Deep learning to extract Breast Cancer diagnosis concepts

Conference Paper · July 2022

DOI: 10.1109/CBMS55023.2022.00010

6 authors, including:

Oswaldo Solarte
Universidad del Valle (Colombia)
29 PUBLICATIONS **100** CITATIONS
SEE PROFILE

Maria Torrente
Hospital Universitario Puerta de Hierro-Majadahonda
61 PUBLICATIONS **560** CITATIONS
SEE PROFILE

Mariano Provencio
Hospital Universitario Puerta de Hierro-Majadahonda
978 PUBLICATIONS **21,214** CITATIONS
SEE PROFILE

Ernestina Menasalvas
Universidad Politécnica de Madrid
259 PUBLICATIONS **2,631** CITATIONS
SEE PROFILE

Deep learning to extract Breast Cancer diagnosis concepts

1st Oswaldo Solarte-Pabon
Centro de Tecnología Biomédica
Universidad Politécnica de Madrid
Madrid, Spain
oswaldo.solartep@alumnos.upm.es

2nd Maria Torrente
Hospital Universitario Puerta de Hierro
Madrid, Spain
maria.torrente@salud.madrid.org

3rd Alvaro Garcia-Barragán
Centro de Tecnología Biomédica
Universidad Politécnica de Madrid
Madrid, Spain
alvaro.gbarragan@upm.es

4nd Mariano Provencio
Hospital Universitario Puerta de Hierro
Madrid, Spain
mariano.provencio@salud.madrid.org

5nd Ernestina Menasalvas
Centro de Tecnología Biomédica
Universidad Politécnica de Madrid
Madrid, Spain
ernestina.menasalvas@upm.es

6ndVictor Robles
Centro de Tecnología Biomédica
Universidad Politécnica de Madrid
Madrid, Spain
victor.robles@upm.es

***Abstract*—The wide adoption of electronic health records (EHRs) provides a potential source to support clinical research. The Bidirectional Encoder Representations from Transformers (BERT) has shown promising results in extracting information in the biomedical domain, including the cancer field. However, one of the challenges in the cancer domain is annotating resources to support information extraction. In this paper, we will show how models trained in a lung cancer corpus can be used to extract cancer concepts even in other cancer types. In particular, we will show the performance of BERT models on breast cancer data that was not used to train the models. Results are very promising as they show the possibility of applying deep learning-based models to predict cancer concepts in a different dataset to the one they were trained on, representing a considerable save of time and resources.**

***Index Terms*—Natural Language Processing (NLP), Information extraction, Cancer Diagnosis extraction, Breast cancer.**

I. Introduction

Cancer remains one of the main public health problems, ranked as the second leading cause of death globally. In particular, breast cancer is the most common cancer in women worldwide and the second most common cancer overall [1]–[3]. The process of diagnosing and treating cancer patients generates a huge amount of information. Physicians register this information in Electronic Health Records (EHR) using clinical notes written in narrative form. Extracting and mining information from EHR is crucial to support clinical and epidemiological studies in the cancer domain [4]. However, obtaining this information automatically is a challenge due to the unstructured nature of clinical notes.

In the oncology domain, one of the first challenges is to extract information related to cancer diagnosis [5]. The diagnosis is a crucial factor for defining treatment plans for each patient and for evaluating the patient outcomes [6], [7]. In clinical notes, mentions of cancer vary from very precise (exact description of the tumor) to more general ones, for example, when one refers to the family oncological antecedents of the patient. Extracting the cancer diagnosis includes obtaining the most specific tumor mention, the stage of the tumor, and the date of the diagnosis. These pieces of information are crucial for relating them to treatments and survival rates [8]. Several Natural language processing (NLP) approaches have been proposed to extract information related to cancer diagnosis, including rule-based [9] and machine learning-based approaches [10].

Recently, deep learning-based approaches have shown their feasibility in obtaining accurate information on different cancer types such as lung cancer [11], breast [12] and prostate cancer [13]. However, one of the most important challenges for using deep learning models is to have annotated corpora to train these models. Creating annotated resources in the cancer domain is time-consuming and costly since this domain is complex and very specialized [14]. Moreover, most of the proposals to extract information in the cancer field have focused on the English language [15] and recently in Chinese [16]. In the case of Spanish, the first studies carried out have focused on extracting lung cancer information [9], [11].

