Comparing Transformer-based NER approaches for analysing textual medical diagnoses

Marco Polignano, Marco de Gemmis and Giovanni Semeraro

University of Bari Aldo Moro
Via E. Orabona 4, 70125, Bari, Italy

Abstract

The automated analysis of medical documents has grown in research interest in recent years as a consequence of the social relevance of the thematic and the difficulties often encountered with short and very specific documents. In particular, this fervent area of research has stimulated the development of several techniques of automatic document classification, question answering, and name entity recognition (NER). Nevertheless, many open issues must be addressed to obtain results that are satisfactory for a field in which the effectiveness of predictions is a fundamental factor in order not to make mistakes that could compromise people's lives. To this end, we focused on the name entity recognition task from medical documents and, in this work, we will discuss the results we obtained by our hybrid approach. In order to take advantage of the most relevant findings in the field of natural language processing, we decided to focus on deep neural network models. We compared several configurations of our model by varying the transformer architecture, such as BERT, RoBERTa and ELECTRA, until we obtained a configuration that we considered the best for our goals. The most promising model was used to participate in the SpRadIE task of the annual CLEF (Conference and Labs of the Evaluation Forum). The obtained results are encouraging and can be of reference for future studies on the topic.

1. Introduction

Different definitions of *digital health* [1] are provided in the scientific literature, and many of them focus on the use of smart systems that enable the delivery of medical and health services directly to the patients. Those systems are grounded on the use of innovative technologies that not only allow providing innovative functionalities to users such as recommendations, monitoring services, and document archiving, but they also collect an enormous amount of data that needs to be automatically processed to be correctly dispatched to physicians, specialist doctors or the national health network.

Generally speaking, eHealth approaches could be grouped into two main macro-categories by considering their domain of application:

- **Wellness**: all those applications for fitness and to support a correct lifestyle;
- **Disease & Treatment Management**: all those systems for the management of a pathology, supporting all the phases from diagnosis to treatment and monitoring of the same.

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

marco.polignano@uniba.it (M. Polignano); marco.degemmis@uniba.it (M. de Gemmis);
giovanni.semeraro@uniba.it (G. Semeraro)
0000-0002-3939-0136 (M. Polignano); 0000-0002-2007-9559 (M. de Gemmis); 0000-0001-6883-1853
(G. Semeraro)

Both these two main kind of eHealth systems,have recently gained a large amount of new young consumers. eHealth systems and applications are quickly increasing, and the market of eHealth related systems and mobile application is expected to constantly grow of the 48.44% each year for the next four years [1]. The amount of user of eHalth platforms, which today is around 825 million worldwide, will cross the billion mark in the next four years. Unfortunately, despite the increasing amount of data in healthcare available, the percentage of those annotated and usable for supervised machine learning systems is still very small. As already stated in [2] this behavior is more frequent to observe in the medical domain than in others due to some critical limits about the ownership of sensitive data and the deep knowledge need for correctly annotating them. Only a few years ago, in Europe, the regulation regarding the use of sensitive user data (GDPR) [3] was defined and some guidelines regarding how to properly manage such data in a uniform manner among the member states of the European Union have been provided. Nevertheless, often a simple anonymization of data is not enough and a careful procedure of requesting consents for research purposes is necessary. These limitations make it necessary for artificial intelligence techniques that can make good use of the little data available, inheriting, where possible, knowledge from other similar application domains. Fortunately, in recent years, transfer learning has become a common practice to address various tasks of natural language processing. This opportunity has allowed the growth of several applications [4] in the field of eHealth including name entity recognition [5, 6, 7], the main topic of interest of this work. In particular, we decided to focus on NER's task regarding medical diagnoses in text format due to the substantial amount of available data provided for the SpRadIE 2021 [8, 9] competition co-located with the Conference and Labs of the Evaluation Forum 2021. In detail the main contribution of this work is about:

- the definition of an hybrid NER model for analysing textual medical diagnoses, based on a deep learning transformer based models [10] and conditional random field (CRF) [11];
- the comparison of different model configurations in order to detect the better performing transformer based model among BERT, BERT-mul, BETO, RoBERTa, XLM-RoBERTa, ELECTRA, ELECTRA-ES, ELECTRA-MED, BioBert, BioClinBERT.

Following we will present some relevant work useful for better understanding the technologies used in the model described in Chapter 3. In Chapter 4, we will discuss the configurations performed and the results obtained, until we present the one that performed best on the data at hand. We will conclude with our considerations and directions for future work.

