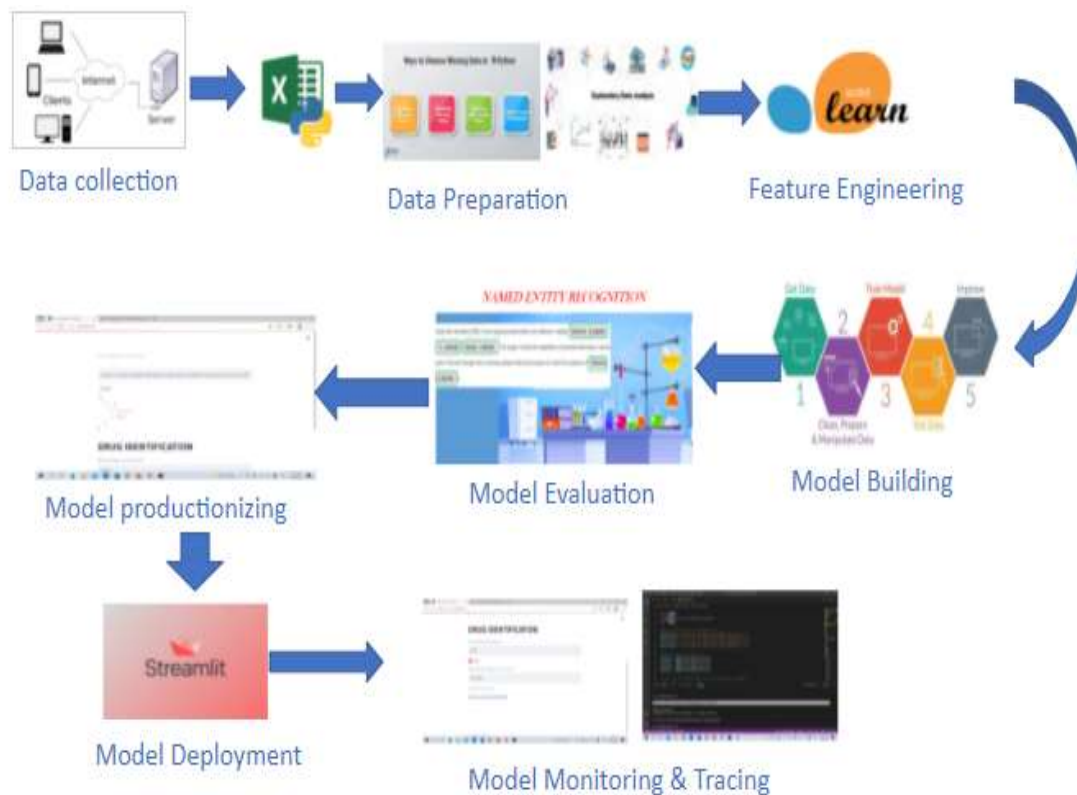
Machine Learning (to its best) biomedical field, letting one focus on Extract Name from article tasks.

Keeping all the things mentioned here and other one's previously under consideration, one thing is for sure to happen – market, business & customer needs are met to a greater extent, for sure. Now, speaking of biomedical field is highly multi-dimensional because it not only covers Drug but equally focuses on Combination of Drug and Doses and strength.

10. Concept Development:

The purpose of extracting of Bio-Medical Entities is to recognize the particular entities, whether word or phrases, from the unstructured data contained in the text. This work proposes different approaches and methods, i.e. Machine Learning Hybrid Classification, Rule Based Non-tested Generalized Exemplars and Partial Decision Tree (PART) Learners for Bio-Medical Named Entity Recognition.