In [14] the authors showed the feasibility of training deep learning models and exploiting them in other datasets in the medical domain. That proposal aims to reduce the cost of annotating medical text resources to support information extraction. Following the idea described in [14], in this paper, we propose a deep learning-based approach for extracting breast cancer information from Spanish clinical notes that leverages a model obtained using an annotated lung cancer corpus. This approach takes advantage of the lung cancer corpus described in [17] to fine-tune a BERT (Bidirectional Encoder Representations from Transformers) model. This BERT-based model is applied to extract breast cancer diagnosis concepts, and results are discussed.

The main contributions of this approach are:

This paper is supported by European Union's Horizon 2020 research and innovation program under grant agreement No. 875160, project CLARIFY.

- A deep learning-based approach for extracting breast cancer information from Spanish clinical notes. To the best of our knowledge, this is the first approach aiming to extract breast cancer information from clinical texts written in Spanish.
- An approach that leverages models trained in lung cancer annotated corpora and exploits them in another cancer type. This approach aims to reduce the time in annotating resources to support cancer diagnosis extraction.

The rest of the paper has been organized as follows: Section II shows a review of relevant studies about cancer concept extraction, Section III describes data and methods of the proposed approach. Section IV presents the results of the experiments, and Section V presents the main conclusions and outlook for future work.

II. Related Works

Extracting information related to cancer using NLP approaches has grown in recent years because it can be used to perform evidence-based medicine, health quality improvement, and patient-centered treatments [18]. In the oncology field, the first studies used rule-based approaches and aimed to extract the cancer tumor stage. The cancer stage indicates the grade and size of the tumor when the cancer was diagnosed [19]–[21]. The study carried out in [22] uses a set of regular expressions and the Unified Medical language system (UMLS dictionary) to extract lung cancer concepts. The main weakness of these approaches is that they rely on hand-crafted rules and dictionaries of medical terms. These approaches are limited to the dictionary terms and are difficult to adapt to other medical fields.

Machine learning-based approaches aim to extract information from clinical notes using annotated data and a set of defined features. In these approaches, the extraction of cancer concepts is defined as a classification problem where each word in a sentence is labeled with a pre-defined label. In [10] are described two machine learning-based models to extract information from cancer pathology reports using the Support Vector Machine Machine (SVN) algorithm. The first model was used to classify notes into internal (primary review) and external (consultation) reports. The second model was used to extract dates and tumor location in patients diagnosed with gastroesophageal cancer. The Conditional Random Field (CRF) algorithm [23] has also been used to extract information in the cancer field. A disadvantage of these approaches is they require a considerable set of hand-crafted features which frequently depends on humans and is a time-consuming task [24], [25].

Recently, deep learning-based methods have been shown to improve performance in the processing of natural language texts [26], [27]. One of the core components in these approaches is the use of word embeddings, a dense low-dimensional word vector representations [28], [29]. Word embeddings allow words with similar meanings to have a similar representation. Moreover, transformer-based architectures such as XLNet [30], and BERT [31], [32] have also shown that such representations are able to improve performance in extracting medical concepts from clinical narratives. The main advantage of deep learning approaches is the ability to automatically learn high-level features from the text to reduce the time-consuming burden of the human feature engineering process.

Deep-learning methods have allowed researchers to extract more concepts related to cancer from clinical notes. These concepts include the diagnosis [17], cancer medications [11], radiotherapy, and chemotherapy treatments [33]. In [4] is described a deep convolutional neural network (CNN) to extract lung cancer stages, histology, tumor grades, and therapies (chemotherapy, radiotherapy, surgery) from clinical notes written in English. In [12], the authors describe a deep learning-based model to extract breast cancer information from clinical notes written in Chinese. This model uses BERT to extract a comprehensive set of concepts related to breast cancer. Most of these proposals have focused on the English language [4], [33] and Chinese [16]. Recently there is also growing interest in extracting cancer-related information in other languages such as Italian [34], French [25] and Bulgarian [35].