2. Related Work

In recent years, natural language processing has gained a lot of attention, quickly becoming one of the most trending research areas. In particular, technologies based on deep learning have made it possible to learn models on a vast amount of data that are highly efficient and versatile in many scenarios of text analysis and classification. The most famous of such transfer learning approaches is BERT [12], a transformer model that uses a bidirectional encoding model

[1] https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-era-of-exponential-improvement-in-healthcare

to generate a text representation that takes into account the context of the use of each term. Its applicability has been demonstrated in several works by adopting it for facing tasks as sentiment analysis [13, 14, 15], question-answering, name-entity recognition, text summarization, and many more [16]. The use of BERT is not only limited to the English language. Different versions of it, trained on different sets of data, have been presented to make it suitable in scenarios where it needs to work correctly with multilingual text (BERT multilang). Still, it has also been trained on many single-language data in order to correctly work on specific languages different than English, such as Italian (AlBERTo) [17] and Spanish (BETO) [18]. Different specializations of BERT have been proposed for a specific domain of use, including the medical one. Indeed, BioBERT [19] has been trained on biomedical literature, making it suitable for many tasks of text analysis in the biological domain. BioClinicalBERT [20] starting from BioBert specialized the vocabulary on clinical terms. Taking inspiration from the BERT architecture, other models based on transformer have been realized. Among these, we would like to cite RoBERTa [21], a model that evolves BERT with the aim of making the learning phase less demanding from the computational point of view while maintaining performances comparable to those of the original approach. In addition, to ensure its easy use in contexts where more than one language is used in the text, the XLM-RoBERTa [22] version was created. Similar to RoBERTa, ELECTRA [23] follows the BERT architecture by replacing its training model. Instead of using a masking approach, it uses another neural network that attempts to trick the model during the training phase by replacing random tokens with fake tokens. This strategy allows ELECTRA to train a more effective final language understanding model.

The idea of using BERT as the basis for a name entity recognition task in the medical domain is not entirely new. The name entity recognition (NER) task consists of identifying n-grams of a textual content that refers to entities or names relevant for the domain of application. It is common to find in literature NER systems able to detect in text locations, organizations, companies, person names but having enough training data it is possible to train those systems on every entity of interest. The task is a subproblem of information extraction and involves processing structured and unstructured documents [24]. Since it does not require the annotation of strictly personal and sensitive data, but generic texts are sufficient, this task is widely studied also in the literature even in the medical domain of application. Mao et al. [25] use BERT and the Conditional Random Field (CRF) models to identify the name of hospitals, medical staff, and territory in medical records. Similarly, Xue wt al. [26] fine-tuned the BERT model in order to detect entities and relations in Chinese medical documents. Vunikili et al. [27] explored the use of Bidirectional Encoder Representations from Transformers (BERT) based contextual embeddings trained on general domain Spanish text to extract tumor morphology from clinical reports written in Spanish. In a similar way, its not difficult to find attempts that use the RoBERTa model [28], ELECTRA [29] or BioBERT [30]. All those work fall in not evaluating the efficacy of different transformer models before choosing one to use in their models. On the contrary, we focus on developing a hybrid model that merges an arbitrary transformer based architecture with a classic CRF layer. In this way, we will not limit ourselves to use with blinded eyes only a single transformer based model. Still, we will evaluate different ones in order to identify the one that best performing in the medical domain and, in particular, on documents written in Spanish as provided us by the organizers of the SpRadIe challenge [8, 9].

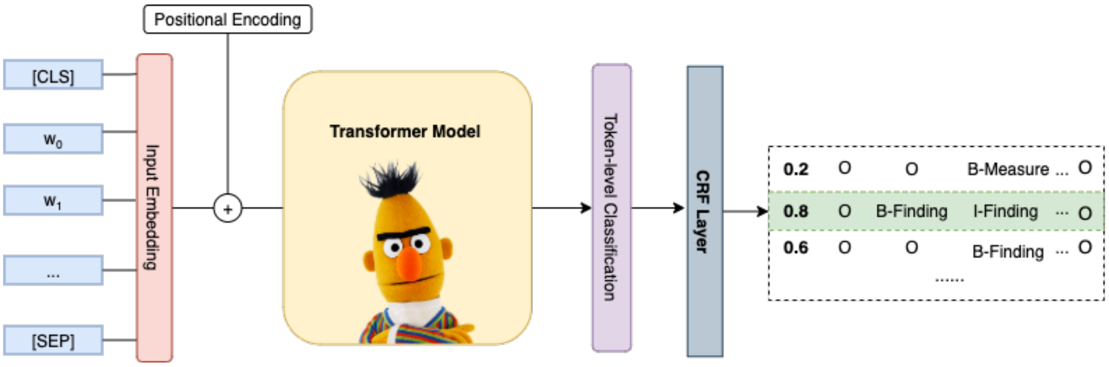

Figure 1: Hybrid architecture proposed

3. Transformer based NER Model

The hybrid model we propose is inspired by a common architecture largely used for dealing with a NER task. In particular, we are looking at the architecture proposed by Huang [31] that is composed by the concatenation of a Long Short Term Memory (LSTM) model and a Conditional Random Field layer (CRF) [11]. This approach has been demonstrated to be very effective for dealing with the identification of entities and relations into the text, and consequently, it has been our starting point. We decided to combine a transformer based model with a CRF layer, similarly to the approach discussed in [32].

