ADOP FERT-Automatic Detection of Occupations and Profession in Medical Texts using Flair and BERT

Fazlourrahman Balouchzahi[a,1], Grigori Sidorov[b,1], and Hosahalli Lakshmaiah Shashirekha[c,2]

[1] Center for Computing Research, Instituto Politécnico Nacional, CDMX, Mexico
[2] Department of Computer Science, Mangalore University, Mangalore - 574199, India
[a] *frs_b@yahoo.com*, [b] *sidorov@cic.ipn.mx*, [c] *hlsrekha@gmail.com*

Abstract. Technological developments in healthcare industry are generating lots of electronic health records as well as text data which is usually referred as medical text data. Processing medical text data in unstructured form is not only challenging but also has lot of applications. Named entity recognition, the task of extracting named entities and classifying them into predefined categories is an important preprocessing step in the NLP pipeline. Extracting named entities from medical text is very useful for many applications and at the same time very challenging because of the characteristics of medical text data. Considering the gravity of medical text processing, in this paper, we (Team MUCIC) describe the models submitted to "MEDical DOocu-ments PROFessions recognition" (MEDDOPROF), a first shared task consisting of three Tracks, namely: Track 1: MEDDOPROF-NER, Track 2: MEDDOPROF-CLASS, and Track 3: MEDDOPROF-NORM, in Spanish language. We participated in Track 1 and 2 and proposed two models based on fine-tuning BERT embeddings using i) BertForTokenClassification from transformers and ii) Flair framework, for the automatic detection of Occupations and Professions in medical text. The model using BertForTokenClassification reported micro F1 scores of 0.629 and 0.598 for Track 1 and 2 respectively, while the Flair framework model reported micro F1 scores of 0.8 and 0.764 for Track 1 and 2 respectively. Further, the Flair framework model for MEDDOPROF-NER track became one among the best models.

Keywords: Profession· Medical Documents· NER· BERT· Flair· Embeddings.

1 Introduction

The recent updates on medical and health-care information systems are generating large amount of Electronic Health Records (EHRs) [1] as well as text data

IberLEF 2021, September 2021, Málaga, Spain.

in medical domain. Despite the popularity of existing systems to manage EHRs there is a massive amount of unstructured medical text data that are required to be transformed into a more structured format for further processing [2]. Medical text processing or text analytics is one of the exciting areas of research in NLP world that deals with various applications like Text Classification (TC) (classification of medical records, classification of medical news articles), Text Summarization (automatic generation of summaries from medical news articles, summarization of clinical in-formation), Hypothesis Generation and Knowledge Discovery and so on.

One of the most popular text processing applications is Named Entity Recognition (NER), which is used to automatically recognize and classify Named Entities (NEs) [3] representing names of persons, and organizations, locations and so on from a given natural language text. NER is a crucial step in NLP pipeline as performance of the NER module decides the performance of subsequent modules [4] and NER systems also act as a preprocessing step for tasks like Relation Extraction [5]. Medical NER which deals with extracting medical NEs such as disease names, symptoms, medical conditions, medications, medical professions, employment status, etc., from medical texts is challenging due to specialized terminology, huge number of alternate spellings, and multi-word NEs. Even though a variety of works have been explored for processing medical texts in diverse aspects, very few works are reported in the literature on processing texts related to medical profession and employment status in general and in particular to identify and classify the NEs describing medical occupations in medical documents.

To address the challenges of identifying and classifying the NEs describing medical occupations and employment status in Spanish medical documents, in this paper, we (Team MUCIC) describe the models submitted to two Tracks of MEDical Documents PROFessions recognition (MEDDOPROF)[6] task. MEDDOPROF is a first shared task of its kind that consists of three sub-tracks and the description of Tracks 1 and 2 (the ones in which we participated) are briefly given below:

- **Track 1 - MEDDOPROF-NER:** includes identifying the portion of texts that mentions an occupation and classifying them into one of three predefined categories, namely: PROFESION (PROFESSION), SITUACION_LABORAL (WORKING_STATUS) or ACTIVIDAD (ACTIVIDAD).
- **Track 2 - MEDDOPROF-CLASS:** includes automatically finding the beginning and end of occupation mentions and classifying them into one of the further categories, namely: PACIENTE (Patient), FAMILIAR (Family member), SANITARIO (Health professional) or OTRO (Other).