In the case of the Spanish language, previous studies have used deep learning methods to extract cancer diagnosis [17] and treatments [11] from lung cancer clinical notes. In [17] the authors describe a Bidirectional Long Short Memory Network (BiLSTM) to extract lung cancer diagnosis information. This proposal performs an F-score of 90%, showing the feasibility of deep learning methods to extract cancer information in the Spanish language. However, the above proposals have required great effort by oncology experts to manually create lung cancer annotations. Moreover, these proposals have focused only on lung cancer information extraction. There are no previous proposals to extract breast cancer from Spanish clinical texts.

Motivated by the success of the deep learning models to extract information from clinical narratives, in this paper, we propose an approach to extract breast cancer information from Spanish clinical texts. This approach leverages models obtained with an annotated corpus of lung cancer [11], [17] and uses them to extract breast cancer concepts. This proposal thus takes advantage of resources previously annotated.

III. Datasets and methods

The proposed approach aims to extract breast cancer diagnosis concepts using a BERT-based model that has been trained with a lung cancer corpus. In this section, the datasets used in the proposed approach are presented. Then, we describe the methods to extract breast cancer diagnosis concepts.

A. Datasets

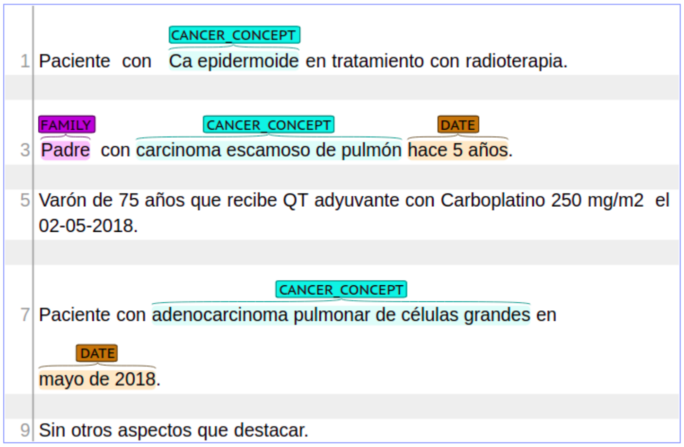
In the proposed approach we used one annotated dataset and another partially annotated dataset, as follows:

- **Lung cancer corpus:** is a dataset manually annotated which contains clinical notes of patients from “Hospital Universitario Puerta de Hierro, Madrid, Spain”. This corpus was described in the study performed in [17]. The corpus contains 14,000 annotated sentences of patients diagnosed and treated with lung cancer. The corpus was manually annotated using the BRAT annotation tool[1]. This corpus will be used for fine-tuning the core BERT model to perform clinical concept extraction. Table I shows a summary with the labels annotated in this corpus. Moreover, Figure 1 shows a set of annotations in the lung cancer corpus, which will be used for training a deep learning model.

- **Breast cancer corpus:** is a dataset that contains clinical notes of patients treated with breast cancer in "Hospital Universitario Puerta de Hierro, Madrid, Spain". This dataset contains more than 350,000 text sentences. We selected 200 sentences from this dataset and annotated the breast cancer diagnosis manually. Figure 2 shows a set of annotations in the breast cancer corpus. This corpus will be used for validating the deep learning model.

TABLE I
ANNOTATED LABELS ON THE CORPUS

Label	Description
Cancer entity	Used for labeling cancer mentions together with the anatomical location. Eg. Carcinoma escamoso en pulmón derecho
Dates	Represents dates and time expressions mentioned in clinical notes. This concept is used to follow the evolution of cancer in the patient.
Family members	Represents concepts about family members of the patient. This concept is essential to differentiate which cancer concepts belong to the patient and which to their family.



Paciente con Ca epidermoide en tratamiento con radioterapia.

Padre con carcinoma escamoso de pulmón hace 5 años.

Varón de 75 años que recibe QT adyuvante con Carboplatino 250 mg/m2 el 02-05-2018.

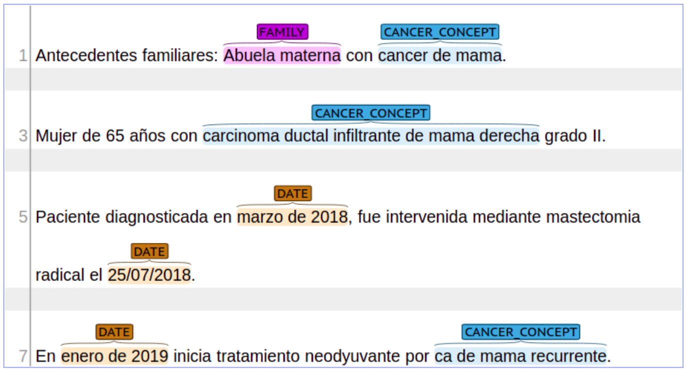
Paciente con adenocarcinoma pulmonar de células grandes en mayo de 2018.