The architecture of our proposed hybrid model is showed in Fig. 1. The first step consists of pre-processing the text to make it suitable for the transformer based model. In particular, we split the text into sentences, and we added special tokens at the beginning and end of the sentence. The [CLS] token is used for indicating the beginning of the sentence, [SEP] for indicating its end, and [PAD] tokens are added before the [SEP] for making the input length uniform. For each token, the transformer based model requires input *ids*, a sequence of integers identifying each input token to its index number in the tokenizer vocabulary provided by the specific pre-trined model. Consequently, the tokens are transformed in their numerical counterpart in order to be properly encoded using token embeddings, sentence embeddings, and positional vectors [33]. The formatted input of length *n* is denoted as $w =< w_1, w_2, .., w_n >$, and the corresponding tag sequence is denoted as $t =< t_1, t_2, ..., t_n >$. The tag annotation schema used for the model is *BIO*. The schema allows to learn a NER model able to identify tokens in text that can represent one of the following tag: *(i)* the *Beginning* (B) of the entity; *(ii)* the word *Inside* (I) the entity; *(iii)* words *Outside* (O) entities [34].

When the data is ready, it is given as input to the chosen transformer based model. Each one of them has its internal architecture, but all of them allow us to obtain from an embedding representation of each token present in the input sentence. Generally speaking, *transformer* is an architecture for transforming one sequence into another one by using two modules: the encoder and the decoder. Both encoder and decoder are composed of blocks that can be stacked on top of each other multiple times in order to obtain the desired depth of the network. Each block is mainly composed of a multi-head attention layer, followed by a normalization layer and a feed-forward layer. Using only the encoder module, we can obtain a word embedding representation of each token of the input after any possible arbitrary number of encoding blocks. In particular, BERT, in its basic version, is trained on a Transformer network made of only 12 encoding blocks, 768 dimensional states and 12 heads of attention for a total of 110M of parameters trained on BooksCorpus [35] and Wikipedia English for 1M of steps. The learning phase is performed by scanning the span of text in both directions, from left to right and from right to left, as was already done in Bidirectional LSTM. Moreover, BERT uses a "masked language model": during the training, random terms are masked in order to be predicted by the net. Jointly, the network is also designed to potentially learn the next span of text from the one given in input. These peculiarities allow BERT to be the current state of the art language understanding model.

By following the logical flow of our hybrid model, we used the embedding generated by the transformer based model to perform a token-level classification task. In particular, an hidden layer has been added on top of the stack of the model in order to obtain a prediction matrix $\mathbb{P} \in \mathbb{R}^{n \times k}$ (i.e. one prediction vector for each on of the n tokens) for the given input sequence. Formally speaking, the classification layer projects each token's encoded representation to the tag space $\mathbb{R}^H \mapsto \mathbb{R}^k$, where k is the number of tags which varies in accordance with the number of classes and the tagging scheme. The CRF layer consequently learns only the transition probability of the output labels, $A \in \mathbb{R}^{K+2 \times K+2}$ where +2 indicates one tag each for start and end marker. The matrix A is such that $A_{i,j}$ represents the score of transitioning from tag i to tag j [36]. For an input sequence $x =< x_1, ..., x_n >$ and a sequence of tag predictions $y =< y_1, ..., y_n >, y_i \in 1, ..., k$, the score of the sequence is defined as [11]:

$$S(x, y) = \sum_{i=o}^{n} A_{y_i, y_{i+1}} + \sum_{i=o}^{n} \mathbb{P}_{i, y_i} \tag{1}$$

The objective function is the maximum likelihood of the probability distribution denoted as:

$$ln(\ p(y|x)\ ) = S(x, y) - ln \left(\sum_{y' \in y} e^{S(x, y')} \right) \tag{2}$$

We use $y^\star$ to represent the most likely tag sequence of x that will be considered as output of the model:

$$y^\star = \arg\max_{y' \in y} (S(x, y')) \tag{3}$$

During the evaluation, the most likely sequence is obtained by the Viterbi algorithm. The output obtained is re-processed to group back sub-tokens obtained due to the tokenization approach used by transformers models [37]. Only the tag of the first sub-token is considered for the final tag sequence.

4. Experimental setting and discussion of results

4.1. Implementation

The model architecture has been implemented by using the Python programming language. In particular we used the Pytorch [38] syntax and the Transformer Huggingface library [39]. The CRF layer used is the one provided by the library TorchCRF[2]. We set AdamW as optimizer; a modified version of the Adam optimizer that uses a weight decay factor. Consequently we set a weight decay of 0.01, a learning rate of 1e-5, a number of epochs equal to 500, a number of steps equal to 2000, and a batch size of 16. The code has been run on Google Colab environment[3] by using a Tesla T4 Nvidia GPU with 16GB of RAM.

4.2. Dataset

The dataset is provided by the organizers of the SpRadIe 2021 [8, 9] competition. It consists of 513 ultrasonography reports provided by a pediatric hospital in Argentina. Reports are unstructured and have abundance of orthographic and grammatical errors and have been anonymized in order to remove patient IDs, and names and the enrollment numbers of the physicians. Reports were annotated by clinical experts and then revised by linguists. Annotation guidelines and training were provided for both rounds of annotations. Automatic classifiers will be expected to perform well in those cases where human annotators have strong agreement, and worse in cases that are difficult for human annotators to identify consistently. Annotations are provided in brat format. More details are reported in [2]. The annotated dataset is constituted as follows: *training set* (175 reports); *Same-sample development set* (47 reports); *Held-out development set* (45 reports); *Held-out test set* (207 reports).