Based on the description of Tracks and categories, Track1 and 2 can be modeled as an NER task of identifying the NEs (tokens) which could be either a single word or multi-word and then classifying/labeling them into one of the pre-defined categories according to the Tracks. Of late transformer based models are achieving state-of-the art results for many NLP tasks compared to various Machine Learning (ML) and Deep Learning (DL) models. To explore transformers[18], we have proposed two models based on fine tuning Bidirectional Encoder

Representations from Transformers (BERT) [7] embeddings using i) BertForTokenClassification class from transformer library and ii) Flair framework, for the task of automatic detection of occupations and professions in Spanish medical texts for Track 1 and 2 of MEDDOPROF.

BERT as a language representation model employs bidirectional representations from text to pre-train both left and right context. It also can be fine-tuned for downstream tasks such as NER and TC, only by adding a specific output layer [8]. The difference between BERT and Embeddings from Language Models (ELMo) [9] that uses pre-trained language models is that ELMo uses the language model as additional features whereas BERT enables fine-tuning all parameters of pre-trained language model to make it task-specific on downstream task [7].

Flair framework provides a standard model for training along with hyperparameter selection and unified interface that reduce complexity of using various embeddings and enables researchers to mix the embeddings effectively. It also offers various embeddings that are publically available in HuggingFace[19]. In the current work Flair is used with BERT embeddings [10]. Generative Pre-trained Transformer (OpenAI GPT) is another architecture that allows fine-tuning. However, it is limited in unidirectional representation whereas BERT utilize bidirectional representation which effectively overcomes the restrictions of OpenAi GPT's architecture [7].

The rest of paper is organized as follows: Section 2 gives an overview of the works carried out in the related area and Section 3 describes the proposed methodology. While Section 4 presents Experiments and Results, Section 5 gives the conclusion and throws light on future work.

2 Related Work

ML classifiers have reported reasonable and competitive performance for various TC applications such as NER, Sentiment Analysis, Opinion Mining, etc. However, these days' Neural Network (NN) based systems are commonly used for various TC applications in various domains including medical domain. Some recent adventures in medical text processing are described below:

Yepes et al. [11] developed a NN based system for the identification of medical NEs from Twitter posts. The authors used 148 million tweets collected to generate a CBOW word embeddings that is used as weights in model construction. Two LSTMs are used to construct a sequence to sequence model where first LSTM acts as an encoder to encode the texts to vectors and second LSTM as main classification model to classify the tokens. The proposed model on Micromed[20] dataset containing 1300 tweets obtained F1 scores of 0.665, 0.682, and 0.718 on diseases, pharmacological substances and symptoms entities respectively.

Li et al. [12] presented a NN based model for medical NER in Chinese texts. The authors used character level and word level embeddings to capture orthographic and lexicosemantic features along with POS tags as word information

features. A Chinese medical corpus containing 12,498 records is used and 1739 records out of them were manually annotated into two categories, namely, subject and lesion where symptoms related to body are considered as subject and lesion refers to the pathological changes of the subjects. The dataset is transformed into BIESO NE representation where B, I, E, and O indicates the beginning, inside, end, and outside of the entity respectively and S illustrates that the entity consists of only a single word. RNN, LSTM, GRU, BLSTM and BGRU are experimented with various configurations and feature combinations. Among all, BGRU without employing any embeddings and only POS tags features, had the best performance with 90.36 and 90.48% F1 scores for subject and lesion detection tasks respectively.

Feature engineering step is one of the important steps in any NLP task as it aims to improve the performance of the system. Weegar et al. [13] explored the impact of simple feature engineering in NER systems for medical texts of three languages namely, English, Swedish, and Spanish. The authors examined some basic features including POS and semantic tags along with prefixes, window size, and capitalization. Averaged structured perceptron algorithm is used with SemEval-2014 Task 7 Analysis of Clinical Text Shared Task dataset containing 9,694 disease NEs for English, EHRs consisting of patient records developed by Oronoz et al. [14] containing 3,362 instances of diseases and 1,406 drugs entities as Spanish dataset and a dataset containing 4,000 entities corresponding to body parts, disorders and findings from over 500 different clinical units at Karolinska University Hospital for Swedish released by Dalianis et al. [15]. The observation of the results illustrates that in many cases simple but neglected features can significantly enhance the performance of the systems. Their best performing systems which obtained F1 scores of 66.40 and 68.41, and 68.22 for English, Swedish, and Spanish languages respectively used specialized medical dictionaries.