Sin otros aspectos que destacar.

Fig. 1. Lung cancer annotations (Training corpus).

B. Deep learning methods to extract cancer concepts

The proposed approach addresses cancer concept extraction as a sequence-labeling task, where each token in a sentence is classified according to a set of labels defined in a corpus. Figure 3 shows the proposed approach, which consists of three steps: Preprocessing, BERT fine-tuning, and Model validation.



Antecedentes familiares: Abuela materna con cancer de mama.

Mujer de 65 años con carcinoma ductal infiltrante de mama derecha grado II.

Paciente diagnosticada en marzo de 2018, fue intervenida mediante mastectomia radical el 25/07/2018.

En enero de 2019 inicia tratamiento neodyuvante por ca de mama recurrente.

Fig. 2. Breast cancer annotations (Validation corpus).

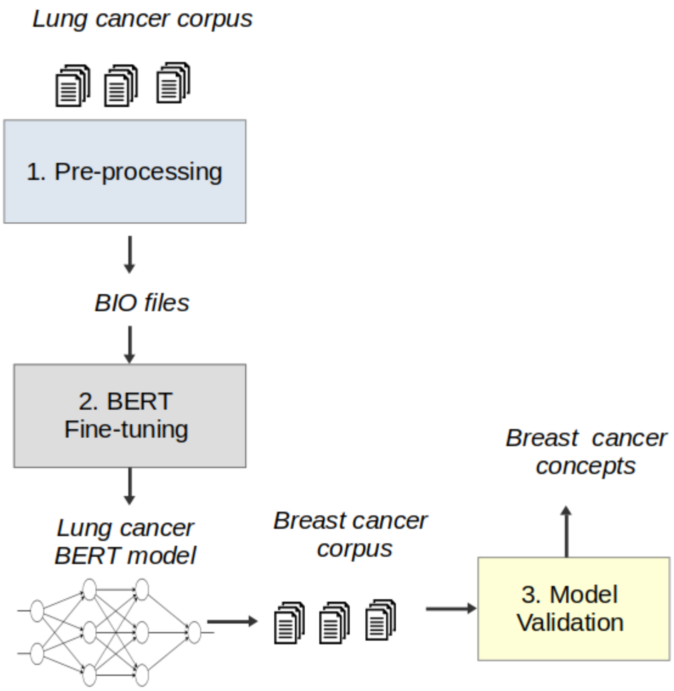


Fig. 3. Proposed approach to extract breast cancer concepts.

1) **Preprocessing**: transform annotations in the corpus into the BIO tagging format. In this format, each annotated token is labeled with B (at the beginning of the entity), I (inside the entity), or O (Outside the entity). For instance, the sentence *"Mujer con carcinoma ductal infiltrante de mama ."* (The cancer concept is underlined) can be formatted as follows:

   {'O', 'O', 'B-CANCER', 'I-CANCER', 'I-CANCER', 'I-CANCER', 'I-CANCER', 'O'}

2) **BERT fine-tuning**: we fine-tune the BERT model with a classification layer on top to perform token classification. In this step, each token in a sentence is

[1] https://brat.nlplab.org/

classified according to the set of labels defined in the corpus. We used Multilingual BERT [31] as pre-trained contextual embeddings. Figure 4 shows the process of extracting cancer concepts using BERT. This process consists of three steps: Tokenization, BERT Processing, and Classification & Post-processing.