The following entities are distinguished:

- **Anatomical Entity**: entities corresponding to an anatomical part, for example breast (pecho), liver (hígado), right thyroid lobe (lóbulo tiroideo derecho);
- **Finding**: a pathological finding or diagnosis, for example: cyst, cyanosis;
- **Location**: it refers to a location in the body. Examples of locations are: walls, cavity, longitudinal, frontal, occipital, cervicodorsolumbosacra, lumbosacral, intracanalar, subcutanea;
- **Measure**: expression indicating a measure;
- **Type of Measure**: expression indicating a type of measure;
- **Degree**: It indicates the degree of a finding or some other property of an entity, for example, "leve", "levemente" (slight), "mínimo" (minimal);
- **Abbreviation**: acronyms or abbreviations to indicate a medical concept;
- **Negation**: hedge cues indicating negation;
- **Uncertainty**: hedge cues indicating a probability (not a certainty) that some finding may be present in a given patient.

[2] https://pypi.org/project/TorchCRF/

[3] https://colab.research.google.com/

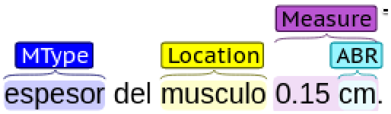

Figure 2: Example of annotated piece of diagnosis.

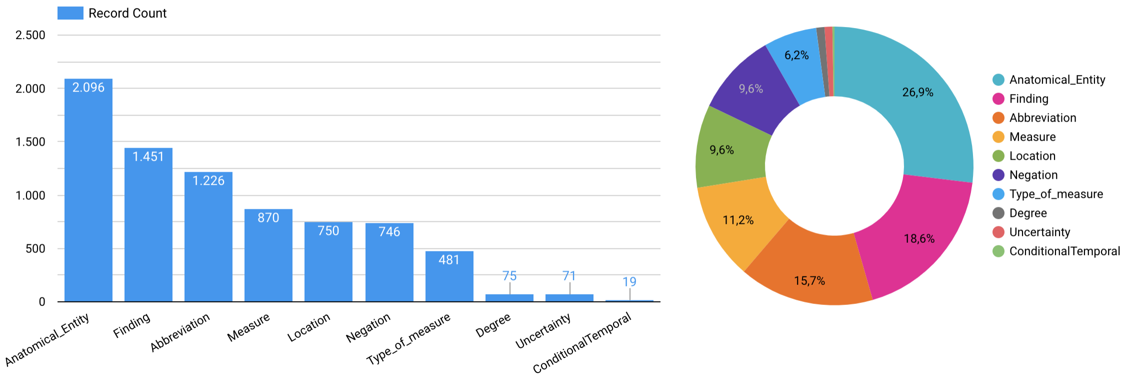

Figure 3: Distribution of tag annotations among the training dataset.

- **Conditional Temporal**: hedge cues indicating that something occurred in the past or may occur in the future.

In Fig. 2 is reported an example of annotated piece of diagnosis. In Fig. 3 it is showed the distribution of different entities among the training dataset. The entity type *Finding* is particularly challenging, as it presents great variability in its textual forms. It ranges from a single word to more than 10 words in some cases, and comprising all kinds of phrases. However, this is also the most informative type of entity for the potential users of these annotations. Other challenging phenomena are the regular polysemy observed between *Anatomical entities* and *Locations*, and the irregular uses of *Abbreviations*. Moreover the same token can be labeled with different tags at the same time making the automatic annotation a complex task.

4.3. Setting and metrics

In order to train the proposed hybrid NER model, we merged the train set with the *Same-sample development set* resulting in a single set of 222 diagnoses (i.e. 18428 tokens obtained by using a white space splitting). Consequently we used only the *Held-out development set* for evaluating the performances of the model on 45 diagnoses (i.e. 5090 tokens). The performances of the models are evaluated by using classic metrics such as *Precision*, *Recall* and *F1-score*. As a consequence of the unbalancing of the dataset, we decided to consider the *micro average* of results obtained for each tag category.

We chose ten different transformer models imported through the Transformer HuggingFace library. They are listed in the following:

- *bert-base-uncased* (**BERT**): it is the standard base version of BERT (12-layer, 768-hidden, 12-heads, 168M parameters) with an uncased English vocabulary;
- *bert-base-multilingual-uncased* (**BERT-mul**): it is the base version of BERT trained on lower-cased text in the top 102 languages;
- *dccuchile/bert-base-spanish-wwm-uncased* (**BETO**): BETO is a BERT model trained on a big Spanish corpus. BETO is of size similar to a BERT-base and was trained with the Whole Word Masking technique;
- *roberta-base* (**RoBERTa**): it is a standard RoBERTa transformers model pretrained on a large corpus of English data in a self-supervised fashion;
- *xlm-roberta-base* (**XLM-RoBERTa**): it is a multilingual RoBERTa model trained on 100 different languages.
- *google/electra-small-discriminator* (**ELECTRA**): is a new pretraining approach which trains two transformer models: the generator and the discriminator. It has been trained on English data;
- *skimai/electra-small-spanish* (**ELECTRA-ES**): it is the ELECTRA model trained on Spanish data;
- *fspanda/electra-medical-small-discriminator* (**ELECTRA-MED**): ELECTRA model trained on English scientific paper in medical domain published on PubMed and PubMEdCentral;
- *emilyalsentzer/Bio_ClinicalBERT* (**BioCliBERT**): The Bio_ClinicalBERT model is initialized with BERT-Base and it was trained on all notes from MIMIC III, a database containing electronic health records from ICU patients at the Beth Israel Hospital in Boston, MA;
- *seiya/oubiobert-base-uncased* (**BioBERT**): Bidirectional Encoder Representations from Transformers for Biomedical Text Mining by Osaka University (ouBioBERT) is a language model based on the BERT-Base architecture.

