Tracking the Evolution of Covid-19 Symptoms through Clinical Conversations

Ticiana L. Coelho da Silva, José Antônio F. de Macêdo, Régis Pires Magalhães
Insight Data Science Lab
ÍRIS Lab, Ceará, Brazil
{ticianalc, jose.macedo, regis}@insightlab.ufc.br

Abstract

The Coronavirus pandemic has heightened the demand for technological solutions capable of gathering and monitoring data automatically, quickly, and securely. To achieve this need, the Plantão Coronavirus chatbot has been made available to the population of Ceará State in Brazil. This chatbot employs automated symptom detection technology through Natural Language Processing (NLP). The proposal of this work is a symptom tracker, which is a neural network that processes texts and captures symptoms in messages exchanged between citizens of the state and the Plantão Coronavirus nurse/doctor, i.e., clinical conversations. The model has the ability to recognize new patterns and has identified a high incidence of altered psychological behaviors, including anguish, anxiety, and sadness, among users who tested positive or negative for Covid-19. As a result, the tool has emphasized the importance of expanding coverage through community mental health services in the state.

1 Introduction

The Covid-19 pandemic required efficient and agile measures from governments to mitigate the effects caused by the disease. Plantão Coronavírus (PC) is a chatbot, and one of the solutions developed around April 2020 in Ceará State, Brazil, to hold back the pandemic. PC is an automated tool to converse with patients via text and provide them guidelines on how to proceed based on their level of emergency. The Plantão Coronavirus project was specifically developed by ÍRIS Lab of Innovation and Data in Ceará state, Brazil, in collaboration with the Health Secretary of Ceará State. Its purpose is to create a chatbot with artificial intelligence capabilities. The platform incorporates chatbot technology, enabling users to interact with an artificial intelligence system, and also offers the option to redirect to a virtual service manned by healthcare professionals such as doctors or nurses.

The interactions between patients and healthcare professionals through the PC generated a lot of clinical conversation data that needed to be mined, analyzed, and transformed into valuable information. The Health Secretary was required to track the signs of the disease, and it was not feasible to perform this task manually by reading thousands of texts. In this way, an automated and intelligent solution to classify the Covid-19 symptoms was essential. Especially at the beginning of the pandemic, when little was known about Covid-19.

This work proposes a solution to address the issue of screening the symptoms reported by users in a chat dialogue box. Ultimately, the chatbot classifies the user's health status as mild, moderate, or severe. Based on this classification, the system recommends various services to the user, including medical appointments or Covid-19 tests.

Our solution approaches the identification of symptoms in text as a Named Entity Recognition (NER) problem. NER identifies named entities in documents and categorizes them into predefined classes based on the type of entity (Li et al., 2020). Typically, a neural network is utilized for entity recognition. In this study, the named entity is a symptom.

However, in order to train a NER model for extracting symptoms and diseases from clinical conversations in Plantão Coronavírus, automatic annotation was necessary due to the impracticality of manual annotation given the large size of the dataset. Additionally, there was no publicly available NER model for Brazilian Portuguese that could extract symptoms and diseases. Therefore, we utilized ScispaCy (Neumann et al., 2019), a disease-focused NER model trained on English-language texts, and employed transfer learning to build our model. During the training process, we initially translated the Portuguese texts into English and used the ScispaCy model to analyze each English text, extracting the symptoms identified and

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 41–47

translating them back into Portuguese. The training set comprised the original text and the Portuguese symptoms extracted by the ScispaCy model.

The intelligence developed in this study was crucial in identifying patterns of disease indicators, moreover, new or rare symptoms that had not yet been documented by researchers and health professionals in the state of Ceará. This enabled the tracking of the evolution of Covid-19 findings over time.

The entity recognition process was performed automatically. Related works as (Tarcar et al., 2019) achieved an F1-score of 78.5% and (Neumann et al., 2019) reached 84.94% of F1-score for the symptom/disease discovery model. In contrast, this work achieved F1-score equal to 85.66%. The F1-score is a metric that quantifies the harmonic mean of precision and recall. In our case, we report the F1-score for the test set. Our approach has made significant advances in tackling the disease in Ceará, given the possibility of virus mutations and the consequent appearance of new symptoms. With the capability to recognize new patterns, our model identified a high frequency of altered psychological behaviors, such as anxiety, anguish, and sadness, in both Covid-19 positive and negative users. As a result, our tool highlighted the need for the state to expand its mental health care services to the population through various channels. Our work has enabled the government to develop a public policy to address this need.