11. Final Product Prototype with Schematic Diagram

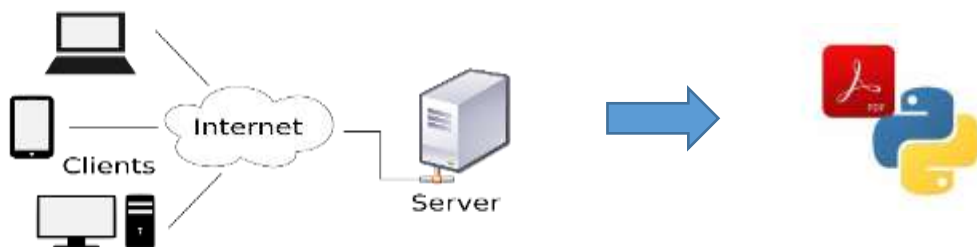


12. Product details

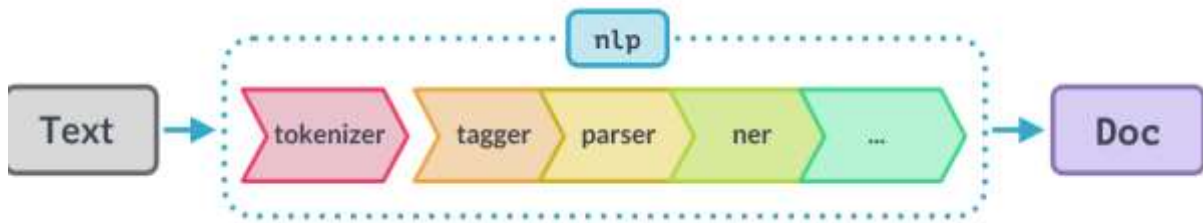
- Data collection from biomedical journal in pdf.

Key Inferences:

- Determine what data is needed.
- Collect data, if required data is not available.
- Explore data.
- Verify data quality.



- Exploratory Data Analysis(EDA):
- Data exploration can help cut down your massive data set to a manageable size where you can focus your efforts on analyzing the most relevant data.
- A feature type has the following attributes that can be overridden:
 - description: A description of the feature type.
 - Name: The name of the feature type



- As you can see in the figure above, the NLP pipeline has multiple components, such as tokenizer, tagger, parser, NER, etc. So, the input text string has to go through all these components before we can work on it.
- Using NLTK to tokenize is a toolkit build for working with NLP in Python. It provides us various text processing libraries with a lot of test datasets

The screenshot shows a Jupyter Notebook interface with the following code cells:

```

[ ] 9879
    98206
    95789
    116098
    162118
    264866

[ ] #removing line break/new line character from corpus
    corpus=corpus.replace("\n"," ")

[ ] import nltk
    from nltk.tokenize import sent_tokenize
    sent=sent_tokenize(corpus)

[ ] print(corpus)

Haematologia, Vol. 30, No. 1, pp. 27 - 30 (2000) (cid:211) VSP 2000. Short communication Serum L-selectin and P-selectin levels in lymphomas I. L. NAZMEDARU "
[ ]

[ ] import spacy
    import pandas as pd
    import warnings
    warnings.filterwarnings('ignore')

[ ] !pip install https://huggingface.co/kornilitsin/es-core-med7-1g/resolve/main/en-core-med7-1g-any-py3-none-any.whl

```

- ML Modelling

Med7: a transferable clinical natural language processing model for electronic health records.

Spacy: is an open-source software python library used in advanced natural language processing and machine learning



```
med7 = spacy.load("en_core_med7_lg")

# create distinct colours for labels
col_dict = {}
seven_colours = ["#ed194b", "#1f77b4", "#ff7f0e", "#ffbb78", "#ff8c00", "#9467bd", "#8c564b"]
for label, colour in zip(med7.pipe_labels['ner'], seven_colours):
    col_dict[label] = colour

options = {'ents': med7.pipe_labels['ner'], 'colors': col_dict}

doc = med7(corpus)

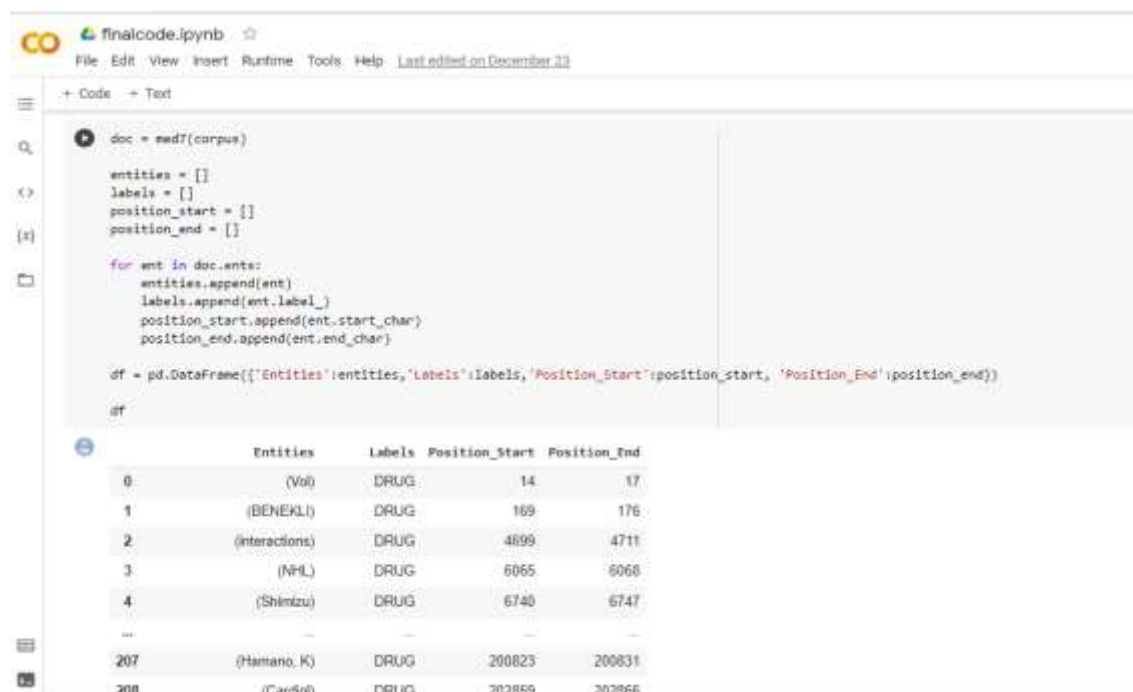
spacy.displacy.render(doc, style='ent', jupyter=True, options=options)

[ent.test, ent.label_] for ent in doc.ents]
```