Sometimes instead of working on features and model construction, proposing new representation for the data might be more efficient. In one of such studies Hamada et al. [4] proposed FROBES Segment Representation (SR) model which is an exten-sion of IOBES model, when NEs are multi-word in nature. In the proposed FROBES model F, R, O, B, and E represents front, rear, outside, begin, and end respectively and S represents a single word. FROBES is an extension of IOBES where tag I in IOBES is replaced with F and R when entity has more than two words. As it considers both halves of entities, first half are annotated with B and F and second half contains R and E. The proposed SR scheme is evaluated using BiLSTM as baseline model on two datasets, namely: i2b2/VA 2010 challenge dataset and JNLPBA 2004 shared task dataset and the results reported by the authors illustrate that using FROBES improved the performance slightly. However, ensembling the baseline models with different SR models, namely: IOB2, IOBES, and FROBES outperformed the baseline models with F1 scores of 71.99 and 83.62 on the same datasets.

3 Methodology

The two proposed models based on fine-tuning BERT embeddings using i) BertForTokenClas-sification from transformers and ii) Flair framework, designed and evaluated for the Tracks 1 and 2 of MEDDOPROF are described in this section.

3.1 Data Transformation

The datasets provided by the organizers of MEDDOPROF shared task for the sub-Tracks are in Brat standoff annotation format. As per this format, for each text file there is a corresponding annotation file consisting of an annotation ID, a label, and beginning and ending offset for each NE which could be of a single word or multi-word. More details of Brat standoff format can be found on its website[21]. As the data in CONLL IOB [22] format is easy to handle, the given data in Brat standoff annotation format is transformed to CONLL format with IOB representation using brat_to_conll.py [23] module. IOB representation assigns the tags I and O for the tokens that are inside and outside the NE respectively and assigns the tag B for the first word of the NE [2, 4]. A snapshot of data in Brat format and CONLL (IOB) format is shown in Figure 1.

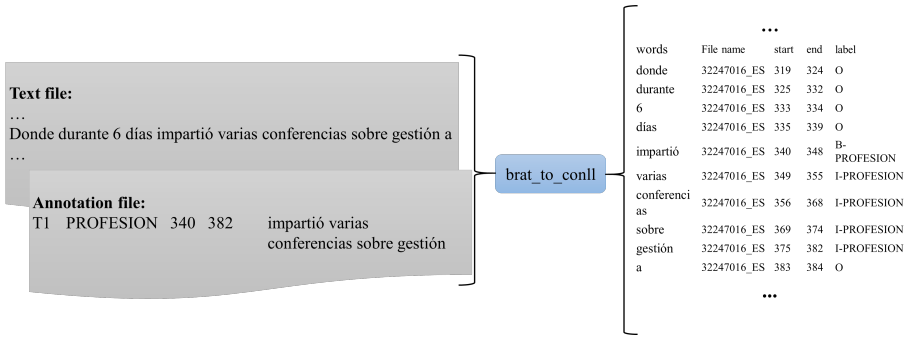

Fig. 1. Snapshot of data in Brat standoff and CONLL IOB format

As the data transformed into CONLL IOB format is used to train the classifier models, the predictions of the models will also be in CONLL IOB format. This requires a post processing step to re-transform the predictions in CONLL IOB format back to Brat standoff annotation format to generate .ann files as output as required by the organizers.

3.2 Models

The main component of the proposed models is BETO [16, 17, 24] which is a Spanish BERT language model trained on a large amount of Spanish unannotated corpora [25]. In this work, we have used bert-base-spanish-wwm-cased[26]

model which is more efficient for NER tasks as capitalization play a major role in identifying NEs.

BertForTokenClassification using Transformer:

The first step of this model is to fine tune the BERT model on downstream task using transformers library. Using the data in CONLL IOB format, the fine-tuned models are further trained for Track 1 and 2 of the shared tasks. For each test dataset, the models generate tagged sequences sentence-wise in IOB annotation format which will be converted back to Brat standoff annotation format manually.