4.4. Discussion of results

The results we obtained are reported in Table 1-3. It is possible to observe that the best performing model, by considering F1-score as main metric, is *BERT* in its *base* version trained on a *multilingual* dataset. The model has obtained a micro F1-score of 0.69632 that outperforms the competitors in a range from 0.59% to 17.94%. The most significant difference is obtained by comparing the results with them of the ELECTRA-MED base model. In particular, these results are unexpected low and required further investigation for understanding their causes. On the contrary, the model based on BETO is performing particularly well obtaining a results comparable with the one obtained by using BERT multilingual. The differences among the two models are very low, probably because they share the same internal architecture, and both of them are based on a vocabulary able to deal with Spanish terms. If from one end BERT multilanguage is trained on many different languages simultaneously creating some noise on data, on the other end, BETO is trained on a smaller set of data losing then in accuracy. The two issues of the two models are able to produce results somehow very close to each other. If we observe the results obtained for precision (Table 1) and recall (Table 2), we can keep an interesting behavior of the two models. BETO, even if for a small gap, obtains better results for precision than BERT-mul. On the contrary, considering the recall, BERT-mul is the best.

Still looking at the precision values, we can observe that, in general, the chosen models show prediction difficulties for the tag classes "Conditional Temporal", "Uncertainly", "Degree" and "Location". In particular, such behavior could derive from the specificity of such classes for the medical field. It is, in fact, evident as for classes of more general character like "Negation", "Type of Measure", "Abbreviation" all the models perform medium-well. The only exception is the class "Anatomical Entity" that even if much specific of the medical domain turns out to be classified correctly in nearly the totality of the models. This could derive from the wide everyday use that is made of terms regarding the parts of the body also in the common language. Looking at the recall, we can see that some classes were easier to find than others, such as "Anatomical Entity", "Negation", "Abbreviation", "Uncertainly". For the first three classes, obtaining good results of precision and recall, we can say that the models are able not only to be precise in the classification but also to identify them correctly in the text. For the class "Uncertainty", however, the models tend to overestimate its presence in the sentences because it has a good recall value and low precision.

If we want to understand if the training language of the model is an issue for a correct classification of individual tokens, we can observe the differences between monolingual and multilingual models. Generally speaking, we can observe that multilingual and Spanish-trained models are better performing than their counterpart monolingual. This result was expected as a consequence of the language of the dataset. It is obvious that a model able to manage words in the dataset language is more efficient than a model trained on English data also if the SentencePiece tokenizer, typical of transformer-based architectures, is alleviating these differences through the use of sub-tokens. If we focus on domain-specific models (BioCliBERT and BioBERT), we can observe that they perform in line with their non-specialized counterparts. The difference in F1-score among BERT-base, BioCliBERT, and BioBERT is only around 0.01. This suggests that the training on the medical domain is not correctly supporting the task of NER for clinical diagnoses, if, as in this case, the model is pre-trained on a different language of the testing dataset (i.e. English). In order for it to be possible to draw conclusions about the usefulness of a model trained on medical data for the specific NER task, it would be necessary as a future development to train a new model with medical data in the Spanish language.

By focusing on the differences in transformer architectures, we can affirm that BERT is commonly performing better than others even if XLM-RoBERTa and ELECTRA are, however, obtaining results not very far from the BERT counterpart. Finally, all the models could not correctly classify *Conditional Temporal* tags, but this problem is almost certainly related with the size of this class in the dataset. Indeed, only 19 instances are available in the training set and only one instance was present in the set used for the analysis.

5. SpRadIe submission

In light of the results, decisions were made to submit to the SpRadIe competition. In particular, the first decision taken concerns the configuration of the hybrid model to use. The obvious choice would have fallen on using BERT-mul as the transformer-based model. On the contrary, XLM-

Table 1
Results obtained for the chosen transformer models considering the ***Precision***.

	BERT	**BERT-MUL**	**BETO**	**RoBERTa**	**XLM-RoBERTa**	**ELECTRA**	**ELECTRA-ES**	**ELECTRA-MED**	**BioCliBERT**	**BioBERT**
Abbreviation	0.79412	0.73874	0.69919	0.66071	0.6774	0.66667	0.71585	0.36111	0.75385	0.81818
Anatomical Entity	0.75776	0.77193	0.76630	0.77844	0.7794	0.74859	0.77273	0.64306	0.79655	0.78049
Conditional Temporal	0.00000	0.00000	0.00000	0.00000	0.0000	0.00000	0.00000	0.00000	0.25000	0.00000
Degree	0.33333	0.45455	0.55556	0.50000	0.5294	0.38095	0.57895	0.27778	0.31579	0.50000
Finding	0.50292	0.63725	0.63462	0.52688	0.5418	0.50581	0.57285	0.56757	0.47181	0.56836
Measure	0.58549	0.71721	0.67932	0.69515	0.7111	0.63687	0.73092	0.66818	0.68456	0.60870
Negation	0.89394	0.89888	0.86408	0.86357	0.8701	0.86765	0.91429	0.80723	0.86667	0.89773
Type of measure	0.89394	0.79167	0.70492	0.80625	0.8421	0.93103	0.74627	0.39130	0.85714	0.82927
Uncertainty	0.50000	0.42857	0.33333	0.50000	0.5000	0.50000	0.26667	0.23077	0.42857	0.28571
Location	0.46400	0.51250	0.63846	0.43810	0.4806	0.54082	0.48503	0.44037	0.44231	0.41104
microAvg	**0.62489**	**0.69462**	**0.69645**	**0.65649**	**0.6635**	**0.63983**	**0.67321**	**0.56189**	**0.63311**	**0.64984**