C:\Users\ishh\anaconda3\envs\med7\lib\site-packages\spacy\util.py:833: UserWarning: [W001] Model "en_core_med7_lg" (1.1.1.1) was trained with spaCy v2.1 an...

Hematologia [Vol, 30, No. 1, pp. 27 - 30 (2000) (nd 211) VSP 2000. Short communication Serum L-selectin and P-selectin levels in lymphomas I. C. HAZNEDARO, GLU. M. BENEKL, [nd 3], O. OZCERE, M. C. SAVA, S. I. H. GULLU, S. V. DUNDAR and S. KIRAZLI Department of Internal Medicine, Hematology and Oncology Division, Hacettepe University Medical School, Ankara, Turkey Abstract—The migration of normal and malignant lymphoid cells is governed by specific adhesion molecules. Selectins comprise a family of adhesion receptors expressed by leukocytes, platelets and endothelial cells. In this study, the serum levels of soluble L-selectin and P-selectin were measured in patients with non-Hodgkin's lymphoma and Hodgkin's disease and

Final Output



```
doc = med7(corpus)

entities = []
labels = []
position_start = []
position_end = []

for ent in doc.ents:
    entities.append(ent)
    labels.append(ent.label_)
    position_start.append(ent.start_char)
    position_end.append(ent.end_char)

df = pd.DataFrame({'Entities': entities, 'Labels': labels, 'Position_Start': position_start, 'Position_End': position_end})
df
```

	Entities	Labels	Position_Start	Position_End
0	(Vol)	DRUG	14	17
1	(BENEKL)	DRUG	169	176
2	(interactions)	DRUG	4699	4711
3	(NHL)	DRUG	6065	6068
4	(Shimizu)	DRUG	6740	6747
...
207	(Hamano, K)	DRUG	200823	200831
208	(Cardio)	DRUG	202859	202866

13. Code Implementation

• Model Description

Libraries Used:

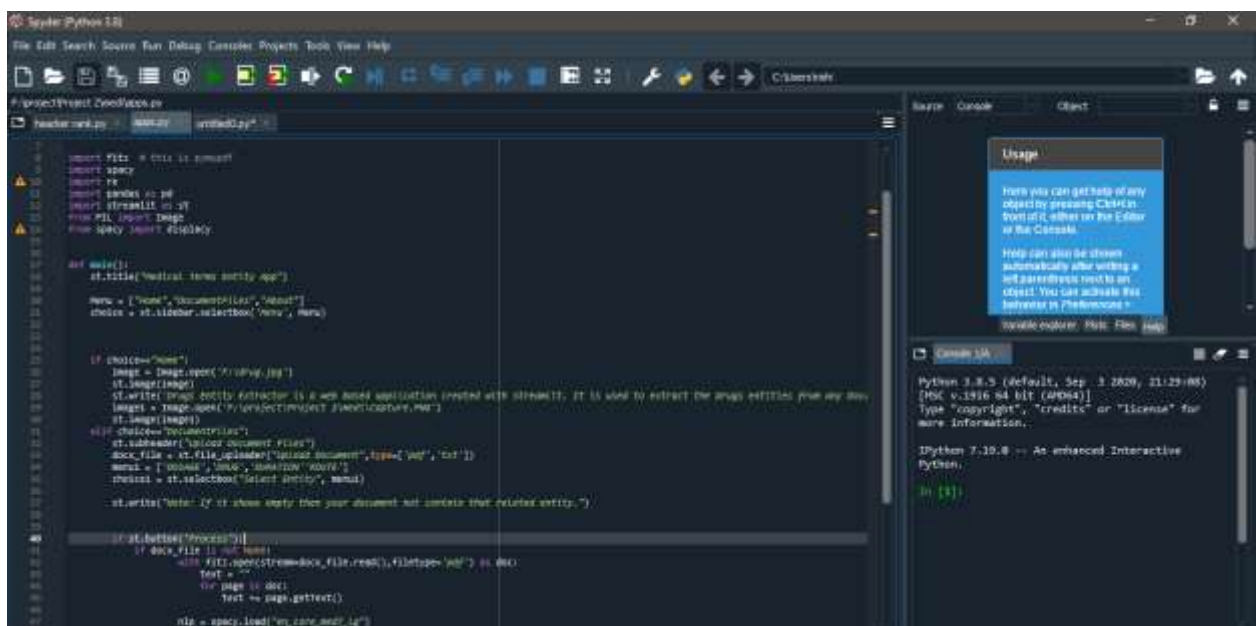
- Pdfminer
- Hugging face
- Spacy
- En-core-med7-trf
- NLTK
- stream lit

Description:

- Pdfminer is used to merge all the medical journals
- Pdfminer is used to extract text from the pdf files.
- Spacy is used to build NLP model to extract features.
- En-core-med7-trf is used to extract specifically medical entities.
- Stream lit is an open-source python framework for building web apps for Machine Learning and Data Science.

Deployment Code:

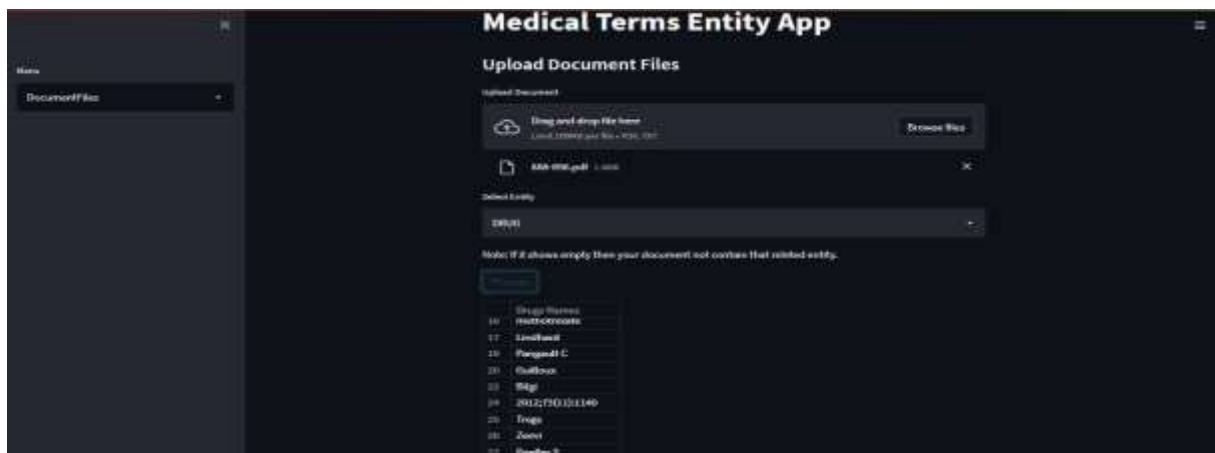
Using stream lit to deployment



[Home page](#)



Input Page



Output in CSV file

[illegible]

- GitHub link to the code implementation

<https://github.com/Rishiverma1993/Feature-Extraction-from-Medical-Journals.git>

14. Conclusion

Because of diverse writing style of clinicians, the rules and patterns are not generalizable. These issues can be addressed by making use of technologies like machine learning. Named entity recognition is grouped into three approaches. Machine learning based approaches, rule-based approaches and dictionary based approaches. The systems that use machine learning based approach focus on choosing effective features for classifier building. Several researchers have extensively used machine learning models for clinical NER.

Databases such as PubMed which include medical publications have generated lot of interest among researchers for applying information extraction techniques to medical literature.

In an attempt to contribute to the research in this area, this work proposed a machine learning model for clinical NER. The model proposed performed better compared to some of the existing methods.