As BERT based models require to be fed with sequences of same length, the maximum length of sequences is set into 510 and the shorter sequences are padded to this length. However, an attention mask is employed to avoid distracting models with padded elements. Similar to Keras[28], BERT support attention masks that are used to allow the model focus on main part of the sequence ignoring padded elements. In other words, mask is typically used for attention when a batch has varying length of sentences. Therefore, it takes real tokens for training by assigning 1 to in sequence tokens and 0 for out of sequence. After assembling training data and corresponding masks using PyTorch [27], BETO will be initialized using BertForTokenClassification class from transformers library which adds a token level predictor on BERT model.

Setting the optimizer to AdamW the models have been trained for 50 epochs. Figure 2 represents training and validation loss where validation set is 10% of training set. Overview of the model based on BERT using transformer library is shown in Figure 3.

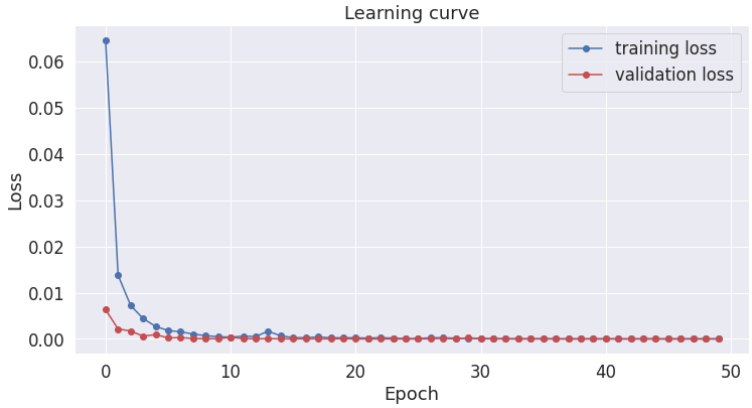

Fig. 2. Training and Validation loss while fine-tuning BERT

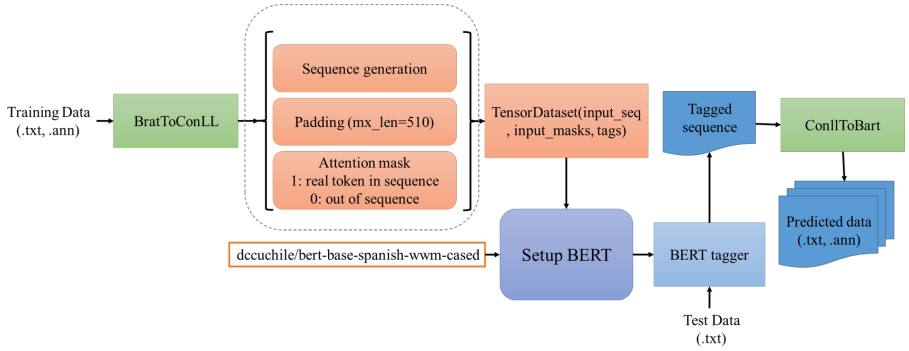

Fig. 3. Overview of BERT model using Transformer

Table 1. Parameters in Flair model

Parameter	Max_len	Hidden_size	Learning_rate	Mini_batch_size	Epochs
Value	512	256	5.0e-6	4	10

Flair with BERT Embedding:

Flair is a PyTorch based NLP tool that provides a model training framework in which various embeddings and language models can be used individually or in combination and fine-tuned for downstream tasks with special support for Medical domain data [10]. However, to compare the performance of this model with that of BertForTokenClassification model, bert-base-spanish-wwm-cased model is used and fine-tuned using Sequence Tagger from Flair which is BiLSTM based backend. It is also possible to use CRF on top of the model, but it is not used in this work. As Flair requires the training data in CONLL format, data from Brat standoff annotation format is transformed to CONLL IOB format as described in Section 3.1 and is loaded using ColumnCorpus class from Flair library. A summary of the layers used in this model is given in our Github page[29]. Parameters of the proposed model are set as given in Table 1 and an overview of proposed Flair model is presented in Figure 4.