Moreover, less research is available for clinical texts in low resource languages as Brazilian Portuguese (Schneider et al., 2020). One can argue whether or not Brazilian Portuguese is an low resource language, but we say that, at least for some tasks, there are fewer resources compared to English or other languages, as discussed by (Costa et al., 2020) and (Fischer et al., 2022).

2 Background

2.1 Plantão Coronavírus Dataset

During the COVID-19 pandemic, the Brazilian state government of Ceará introduced Plantão Coronavírus, a web-based system designed to facilitate online patient consultation[1].

To better understand how the Plantão Coronavírus dataset was built, it is necessary to explain some aspects of the Plantão Coronavírus system. When a user initiates an interaction with the system, a virtual screening protocol categorizes the user into one of three risk profiles: severe, moderate, or mild. In the severe risk profile, the user reports severe symptoms directly related to Covid-19, such as shortness of breath, fever above 39ºC for more than 48 hours, etc. The moderate risk profile is for users who do not report severe symptoms of Covid-19 but may be at an increased risk for getting very sick from Covid-19, such as elderly over 70 years old, those with diabetes, asthma or chronic lung disease, sickle cell disease, or those who are immunocompromised, pregnant women, etc. The low-risk or mild-risk profile is for users who report being asymptomatic or having mild symptoms and do not belong to high-risk groups for Covid-19, such as a stuffy or runny nose, headaches, pain, etc. After categorization, the user interacts with a nurse who answers questions on several topics related to Covid-19, primary care, testing locations, etc.

Since the launch of Plantão Coronavírus, numerous consultations have been conducted, resulting in the recording of several clinical dialogues between patients, nurses, and doctors. A subset of these recordings was utilized to create a dataset for training and testing a neural network that identifies Covid-19 symptoms in Brazilian Portuguese texts. To the best of the author's knowledge, no existing model was available to recognize symptoms in Brazilian Portuguese texts.

The criteria for classifying a user into one of the three categories have been subject to changes over time, following new guidelines from the Health State Secretary and the World Health Organization (WHO). Moreover, as the people's expertise in dealing with the pandemic has grown, the understanding of categorizing users has also evolved. Consequently, the recorded dialogues between users and healthcare professionals vary significantly depending on the period considered. This variability can introduce noise into the dataset. To minimize this effect, we used data from two months (April/20 and May/20) when the same Covid-19 protocol was followed in the state of Ceará. This was the longest continuous period we could identify when the same protocol was in place, allowing us to build a more robust dataset.

Table 1 presents an example of a clinical conversation between a patient and a doctor. The dataset, we used in this work, includes approximately 27,690 dialogues, with 577,814 utterances,

[1] https://coronavirus.ceara.gov.br/

an average of 21.48 turns per dialogue, and an average of 9.9 words per utterance. As a reminder, an utterance in a dialogue refers to a complete unit of speech produced by one speaker, which can be a sentence, phrase, or even a single word. In contrast, a turn refers to a sequence of utterances produced by one speaker before the other speaker takes a turn in the conversation. A turn may consist of one or more utterances. For privacy reasons, data is not publicly available.

Doctor	Hi, I'm Doctor Fabio. How can I help you?
Patient	I'm experiencing three days of fever and a dry cough.
Doctor	Did you get to take the temperature?
Patient	Yes.
Doctor	Do you feel shortness of breath?
Patient	During this period of fever, I took Dipirone...
Patient	A little, I think because of the cough
Doctor	right
Doctor	Do you feel shortness of breath when you walk?
Patient	only when I cough.
Doctor	I recommend taking a COVID test...
...	...