- *Tokenization:* the goal of this step is to tokenize a text sentence using a WordPiece Tokenization method [36]. For each word in the sentence, this method decides to keep the whole word or to split it into a set of sub-words. Additionally, in this step, two special tokens are added to the sentence: [CLS] and [SEP]. The [CLS] token always appears at the beginning of the text, and the [SEP] token is used to separate sentences.
- *BERT Processing:* in this step, the approach takes as input a tokenized sentence from the previous step and obtains an embedding representation (E_n) for each word in the sentence. This representation is created using three embeddings: token, segment, and position embeddings. Then, the BERT Transformer Block takes the embedding representation as input (E_1, E_2, E_n) and produces a final representation (R_n) for each token in the processed sentence. This representation is a score calculated by BERT and represents a contextualized value for a specific word in relation to all other words in the sentence.
- *Classification & Post-Processing:* in this step, the approach takes as input the predicted BERT representations (R_1, R_2, R_n) and feeds them into the softmax function. This function obtains a label for each token in the sentence. A post-processing step is needed to convert BERT predictions into BIO format labels.

3) **Model Validation:** In this step, we used the BERT model fine tuned with the lung cancer corpus to extract breast cancer concepts. We load the BERT model trained in the above step and for each sentence in the breast cancer corpus this model predicts a set of BIO labels. The model predicts a BIO label for each token in the sentence.

IV. Experimentation and Results

In this section, we will first describe the evaluation methodology to validate our approach, the hyperparameters configuration, and finally, the results that were obtained.

A. Evaluation methodology

To train the BERT model with the lung cancer corpus, we followed a cross-validation strategy with $k = 10$. The performance was calculated as the average of all ten folds executed by the cross-validation strategy.

To evaluate the performance of the proposed approach, we used the following standard metrics: Precision (P), Recall (R), and F-score (F1). The F-score is calculated as a weighted average of the Precision and Recall measurements. A token is correctly classified when the predicted label is equal to the label indicated by the annotated corpus. The performance for the medical concept extraction task is measured at token level. Thus, each token is correctly classified when it corresponds to the token manually annotated in the corpus and according to the BIO tags (e.g., B-cancer, I-cancer).

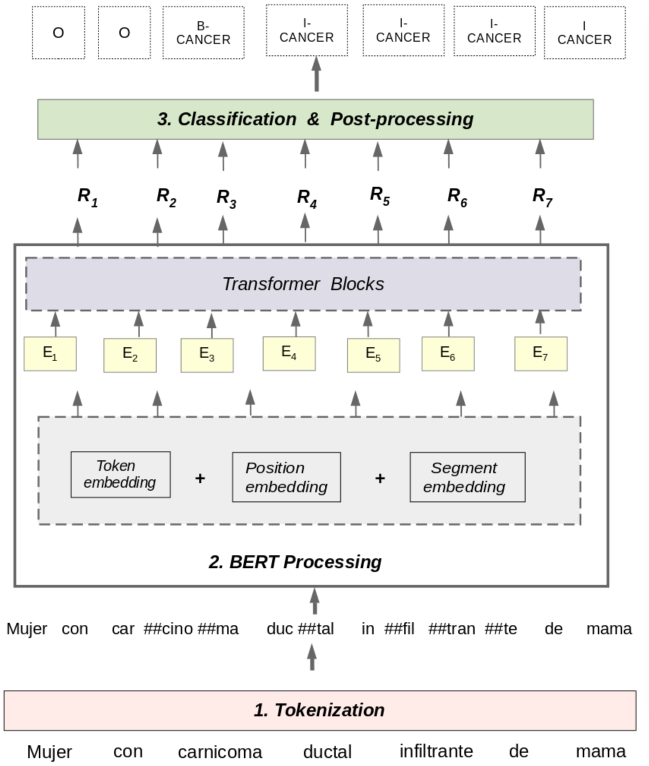


Fig. 4. Extracting cancer concepts using multilingual BERT.

$$\textbf{Precision} = \frac{\text{Number of tokens correctly predicted}}{\text{Number of predicted tokens}} \tag{1}$$

$$\textbf{Recall} = \frac{\text{Number of tokens correctly predicted}}{\text{Number of tokens in the dataset}} \tag{2}$$

$$\textbf{F-score} = 2 * \frac{\text{Precision * Recall}}{\text{Precision + Recall}} \tag{3}$$

B. Hyper parameters setting

To implement the proposed approach, we used Python 3.7, TensorFlow[2] and Keras[3].

The BERT fine-tuning was performed with a sequence length of 256 tokens, a batch size of 64, and 5 epochs. These values were established after training the model's different hyperparameters and checking for the best performance. The code developed to perform this study can be found on Github[4].