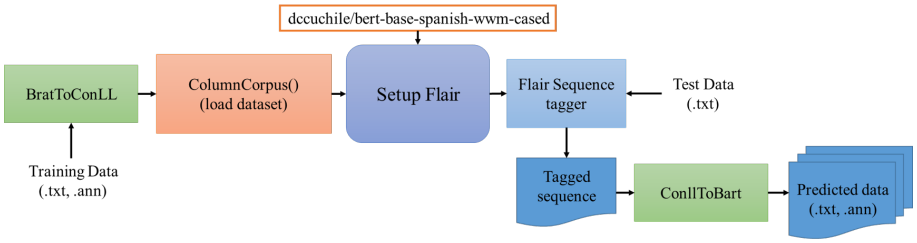

Fig. 4. Overview of proposed Flair mode

4 Experiments and Results

The main requirement of any task is an annotated dataset for training the models. MEDDOPROF corpus provided by the organizers contains 1844 clinical cases with more than 20 specialties annotated manually by clinical and linguistics experts following strict guidelines. Each clinical case is stored as a single text file along with a corresponding Brat standoff annotation file. The description of the dataset is available in the task website [30] and the descriptions of the labels for both the Tracks are given in Table 2.

Evaluating models' performance is the most important task. As per the submission guidelines [6], the predictions for each test file should be in Brat standoff annotation format, i.e., the annotation file should have an extension .ann and should consist of an annotation ID, label, correct beginning and ending offset for each predicted NE, in one line, similar to the annotation file given in the training set. However, the value of annotation ID is generated at random as it does not have any influence on the prediction. Annotation files are generated for each file in the test set and submitted to the task organizer for evaluation. The performance of the models is evaluated in terms of Micro average Precision, Recall and F1-score.

Organizers provided the results obtained by a simple lookup system and the annotations from the training data as baseline. Therefore baseline's results and performances of the proposed models reported by the organizer for both the Tracks with Micro average scores are shown in Table 3. The results illustrate that the pro-posed models obtained quite good performances for both the Tracks. Further, the results also illustrate that the proposed models performed better for MEDDOPROF-NER task. In addition, the model using Flair framework and BERT embeddings outperformed the other proposed model and became one of the best performing models in the shared task.

Table 2. Labels description in MEDDOPROF-NER and MEDDOPROF-CLASS

Track	**Labels**	**Token Description**
MEDDOPROF-NER	PROFESION	Indicates a profession
	SITUACION_LABORAL	Indicates an employment status
	ACTIVIDAD	Indicates an activity
MEDDOPROF-CLASS	PACIENTE	Token is related to the patient
	FAMILIAR	Token is related to a family member
	SANITARIO	Token is related to a health professional
	OTROS	Token is related to someone else

Table 3. Performances of proposed models (Micro average)

Subtask	Model	Precision	Recall	F1-score
MEDDOPROF-NER	Baseline	0.465	0.508	0.486
	BERT	0.809	0.515	0.629
	Flair-BERT embeddings	**0.813**	**0.788**	**0.8**
MEDDOPROF-CLASS	Baseline	0.391	0.377	0.384
	BERT	0.77	0.488	0.598
	Flair-BERT embeddings	**0.77**	**0.75**	**0.764**

5 Conclusion and Future Work

Medical text processing is one of more exciting as well as vital task in NLP. Considering its importance MEDDOPROF has called for a shared task with three Tracks and we participated in two of them, namely: MEDDOPROF-NER and MEDDOPROF-CLASS for the automatic detection of occupations and profession in Spanish medical texts. We (team MUCIC) proposed two models using BERT embed-dings, namely: BertForTokenClassification from transformers and Flair framework. The results illustrate that the models performed better in NER and Flair model outperformed the other model in both the Tracks and also obtained very good results with micro F1-scores of 0.8 and 0.764 for MEDDOPROF-NER and MEDDOPROF-CLASS respectively. Further, the Flair model for MEDDOPROF became one of the best per-forming models in the shared task. As future work it is planed for exploring the Language Understanding with Knowledge-based Embeddings (LUKE) model which is a new pre-trained contextualized representation of words and entities based on transformer. Improving the performances of system with modifications in NEs representations and also exploring various learning approaches for task of NER in medical texts are other plans of future work.

6 Acknowledgements

Team MUCIC deeply appreciates the organizers of MEDDOPROF shared task for their efforts, guidance and supports during the task and the anonymous reviewers for their valuable comments.

References

1. Nayel H, Shashirekha HL. Improving NER for Clinical Texts by Ensemble Approach using Segment Representations. In Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017) 2017 Dec (pp. 197-204).
2. Cui M, Bai R, Lu Z, Li X, Aickelin U, Ge P. Regular Expression Based Medical Text Clas-sification using Constructive Heuristic Approach. IEEE Access. 2019 Oct 11; 7:147892-904.
3. Balouchzahi Fazlourrahman, H. L. Shashirekha. PUNER-Parsi ULMFiT for Named-Entity Recognition in Persian Texts. No. 4224. EasyChair, 2020.