Table 1: Plantão Coronavírus Dialogue example

2.2 Related Works

Researchers have lately worked to create Covid-19-related chatbots due to the significant demand for patient follow-ups. Using research papers from the Covid-19 Open Research Dataset and CORD-19 (Wang et al., 2020), (Lei et al., 2021) trained a NER model. The research group used the papers to extract entities to identify symptoms in the patient's written sentences. The most prevalent symptoms in the articles are found using word clouds, and a knowledge graph was created using the chatbot NLU model to keep track of follow-up appointments with returning patients. (Fazzinga et al., 2021) uses argumentation graphs and natural language to create dialog systems explaining Covid-19 vaccination.

The paper (Miner et al., 2020) outlines issues and queries that a chatbot may handle during a pandemic like Covid-19. Initiatives like Clara [2] from the CDC in the United States aim to combat the proliferation of contradicting information brought on by ignorance and fake news, which can eventually make handling the pandemic crisis much more challenging.

(Schaeffer et al., 2022) provides a dataset based on three corpora: the first one contains 70 carefully annotated tweets and 10 transcriptions of YouTube videos. The second corpus comprises the same textual material with the named entities annotated. 100 YouTube transcriptions that were automatically tagged using NER models included the third corpus. The dataset provides geographic information, city names, and epidemiological data, such as diseases, symptoms, and virus entities. (Schaeffer et al., 2022) can be used to train and evaluate different NER models.

(Beltagy et al., 2019) proposes SciBERT, a pretrained BERT-based language model for scientific data. To enhance performance on downstream scientific NLP tasks, SciBERT uses unsupervised pretraining on a large multi-domain corpus of scientific literature. The paper (Beltagy et al., 2019) evaluates the performance of SciBERT for different NLP tasks such as NER, Relation Classification, and Text Classification, among others. BioBERTpt (Schneider et al., 2020) was developed using clinical notes and scientific abstracts. BioBERTpt is NER model BERT-based for Portuguese texts. However, the paper does not clarify the quality of the model to extract the named entity disease. We might experiment BioBERTpt and maybe enrich our annotated data, instead of using only ScispaCy to annotated our training data. It might be a future study direction.

The authors (Lopes et al., 2019) manually collected and annotated a corpus of Portuguese clinical texts, identifying named entities such as characterization, test, evolution, genetics, additional observations, results, date and time, therapeutics, among others. They also evaluated the effectiveness of various state-of-the-art models for named entity recognition. While their work is relevant to ours, we are particularly interested in named entities related to diseases or symptoms, which were not included in the dataset used in (Lopes et al., 2019).

(Schäfer et al., 2022) explores two main approaches. Firstly, the authors investigate the application of English models to translated texts, followed by the transfer of predicted annotations back

[2] https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/testing.html

to the source language. This direction closely aligns with the approach adopted in our paper to create our training set. Secondly, the authors explore the possibility of utilizing existing high-quality annotations to train NLP models in the target language, going beyond mere translation. Given the scarcity of resources for low-resource languages, the ideia is to employ English models and external biomedical and clinical datasets as substitutes. The primary objective is to assess the potential benefits for low-resource languages by leveraging the existing resources available in English. The findings in (Schäfer et al., 2022) indicate that English language models can indeed be applied to other languages in clinical contexts. Translated training data can serve as a solid foundation in languages where resources are otherwise lacking. The success of the second approach depends on both the annotation standards and the similarity between English and the low-resource language in terms of grammar and morphology.

The study (Schäfer et al., 2022) further supports the suitability of our methodology. A potential future direction could involve evaluating our translation phase, similar to the approach described in (Schäfer et al., 2022), through word alignment using contextualized embeddings with the assistance of multilingual BERT.

3 Data and Methods

The symptom capture mechanism is the primary contribution of this work and is integrated into the entire triage process, from the chatbot to the teleservice with health professionals. Our solution is a technology that leverages Plantão Coronavírus data to process and identify symptoms contained in natural language texts. In the following sections, we describe the pipeline for building a tracker that monitors the evolution of Covid-19 symptoms.

Note that the sentences from Plantão Coronavírus were not annotated with symptoms. The main challenge at this point was to build a training set using this data. The pipeline from data collection to implementation of the NER model is shown in Figure 1 and described as follows.