RoBERTa was chosen. This decision is motivated by our feeling that the validation dataset was not large enough for containing a representative subset of all the possible variations in data we could have in validation phase, especially regarding the less numerous classes. Indeed, a model best performing on validation set is not always the best performing on test set, especially when the validation set is very small. Therefore, it was decided to opt for a model with intermediate performance in F1-score, following the idea that a less specialized model could better deal the

Table 2
Results obtained for the chosen transformer models considering the ***Recall***.

	BERT	**BERT-mul**	**BETO**	***RoBERTa***	***XLM-RoBERTa***
Abbreviation	0.87097	0.85417	0.81132	0.86071	0.8630
Anatomical Entity	0.80795	0.78338	0.81503	0.80069	0.8179
Conditional Temporal	0.00000	0.00000	0.00000	0.00000	0.0000
Degree	0.37500	0.58824	0.55556	0.50286	0.5294
Finding	0.42786	0.53830	0.52800	0.41294	0.4467
Measure	0.59162	0.75431	0.68511	0.59553	0.6038
Negation	0.85507	0.83333	0.85577	0.83869	0.8481
Type of measure	0.85507	0.76000	0.76786	0.75152	0.7619
Uncertainty	1.0	0.75000	0.75000	0.66667	1.0
Location	0.65169	0.75926	0.72807	0.64750	0.6526
microAvg	**0.63077**	**0.69803**	**0.68801**	**0.63621**	**0.6404**

	ELECTRA	**ELECTRA-ES**	**ELECTRA-MED**	**BioCliBERT**	**BioBERT**
Abbreviation	0.83871	0.85065	0.75581	0.87500	0.70130
Anatomical Entity	0.87748	0.72070	0.67964	0.81338	0.78528
Conditional Temporal	0.00000	0.00000	0.00000	1.0	0.00000
Degree	0.50000	0.55000	0.29412	0.46154	0.41176
Finding	0.43284	0.51805	0.43659	0.43923	0.45887
Measure	0.59686	0.71094	0.66516	0.58286	0.58065
Negation	0.85507	0.80672	0.72043	0.83871	0.89773
Type of measure	0.77143	0.64935	0.73469	0.80000	0.77273
Uncertainty	1.0	0.80000	1.0	1.0	0.66667
Location	0.59551	0.61364	0.45714	0.58974	0.67000
microAvg	0.64530	0.65794	0.58129	0.63252	0.62697

variability on new unseen data. In particular, we decided to use this strategy as a regularization one, similarly to what is commonly performed with early-stopping for reducing overfitting.

Our second decision concerned the number of models to be learned. In particular, our study of the dataset revealed that the same token could be annotated with more than one tag. For example, the term "cm" could be both parts of a "Measure" tag and an abbreviation tag. To this end, we decided to create four groups of tags with as little overlap as possible and to create

Table 3
Results obtained for the chosen transformer models considering the ***F1-score***.

	BERT	**BERT-mul**	**BETO**	**RoBERTa**	**XLM-RoBERTa**
Abbreviation	0.83077	0.79227	0.75109	0.74071	0.7590
Anatomical Entity	0.78205	0.77761	0.78992	0.77513	0.7982
Conditional Temporal	0.00000	0.00000	0.00000	0.00000	0.0000
Degree	0.35294	0.51282	0.55556	0.51053	0.5294
Finding	0.46237	0.58361	0.57642	0.46865	0.4896
Measure	0.58854	0.73529	0.68220	0.62508	0.6531
Negation	0.87407	0.86486	0.85990	0.84060	0.8590
Type of measure	0.87407	0.77551	0.73504	0.75385	0.8000
Uncertainty	0.66667	0.54545	0.46154	0.57143	0.6667
Location	0.54206	0.61194	0.68033	0.50979	0.5536
microAvg	**0.62782**	**0.69632**	**0.69220**	**0.64619**	**0.6517**