4. Nayel H, Shashirekha HL, Shindo H, Matsumoto Y. Improving Multi-Word Entity Recog-nition for Biomedical Texts. arXiv preprint arXiv:1908.05691. 2019 Aug 15.
5. Shashirekha HL, Nayel HA. A Comparative Study of Segment Representation for Biomedi-cal Named Entity Recognition. In 2016 International Conference on Advances in Compu-ting, Communications and Informatics (ICACCI) 2016 Sep 21 (pp. 1046-1052). IEEE.
6. Salvador Lima-López, Eulàlia Farré-Maduell, Antonio Miranda-Escalada, Vicent Brivá-Iglesias, Martin Krallinger. "NLP applied to occupational health: MEDDOPROF shared task at IberLEF 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts". Procesamiento del Lenguaje Natural 67, 2021.
7. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-Training of Deep Bidirectional Trans-formers for Language Understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.
8. Campillos-Llanos L, Valverde-Mateos A, Capllonch-Carrión A, Moreno-Sandoval A. A Clinical Trials Corpus Annotated with UMLS Entities to Enhance the Access to Evidence-Based Medicine. BMC medical informatics and decision making. 2021 Dec; 21(1):1-9.
9. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep Con-textualized Word Representations. arXiv preprint arXiv:1802.05365. 2018 Feb 15.
10. Akbik A, Bergmann T, Blythe D, Rasul K, Schweter S, Vollgraf R. FLAIR: An Easy-To-Use Framework for State-Of-The-Art NLP. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations) 2019 Jun (pp. 54-59).
11. Yepes AJ, MacKinlay A. NER for Medical Entities in Twitter using Sequence to Sequence Neural Networks. In Proceedings of the Australasian Language Technology Association Workshop 2016 2016 Dec (pp. 138-142).
12. Li J, Zhao S, Yang J, Huang Z, Liu B, Chen S, Pan H, Wang Q. WCP-RNN: a Novel RNN-based Approach for Bio-NER in Chinese EMRs. The journal of supercomputing. 2020 Mar; 76(3):1450-67.
13. Weegar R, Casillas A, de Ilarraza AD, Oronoz M, Pérez A, Gojenola K. The Impact of Sim-ple Feature Engineering in Multilingual Medical NER. In Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP) 2016 Dec (pp. 1-6).
14. Oronoz M, Casillas A, Gojenola K, Perez A. Automatic Annotation of Medical Records in Spanish with Disease, Drug and Substance Names. In Iberoamerican Congress on Pattern Recognition 2013 Nov 20 (pp. 536-543). Springer, Berlin, Heidelberg.
15. Dalianis H, Henriksson A, Kvist M, Velupillai S, Weegar R. HEALTH BANK-A Work-bench for Data Science Applications in Healthcare. CAiSE Industry Track. 2015 Jun 11; 1381:1-8.
16. Canete J, Chaperon G, Fuentes R, Pérez J. Spanish Pre-trained BERT Model and Evaluation data. PML4DC at ICLR. 2020; 2020.
17. Wu S, Dredze M. Beto, bentz, becas: The Surprising Cross-lingual Effectiveness of BERT. arXiv preprint arXiv:1904.09077. 2019 Apr 19.
18. BertForTokenClassification, https://huggingface.co/transformers/model_doc/bert.htmlbertfortokenclassification
19. Hugging Face homepage, https://huggingface.co/
20. MedInfo 2015 Dataset, https://github.com/IBMMRL/medinfo2015
21. brat standoff format homepage, https://brat.nlplab.org/standoff.html

22. ://en.wikipedia.org/wiki/Inside%E2%80%93outside%E2%80%93beginning_(tagging)
23. NeuroNER, https://github.com/Franck-Dernoncourt/NeuroNER
24. BETO: Spanish BERT, https://github.com/dccuchile/beto
25. Spanish Unannotated Corpora, https://github.com/josecannete/spanish-corpora
26. Spanish Bert, https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased
27. PyTorch homepage, https://pytorch.org/
28. Keras homepage, https://keras.io/
29. https://github.com/fazlfrs/MUCIC-MEDDOPROF/blob/main/Flair%20arcitecture
30. MEDDOPROF homepage, https://temu.bsc.es/meddoprof/data/