3.1 Annotate the Brazilian Portuguese dataset with Covid-19 symptoms

The detection of symptoms in the Portuguese language was a challenge because, at the beginning of the pandemic, no publicly available model could perform this task, according to the authors' knowledge.

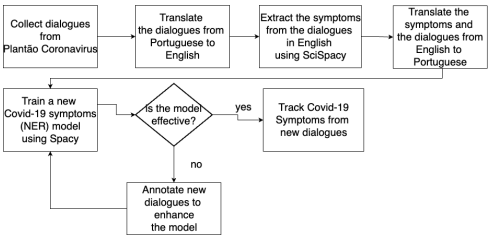

Figure 1: Pipeline to build and deploy the tracker of Covid-19 symptoms evolution.

Our solution is based on Transfer Learning (Pan and Yang, 2009). The technological innovation provided by our solution is a pioneering neural model for recognizing symptoms in Brazilian Portuguese, mainly because the Portuguese language lacks NER models. The transfer learning technique uses the knowledge gained by solving one problem and applying it to a different but related problem, allowing for rapid progress and improved performance when modeling the second task. In other words, transfer of learning is the improvement of learning in a new task by transferring knowledge from a related task that has already been learned.

So after collecting the dialogues from Plantão Coronavírus, we chose to translate the texts that were initially in Portuguese into English. Then, submit each text (in English) to the ScispaCy model (Neumann et al., 2019) as an input parameter. For this work, the model used from ScispaCy was *en_ner_bc5cdr_md*. Then, we analyzed the result generated by this model and translated the symptoms captured by the ScispaCy model from English to Portuguese. All in all, the training set of our NER model comprises the original dialog text and the symptoms captured by the ScispaCy model in Portuguese.

The Google Translate was used in the text translation stages within our pipeline. Nowadays, these translation networks present very accurate results to the expected ones, making the noise insignificant when analyzed in the context of this work.

Another important aspect to consider is the alignment of words or tokens. When translating clinical conversations from Portuguese to English and annotating them using ScispaCy, all named entities are outputted in English by the NER model. Subsequently, we translate them back to Brazilian Portuguese and use a script to locate them in the original text. Our script also generates a training

set in spaCy file format, indicating the occurrences of named entities in the original text from Plantão Coronavírus.

3.2 Train and Evaluate the NER model

As the NER component, we utilize spaCy NER [3]. Given its powerful neural network-based model's cutting-edge performance, this off-the-shelf NER technology is typically chosen for use in many industrial applications (Honnibal et al., 2020). The entity recognizer may be updated with new instances using an existing pre-trained statistical model thanks to SpaCy's support for online learning.

The embed, encode, attend, and forecast steps follow the basic four-step methodology used by the spaCy NLP models, particularly NER. The model first takes the text as input and converts the words into distinct number values. Prefix, suffix, shape, and lowercase characteristics are employed in the embedding step to extract the commonalities between the words. The values are sent through a CNN (Convolutional Neural Network) network to encode the context-free embeddings, creating a context-sensitive sentence matrix. The matrix must travel through the CNN Attention layer before transforming the prediction into a single vector. A common Multi-layer Perceptron (MLP) with a Softmax layer is then utilized as a tag decoder layer for class prediction. After training, the spaCy model is prepared for various NLP tasks.

Initially, we trained the NER model with a total of 27,690 dialogues in Portuguese. The dataset contains at least one symptom annotated per sentence. So, we split the dataset into training (22,152 dialogues) and test (5,538 dialogues) sets following the distribution of 80% and 20%, respectively. We kept training the NER model until it achieved an F1-score equivalent to ScispaCy, i.e., 85.02. In the end, it was possible to reach in terms of F1-score of 85,6 for our NER model.

3.3 Deployment of the NER model

A relevant aspect to point out of the NER model to extract Covid-19 symptoms in this work is the absence of manually annotating the data, usually performed by a human for entity recognition. In a scenario where there was a vast amount of data, and little time to process this information, the gain from optimizing this training step was crucial in supporting decision-making.