	ELECTRA	**ELECTRA -ES**	**ELECTRA -MED**	**BioCliBERT**	**BioBERT**
Abbreviation	0.74286	0.77745	0.48872	0.80992	0.75524
Anatomical Entity	0.80793	0.74581	0.66084	0.80488	0.78287
Conditional Temporal	0.00000	0.00000	0.00000	0.40000	0.00000
Degree	0.43243	0.56410	0.28571	0.37500	0.45161
Finding	0.46649	0.54408	0.49354	0.45494	0.50778
Measure	0.61622	0.72079	0.66667	0.62963	0.59434
Negation	0.86131	0.85714	0.76136	0.85246	0.89773
Type of measure	0.84375	0.69444	0.51064	0.82759	0.80000
Uncertainty	0.66667	0.40000	0.37500	0.60000	0.40000
Location	0.56684	0.54181	0.44860	0.50549	0.50951
microAvg	**0.64255**	**0.66549**	**0.57143**	**0.63282**	**0.63820**

a classifier for each. In particular, the four groups are composed as follows: (i) Finding; (ii) Anatomical Entity, Measure, Degree; (iii) Location, Negation, Type of Measure; (iv) Abbreviation, Uncertainty, Conditional Temporal. In the prediction phase for each tag, we obtained the results of each classifier and chose the one with the highest probability.

If considering the results we obtained from the task organizers reported in Table 4, we can observe that the results are lower than expected, especially for some specific tag class. In

Table 4
Results obtained for the SpRadIe submission considering the "exact" matching and the micro F1-score.

	Abbr.	Anat. Entity	Cond. Temp	Degree	Finding	Location	Measure	Negation	Type measure	Unc.
SWAP	0,49	0,52	0,00	0,48	0,44	0,31	0,51	0,76	0,37	0,17

particular, "Type of measure" is resulted 116.2% lower than what obtained in the preliminary evaluation. Similar values are obtained also for the classes *Uncertainity*, *Abbreviation* and *Anatomical Entity*. In any case it will be considered future work to investigate the reasons for these poorer than expected performances as soon as the annotated test set is released.

6. Conclusion

Automated analysis of textual documents in medical settings is still a research challenge that will be a topic of discussion in the coming years. Recently, the amount of medical data available has grown rapidly due to the deployment of numerous applications and systems to support the patient and wellness enthusiast. Nevertheless, the amount of annotated data is still meager due to their privacy issues. Recently, however, attempts have been made to overcome this problem with transfer learning strategies based on large pre-trained models that use deep learning techniques. Among these, those based on transformers have proved to be the most reliable and versatile. In the wake of these scientific innovations, it was decided to address the problem of name entity recognition in the medical field by exploiting these approaches. Specifically, we proposed a hybrid model that combines the power of representation of transformer models with the predictive power of probabilistic conditional random field models. The model was evaluated in several of its combinations, varying the transformer model at the core of its architecture. Specifically, the models BERT, BERT-mul, BETO, RoBERTa, XLM-RoBERTa, ELECTRA, ELECTRA-ES, ELECTRA-MED, BioBert, BioClinBERT were evaluated. The experimentation was performed on data provided by the SpRadIE 2021 [8, 9] competition co-located with the Conference and Labs of the Evaluation Forum 2021. In particular, the training set counted 222 medical diagnoses in the Spanish language appropriately annotated by experts in the field. The preliminary analysis was carried out on a validation set extracted from a data collection different from the training set. Specifically, this portion of data contained 45 medical diagnoses, also in Spanish. The results obtained showed excellent effectiveness of the transformer-based models trained on multilingual or Spanish data. Specifically, BERT-base in its multilingual version and BETO proved to be the most suitable for the task of NER for medical data. The results of this analysis were used to make implementation decisions regarding the system to be submitted to the SpRadIE 2021 competition. Specifically, it was decided to perform a submission using the transformer based XLM-RoBERTa model with the hope of obtaining satisfactory results through the use of a model with average performance in the preliminary analysis phase. This decision was made to avoid basing our decisions solely on a portion of the data that was not expressive enough due to its size. The results obtained proved to be lower than expected, but further study and investigation will be carried out upon release of the annotated test set.

7. Acknowledgment

This work has been supported by Apulia Region, Italy through the project ”Un Assistente Dialogante Intelligente per il Monitoraggio Remoto di Pazienti” (Grant n. 10AC8FB6) in the context of ”Research for Innovation - REFIN”.

References

[1] H. Oh, C. Rizo, M. Enkin, A. Jadad, What is ehealth?: a systematic review of published definitions, World Hosp Health Serv 41 (2005) 32–40.

[2] V. Cotik, D. Filippo, R. Roller, H. Uszkoreit, F. Xu, Annotation of entities and relations in spanish radiology reports., in: RANLP, 2017, pp. 177–184.

[3] P. Voigt, A. Von dem Bussche, The eu general data protection regulation (gdpr), A Practical Guide, 1st Ed., Cham: Springer International Publishing 10 (2017) 3152676.

[4] H. Suominen, L. Kelly, L. Goeuriot, A. Névéol, L. Ramadier, A. Robert, E. Kanoulas, R. Spijker, L. Azzopardi, D. Li, et al., Overview of the clef ehealth evaluation lab 2018, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2018, pp. 286–301.

[5] R. M. R. Zavala, P. Martínez, I. Segura-Bedmar, A hybrid bi-lstm-crf model for knowledge recognition from ehealth documents., in: TASS@ SEPLN, 2018, pp. 65–70.