[2] https://www.tensorflow.org/?hl=es-419

[3] https://keras.io/

[4] https://github.com/solarte7/breastCancerDiagnosis

TABLE II
RESULTS FOR BERT FINE TUNING IN THE LUNG CANCER DATASET

Label	Precision	Recall	F-score
B-CANCER	0.96	0.97	0.97
I-CANCER	0.98	0.97	0.98

TABLE III
RESULTS OF APPLYING THE BERT LUNG CANCER MODEL TO THE ANNOTATED BREAST CANCER DATASET

Label	Precision	Recall	F-score
B-CANCER	0.93	0.92	0.93
I-CANCER	0.98	0.80	0.88

C. Results

In this section, we will show the obtained results both for fine-tuning the BERT model with the lung cancer corpus and for extracting breast cancer concepts using that model. Thus, Table II shows the results of tuning the model using multilingual BERT with the cancer labels (B-CANCER and I-CANCER). These results show the feasibility of the BERT model to perform clinical concept extraction in the lung cancer corpus obtaining high performance rates. In this paper we show only the results for the label *"Cancer entity"* described in Table I.

On the other hand, Table III shows the results of applying the previous model (trained with lung cancer data) to predict breast cancer data. These results are highly interesting, as they show that the learning of a model on one type of cancer can be extrapolated to other types of cancers with a very high success rate. In this case, the high performance obtained in the experiment may be due to the ability of the BERT-based model to predict instances that were never seen in the training process. For instance, the breast cancer concept **"carcinoma ductal infiltrante de mama derecha"** (See Figure 2) is accurately predicted by the model even though the lung cancer corpus does not contain breast cancer concepts.

D. Additional external validation

In addition to the previously performed experiments, an additional external validation has been performed on breast cancer notes that were not previously annotated to verify that all cancer concepts of the sentences are located and that no relevant information is lost. Thus, the model learned in the lung cancer corpus was run on 200 unlabeled sentences. In this case, the final validation was done by hand, in a double-check process to avoid errors.

Taking into account the 200 total sentences, where 263 cancer concepts have been extracted. Of these, 223 were perfectly found, while in 40 cases (15.20%), the complete concept was not correctly extracted. However, some additional problems were observed. Considering the 223 positive concepts, on 38 cases (17.04%) it was not possible to detect that the cancer was of a non-specific type ("NOS"), and on 32 cases (14.35%) the location of the cancer could not be correctly classified.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we have shown how the application of NLP models trained on a lung cancer annotated corpus can be used to obtain cancer diagnosis from a set of notes of breast cancer patients. The proposed approach in this paper showed the feasibility of deep-learning models trained on a source corpus to predict concepts in another cancer type (target dataset).

Cancer is a complex domain, and annotating resources for each cancer type can be time-consuming and costly. Thus, the proposed approach aimed to reduce this time-consuming by leveraging previous annotated resources.The approach offers the possibility of exploiting deep learning methods and reducing the time in annotating cancer text resources.

The main errors in the extraction of breast cancer were due to: i) the diagnosis of cancer also includes the tumor location in the body. In this case, some errors were observed in which the correct location was not completely annotated; ii) the type of the breast cancer: when the type was not specific, and this was written as "NOS" in Spanish (No Específico), this was not found by the model. Both errors will have to be analyzed, but in any case, the results obtained are promising and show that for many concepts, it is not necessarily to annotate the corpus as the models have good behavior.

In future works, we plan to extend the proposed approach in this paper exploring the following issues:

- We will evaluate the medical concept extraction at entity level to measure how the complete cancer concepts are extracted. Moreover, this approach will be applied to other concepts such as treatments and comorbidities.
- We will explore other medical language models such as BIOBERT [37], or clinical BERT [38] to extract medical concepts in the cancer field.

ACKNOWLEDGMENT

This paper is supported by European Union's Horizon 2020 research and innovation program under grant agreement No. 875160, project CLARIFY.

REFERENCES

[1] P. M. de Groot, C. C. Wu, B. W. Carter, and R. F. Munden, "The epidemiology of lung cancer," *Translational Lung Cancer Research*, vol. 7, no. 3, 2018. [Online]. Available: http://tlcr.amegroups.com/article/view/21996

[2] "Lung Health and Diseases lung disease lookup," https://www.lung.org/lung-health-and-diseases/lung-disease-lookup/lung-cancer/resource-library/lung-cancer-fact-sheet.html, accessed: 2020-01-30.