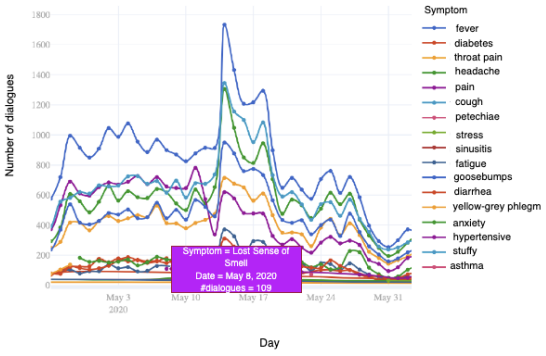

Figure 2: The Evolution of Covid-19 Symptoms and Related Diseases.

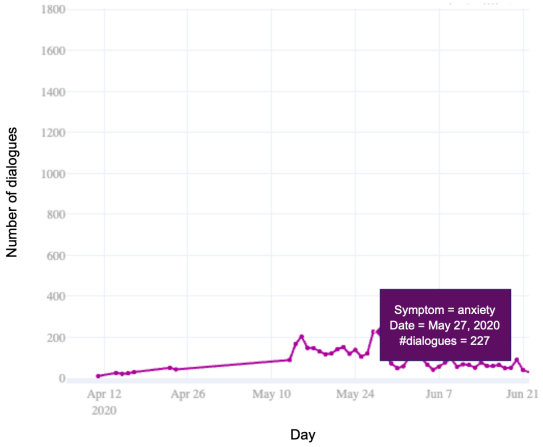

Figure 3: Repeatedly reports of anxiety along the days.

Another innovative aspect of our approach is its ability to recognize mental health behaviors, enabling health professionals to develop and promote public policies to assist individuals affected by issues beyond the scope of epidemiology.

Currently, our solution is used in the Tele-Service platform of the State of Ceará, where it plays a pioneering role in the Health domain [4].

4 Experimental Results

Figure 2 demonstrates the evolution of symptoms in a time series. Not all symptoms could be shown in the figure's legend. World Health Organization (WHO), at the beginning of the pandemic, stated a set of symptoms commonly reported by people who got positive for Covid-19 very related to flu, like nasal flaring, runny nose, malaise, fever, cough, sore throat, diarrhea, headache, and no appetite.

However, new symptoms came along. Through Figure 2, it is possible to identify the detection

[3] https://spacy.io/api/entityrecognizer

[4] https://coronavirus.ceara.gov.br/

of a new symptom of loss of smell on the 8th of May. This symptom appeared and became quite characteristic of Covid-19 after a certain period. Still, on the time series, it is possible to see that the frequency of each symptom is seasonal during the analyzed period.

Diabetes is not tecnically a symptom, however our NER model considered as a symptom, possibly due to contextual factors or biases introduced by the scispaCy model. However, it is worth mentioning that there has been an observed rise in hyperglycemic conditions associated with COVID-19, particularly in patients with diabetes and those receiving steroid treatment (Lim et al., 2021). Nevertheless, this topic falls outside the scope of the current paper.

Another interesting point is that our NER model could capture some symptoms related to altered psychological behaviors, such as anxiety, mental confusion, neurological disorder, and disorientation among the symptoms, as highlighted in Figure 3. Figures 2 and 3 do not represent the entire population as a whole, but rather the number of dialogues on the Plantão Coronavírus platform that report specific symptoms.

As already reported in the previous session, our Covid-19 symptom tracker achieved an F1-score of 85.66, which is competitive compared to the SciSpacy English model, which has an F1-score of 85.02. To mitigate catastrophic forgetting of old knowledge as we update our NER model, we kept including new sentences from Plantão Coronavírus with the following symptoms annotated, such as breathing difficulty, mental confusion, loss of smell, loss of taste, tiredness, anxiety, anosmia, neurological disorder, and disorientation, so the model could learn not only from the frequent symptoms identified by *en_ner_bc5cdr_md* from SciSpacy and the symptoms commonly reported by people who got positive to Covid-19.

5 Conclusion and Future Works

This research provides a NER model to recognize Covid-19 symptoms in Portuguese textual conversations. At the start of the pandemic, no model could automatically identify the symptoms in a text written in Brazilian Portuguese; instead, we utilized ScispaCy, an English-language NER model for diseases, which through transfer learning, trained our NER model.