[6] E. Andrés, O. Sainz, A. Atutxa, O. L. de Lacalle, Ixa-ner-re at ehealth-kd challenge 2020: Cross-lingual transfer learning for medical relation extraction, in: Proceedings of the Iberian Languages Evaluation Forum co-located with 36th Conference of the Spanish Society for Natural Language Processing, IberLEF@ SEPLN, volume 2020, 2020.

[7] N. Garcıa-Santa, K. Cetina, Fle at clef ehealth 2020: Text mining and semantic knowledge for automated clinical encoding, in: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings, 2020.

[8] H. Suominen, L. Goeuriot, L. Kelly, L. A. Alemany, E. Bassani, N. Brew-Sam, V. Cotik, D. Filippo, G. González-Sáez, F. Luque, et al., Overview of the clef eHealth evaluation lab 2021, in: CLEF 2021 - 12th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, 2021.

[9] V. Cotik, L. A. Alemany, D. Filippo, F. Luque, R. Roller, J. Vivaldi, A. Ayach, F. Carranza, L. D. Francesca, A. Dellanzo, et al., Overview of CLEF eHealth Task 1 - SpRadIE: A challenge on information extraction from Spanish Radiology Reports, in: CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, 2021.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[11] J. Lafferty, A. McCallum, F. C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001).

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human

Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423.

[13] P. Basile, V. Basile, D. Croce, M. Polignano, Overview of the evalita 2018 aspect-based sentiment analysis task (absita), volume 2263, 2018. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85058658278&partnerID=40&md5=9719e0512649279a6ed46a8262f050dc, cited By 6.

[14] L. de Mattei, G. de Martino, A. Iovine, A. Miaschi, M. Polignano, G. Rambelli, Ate absita @ evalita2020: Overview of the aspect term extraction and aspect-based sentiment analysis task, volume 2765, 2020. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85097555110&partnerID=40&md5=2122e26ed7367e5a7c40642835591903, cited By 5.

[15] K. Florio, V. Basile, M. Polignano, P. Basile, V. Patti, Time of your hate: The challenge of time in hate speech detection on social media, Applied Sciences (Switzerland) 10 (2020). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85087763770&doi=10.3390%2fAPP10124180&partnerID=40&md5=d4c6b8ba193ed062299bc19f02cb462e. doi:10.3390/APP10124180, cited By 7.

[16] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, Glue: A multi-task benchmark and analysis platform for natural language understanding, arXiv preprint arXiv:1804.07461 (2018).

[17] M. Polignano, P. Basile, M. De Gemmis, G. Semeraro, V. Basile, Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets, in: 6th Italian Conference on Computational Linguistics, CLiC-it 2019, volume 2481, CEUR, 2019, pp. 1–6.

[18] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[19] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.

[20] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, M. McDermott, Publicly available clinical bert embeddings, arXiv preprint arXiv:1904.03323 (2019).

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[22] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).

[23] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, arXiv preprint arXiv:2003.10555 (2020).

[24] A. Mansouri, L. S. Affendey, A. Mamat, Named entity recognition approaches, International Journal of Computer Science and Network Security 8 (2008) 339–344.

[25] J. Mao, W. Liu, Hadoken: a bert-crf model for medical document anonymization., in: IberLEF@ SEPLN, 2019, pp. 720–726.

[26] K. Xue, Y. Zhou, Z. Ma, T. Ruan, H. Zhang, P. He, Fine-tuning bert for joint entity and relation extraction in chinese medical text, in: 2019 IEEE International Conference on

Bioinformatics and Biomedicine (BIBM), IEEE, 2019, pp. 892–897.

[27] R. Vunikili, N. SH, G. Marica, O. Farri, Clinical ner using spanish bert embeddings, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.

[28] K. S. Kalyan, S. Sangeetha, Want to identify, extract and normalize adverse drug reactions in tweets? use roberta, arXiv preprint arXiv:2006.16146 (2020).

[29] I. B. Ozyurt, On the effectiveness of small, discriminatively pre-trained language representation models for biomedical text mining, in: Proceedings of the First Workshop on Scholarly Document Processing, 2020, pp. 104–112.

[30] X. Yu, W. Hu, S. Lu, X. Sun, Z. Yuan, Biobert based named entity recognition in electronic medical record, in: 2019 10th International Conference on Information Technology in Medicine and Education (ITME), IEEE, 2019, pp. 49–52.

[31] Z. Huang, W. Xu, K. Yu, Bidirectional lstm-crf models for sequence tagging, arXiv preprint arXiv:1508.01991 (2015).

[32] M. Liu, Z. Tu, Z. Wang, X. Xu, Ltp: A new active learning strategy for bert-crf based named entity recognition, arXiv preprint arXiv:2001.02524 (2020).

[33] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[34] L. Ratinov, D. Roth, Design challenges and misconceptions in named entity recognition, in: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009), 2009, pp. 147–155.

[35] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 19–27.

[36] R. Panchendrarajan, A. Amaresan, Bidirectional lstm-crf for named entity recognition., in: PACLIC, 2018.

[37] T. Kudo, J. Richardson, Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, arXiv preprint arXiv:1808.06226 (2018).

[38] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, arXiv preprint arXiv:1912.01703 (2019).

[39] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Huggingface's transformers: State-of-the-art natural language processing, arXiv preprint arXiv:1910.03771 (2019).