[3] "Lung Health and Diseases lung disease lookup," https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html, accessed: 2020-02-14.

[4] L. Wang, L. Luo, Y. Wang, J. Wampfler, P. Yang, and H. Liu, "Natural language processing for populating lung cancer clinical research data," *BMC Medical Informatics and Decision Making*, vol. 19, no. Suppl 5, pp. 1–10, 2019. [Online]. Available: http://dx.doi.org/10.1186/s12911-019-0931-8

[5] S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, V. Osmani *et al.*, "Natural language processing of clinical notes on chronic diseases: systematic review," *JMIR medical informatics*, vol. 7, no. 2, p. e12239, 2019.

[6] E. P. Balogh, P. A. Ganz, S. B. Murphy, S. J. Nass, B. R. Ferrell, and E. Stovall, "Patient-centered cancer treatment planning: improving the quality of oncology care. Summary of an institute of medicine workshop," 2011.

[7] A. J. Tevaarwerk, K. B. Wisinski, K. A. Buhr, U. O. Njiaju, M. Tun, S. Donohue, N. Sekhon, T. Yen, D. A. Wiegmann, and M. E. Sesto, "Leveraging electronic health record systems to create and provide electronic cancer survivorship care plans: a pilot study," *Journal of oncology practice*, vol. 10, no. 3, pp. e150–e159, 2014.

[8] S. Shanthi, "A survey on non-small cell lung cancer prediction using machine learning methods," in *2nd EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*. Springer, 2021, pp. 255–266.

[9] M. Najafabadipour, J. M. Tuñas, A. Rodríguez-González, and E. Menasalvas, "Lung cancer concept annotation from spanish clinical narratives," in *Data Integration in the Life Sciences*, S. Auer and M.-E. Vidal, Eds. Springer International Publishing, 2019, pp. 153–163.

[10] T. Oliwa, S. B. Maron, L. M. Chase, S. Lomnicki, D. V. Catenacci, B. Furner, and S. L. Volchenboum, "Obtaining Knowledge in Pathology Reports Through a Natural Language Processing Approach With Classification, Named-Entity Recognition, and Relation-Extraction Heuristics," *JCO Clinical Cancer Informatics*, no. 3, pp. 1–8, 2019.

[11] O. Solarte-Pabón, A. Blazquez-Herranz, M. Torrente, A. Rodríguez-Gonzalez, M. Provencio, and E. Menasalvas, "Extracting cancer treatments from clinical text written in spanish: A deep learning approach," in *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2021, pp. 1–6.

[12] X. Zhang, Y. Zhang, Q. Zhang, Y. Ren, T. Qiu, J. Ma, and Q. Sun, "Extracting comprehensive clinical information for breast cancer using deep learning methods," *International Journal of Medical Informatics*, vol. 132, no. September, p. 103985, 2019. [Online]. Available: https://doi.org/10.1016/j.ijmedinf.2019.103985

[13] T. Hernandez-Boussard, P. D. Kourdis, T. Seto, M. Ferrari, D. W. Blayney, D. Rubin, and J. D. Brooks, "Mining Electronic Health Records to Extract Patient-Centered Outcomes Following Prostate Cancer Treatment," *AMIA ... Annual Symposium proceedings. AMIA Symposium*, vol. 2017, pp. 876–882, 2017.

[14] O. S. Pabón, O. Montenegro, M. Torrente, A. R. González, M. Provencio, and E. Menasalvas, "Negation and uncertainty detection in clinical texts written in spanish: a deep learning-based approach," *PeerJ Computer Science*, vol. 8, p. e913, 2022.

[15] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, "What can natural language processing do for clinical decision support?" *Journal of Biomedical Informatics*, vol. 42, no. 5, pp. 760 – 772, 2009, biomedical Natural Language Processing. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046409001087

[16] X. Li, H. Zhang, and X.-H. Zhou, "Chinese clinical named entity recognition with variant neural structures based on BERT methods," *Journal of biomedical informatics*, vol. 107, p. 103422, 2020.