The texts were initially translated from Portuguese into English as part of the training procedure. The ScispaCy model then processes each English-language input text, and its identified symptoms are subsequently translated from English to Portuguese. The original text and the Portuguese symptoms determined by the ScispaCy model comprise the training set.

On the Plantão Coronavírus dataset, our NER model achieved an F1-score of 85.66, which is competitive with the English model of ScispaCy, which has an F1-score of 85.02. The NER model has brought to light the necessity for the state to increase its coverage of mental health services through the community mental health channel.

As a future research direction, we intend to extend the NER model to other diseases prevalent in Brazil, such as influenza, and explore various neural architectures. Another future work is to investigate the translation performance. One alternative might be manually translate a small sample of our dataset and then compute the BLEU score of the automatic translations on this sample for a more accurate estimate. As mentioned earlier, we selected data from April and May 2020, which corresponds to the early stages of the pandemic when Plantão Coronavírus followed the same protocol. As a future study, we can also examine the evolution of “altered physiological behaviors" in the dialogues that occurred after this time period.

Acknowledgments

The research reported in this work received support from the FUNCAP projects titled “Big Data Platform to Accelerate the Digital Transformation of Ceará State", “Citizen Platform" and “Digital Government" under the numbers 04772314/2020, 04772551/2020 and 04772420/2020, respectively. Part of the results presented in this work were obtained also through the project “Center of Excellence in Artificial Intelligence - AI4WELLNESS", funded by Samsung Eletrônica da Amazônia Ltda., under the scope of the Informatics Law no. 8,248/91”.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Felipe Almeida Costa, Thiago Castro Ferreira, Adriana Pagano, and Wagner Meira. 2020. Building the first english-brazilian portuguese corpus for automatic

post-editing. In *Proceedings of the 28th international conference on computational linguistics*, pages 6063–6069.

Bettina Fazzinga, Andrea Galassi, and Paolo Torroni. 2021. An argumentative dialogue system for covid-19 vaccine information. In *International Conference on Logic and Argumentation*, pages 477–485. Springer.

Marcelo Fischer, Rejwanul Haque, Paul Stynes, and Pramod Pathak. 2022. Identifying fake news in brazilian portuguese. In *International Conference on Applications of Natural Language to Information Systems*, pages 111–118. Springer.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Hannah Lei, Weiqi Lu, Alan Ji, Emmett Bertram, Paul Gao, Xiaoqian Jiang, and Arko Barman. 2021. Covid-19 smart chatbot prototype for patient monitoring. *arXiv preprint arXiv:2103.06816*.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Soo Lim, Jae Hyun Bae, Hyuk-Sang Kwon, and Michael A Nauck. 2021. Covid-19 and diabetes mellitus: from pathophysiology to clinical management. *Nature Reviews Endocrinology*, 17(1):11–30.

Fábio Lopes, César Teixeira, and Hugo Gonçalo Oliveira. 2019. Contributions to clinical named entity recognition in portuguese. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 223–233.

Adam S Miner, Liliana Laranjo, and A Baki Kocaballi. 2020. Chatbots in the fight against the covid-19 pandemic. *NPJ digital medicine*, 3(1):1–4.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Camille Schaeffer, Roberto Interdonato, Renaud Lancelot, Mathieu Roche, and Maguelonne Teisseire. 2022. Labeled entities from social media data related to avian influenza disease. *Data in Brief*, page 108317.

Henning Schäfer, Ahmad Idrissi-Yaghir, Peter Horn, and Christoph Friedrich. 2022. Cross-language transfer of high-quality annotations: Combining neural machine translation with cross-linguistic span alignment to apply ner to clinical texts in a low-resource language. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 53–62.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. Biobertpt-a portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72.

Amogh Kamat Tarcar, Aashis Tiwari, Vineet Naique Dhaimodker, Penjo Rebelo, Rahul Desai, and Dattaraj Rao. 2019. Healthcare ner models using language model pretraining. *arXiv preprint arXiv:1910.11241*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, et al. 2020. Cord-19: The covid-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.