[17] O. Solarte Pabón, M. Torrente, M. Provencio, A. Rodríguez-Gonzalez, and E. Menasalvas, "Integrating speculation detection and deep learning to extract lung cancer diagnosis from clinical notes," *Applied Sciences*, vol. 11, no. 2, p. 865, 2021.

[18] F. S. Saiz, C. Sanders, R. Stevens, R. Nielsen, M. Britt, L. Yuravlivker, A. M. Preininger, and G. P. Jackson, "Artificial intelligence clinical evidence engine for automatic identification, prioritization, and extraction of relevant clinical oncology research," *JCO Clinical Cancer Informatics*, vol. 5, pp. 102–111, 2021.

[19] K. Liu, K. J. Mitchell, W. W. Chapman, and R. S. Crowley, "Automating tissue bank annotation from pathology reports–comparison to a gold standard expert annotation set," in *AMIA Annual Symposium Proceedings*, vol. 2005. American Medical Informatics Association, 2005, p. 460.

[20] J. L. Warner, M. A. Levy, M. N. Neuss, J. L. Warner, M. A. Levy, and M. N. Neuss, "ReCAP: Feasibility and Accuracy of Extracting Cancer Stage Information From Narrative Electronic Health Record Data," *Journal of Oncology Practice*, vol. 12, no. 2, pp. 157–158, 2016.

[21] E. Soysal, J. L. Warner, J. Wang, M. Jiang, K. Harvey, S. K. Jain, X. Dong, H.-Y. Song, H. Siddhanamatha, L. Wang *et al.*, "Developing customizable cancer information extraction modules for pathology reports using clamp," *Studies in health technology and informatics*, vol. 264, p. 1041, 2019.

[22] M. Najafabadipour, M. Zanin, A. Rodriguez-Gonzalez, C. Gonzalo-Martin, B. N. Garcia, V. Calvo, J. L. C. Bermudez, M. Provencio, and E. Menasalvas, "Recognition of time expressions in Spanish electronic health records," *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, vol. 2019-June, pp. 69–74, 2019.

[23] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01. Williamstown, MA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289. [Online]. Available: http://dl.acm.org/citation.cfm?id=645530.655813

[24] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning–based text classification: a comprehensive review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–40, 2021.

[25] V. Jouhet, G. Defossez, A. Burgun, P. le Beux, P. Levillain, P. Ingrand, and V. Claveau, "Automated classification of free-text pathology reports for registration of incident cases of cancer," *Methods of Information in Medicine*, vol. 51, no. 3, pp. 242–251, 2012.

[26] G. Harerimana, J. W. Kim, H. Yoo, and B. Jang, "Deep learning for electronic health records analytics," *IEEE Access*, vol. 7, pp. 101 245–101 259, 2019.

[27] Y. Kim, J. H. Lee, S. Choi, J. M. Lee, J.-H. Kim, J. Seok, and H. J. Joo, "Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records," *Scientific reports*, vol. 10, no. 1, pp. 1–9, 2020.

[28] Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, P. Kingsbury, and H. Liu, "A comparison of word embeddings for the biomedical natural language processing," *Journal of Biomedical Informatics*, vol. 87, no. April, pp. 12–20, 2018. [Online]. Available: https://doi.org/10.1016/j.jbi.2018.09.008

[29] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[30] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[31] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[32] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[33] D. Bitterman, H. Chen Lin, S. Finan, J. Warner, R. Mak, and G. Savova, "Extracting radiotherapy treatment details using neural network-based natural language processing." in *Annual Meeting of the American Society for Radiation Oncology*, Cham, 2020.

[34] S. Martina, L. Ventura, and P. Frasconi, "Classification of cancer pathology reports: A large-scale comparative study," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 11, pp. 3085–3094, 2020.

[35] B. Zhao, "Clinical data extraction and normalization of cyrillic electronic health records via deep-learning natural language processing," *JCO Clinical Cancer Informatics*, vol. 3, pp. 1–9, 2019.

[36] X. Song, A. Salcianu, Y. Song, D. Dopson, and D. Zhou, "Fast wordpiece tokenization," *arXiv preprint arXiv:2012.15524*, 2020.

[37] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 09 2019. [Online]. Available: https://doi.org/10.1093/bioinformatics/btz682

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