viruses

MDPI

Article

Clinical Application of Detecting COVID-19 Risks: A Natural Language Processing Approach

Syed Raza Bashir [1], Shaina Raza [2,*], Veysel Kocaman [3] and Urooj Qamar [4]

[1] Department of Computer Science, Toronto Metropolitan University, Toronto, ON M5B 2K3, Canada
[2] Dalla Lana School of Public Health, University of Toronto, Toronto, ON M5T 3M7, Canada
[3] Data Science, John Snow Labs Inc., Lewes, DE 19958, USA
[4] Institute of Business & Information Technology, University of the Punjab, Lahore 54590, Pakistan
* Correspondence: shaina.raza@utoronto.ca

Abstract: The clinical application of detecting COVID-19 factors is a challenging task. The existing named entity recognition models are usually trained on a limited set of named entities. Besides clinical, the non-clinical factors, such as social determinant of health (SDoH), are also important to study the infectious disease. In this paper, we propose a generalizable machine learning approach that improves on previous efforts by recognizing a large number of clinical risk factors and SDoH. The novelty of the proposed method lies in the subtle combination of a number of deep neural networks, including the BiLSTM-CNN-CRF method and a transformer-based embedding layer. Experimental results on a cohort of COVID-19 data prepared from PubMed articles show the superiority of the proposed approach. When compared to other methods, the proposed approach achieves a performance gain of about 1–5% in terms of macro- and micro-average F1 scores. Clinical practitioners and researchers can use this approach to obtain accurate information regarding clinical risks and SDoH factors, and use this pipeline as a tool to end the pandemic or to prepare for future pandemics.

Keywords: COVID-19; named entities; clinical; non-clinical; social determinants of health; pipeline; de-identification

Citation: Bashir, S.R.; Raza, S.; Kocaman, V.; Qamar, U. Clinical Application of Detecting COVID-19 Risks: A Natural Language Processing Approach. *Viruses* **2022**, *14*, 2761. https://doi.org/10.3390/v14122761

Academic Editors: Amilcar Tanuri and Luciana Jesus Costa

Received: 28 November 2022
Accepted: 8 December 2022
Published: 11 December 2022

1. Background

The COVID-19 pandemic (coronavirus disease 2019) has had a significant impact on society, due to the severity of the disease and the slow implementation of public health measures [1]. Many of these challenges stem from the information overload problem, which is exacerbated by the growing understanding of the disease and a plethora of literature on the subject [2]. COVID-19 Open Research Dataset (CORD19) [3] and LitCOVID [4] are among the pioneering data sources made available by the research community to aid collaboration between the computing community and the many stakeholders in the COVID-19 pandemic. These data sources contain hundreds of thousands of articles, and new articles are added regularly [1,5]. In its current state, it is difficult for researchers, clinical experts, and practitioners to obtain up-to-date information on the most recent findings.

To study the risk factors associated with COVID-19, government organizations and health sectors can always arrange for human resources to convert the pools of information from the literature into a structured format. However, by the time this data is made accessible to the research community, much of the earlier information is outdated. Natural Language Processing (NLP), a branch of artificial intelligence (AI), allows automated processing and analysis of unstructured texts, such as extracting key information and representing it in a structured format appropriate for computational analysis [6].

The goal of this research is to study the clinical factors, such as disease, drugs, treatments, procedures, and non-clinical factors, such as social determinants of health (SDoH) from the biomedical texts. In terms of methodology, we employ the named entity recognition (NER) [7] task of NLP to extract the biomedical factors from the free texts.

Despite being highly useful, the state-of-the-art work [8–11] in biomedical NER is primarily focused on a small number of entities (disease, chemicals, genes, etc.). There are numerous other clinical factors to consider, such as diagnosis, therapies, medical concepts, risks, and vital signs, as well as non-clinical factors such as SDoH. Extracting these biomedical entities (clinical and non-clinical) is important to study the predictors of COVID-19, which is a motivation for this research. Usually, scientific texts, such as clinical reports, medical notes, and Electronic Health Records (EHR) consist of sensitive patient information that must be de-identified. In this work, we also preserve patients' private information through the data obfuscation process.

We have extended our previous work [12] in this paper and performed a more detailed analysis. We also fine-tune a transformer module to create the task-specific embeddings in this work. Our contributions are listed below:

- We develop a biomedical NER pipeline to identify clinical as well as non-clinical named entities from the COVID-19 texts. We attempt to consolidate and explain data science best practices through this pipeline, with numerous convenient features that can be used as it is or as a starting point for further customization and improvement.
- We develop a new dataset by curating a large number of scientific publications and case reports on COVID-19, and we scientifically parse the text from these scientific articles and prepare a dataset from it. We annotate a part of this dataset on biomedical-named entities to prepare a gold-standard dataset to train the NER pipeline. A portion of the gold-standard dataset is also reserved as a test set.
- We de-identify the patients' personal information after identifying the named entities, thus adhering to the Health Insurance Portability and Accountability Act (HIPAA) [13].
- We demonstrate the efficacy and utility of this pipeline by comparing it with the state-of-the-art methods on public benchmark datasets. We also show the key findings related to COVID-19 in the analysis.

2. Previous Work

Named Entity Recognition (NER) is the task of identifying a named entity (a real-world object or concept) in unstructured text and then classifying the entity into a standard category [7]. In the field of biomedicine, NER is the task of identifying entities such as genes, diseases, chemicals, and proteins [11]. Several datasets are proposed for the NER task. These datasets are prepared usually in the CONLL-2003 format [14], a prototypical format for NER datasets. Many machine learning and deep learning based NER models have also been released in the past few years. Below, we summarize the benchmark datasets and methods used for NER in Table 1:

Table 1. Biomedical NER datasets and methods.

Benchmark Datasets		
Corpus	**Entity Types**	**Data Size**
NCBI-Disease [15]	Diseases	793 PubMed abstracts
BC5CDR [16]	Diseases	1500 PubMed articles
BC5CDR [16]	Chemicals	1500 PubMed articles
BC4CHEMD [17]	Chemicals	10,000 PubMed abstracts
BC2GM [18]	Gene/Proteins	20,000 sentences
JNLPBA [19]	Genes, proteins	2404 abstracts

Table 1. *Cont.*

Benchmark Datasets		
Corpus	**Entity Types**	**Data Size**
i2b2-Clinical [20]	Problem, Treatment, and Test.	426 discharge summaries
I2b2 2012 [21]	Clinical (problems, tests, treatments, clinical departments, occurrences (admission, discharge) and evidence).	310 discharge summaries
Benchmark methods		
Method	**Description**	
BiLSTM-CRF [22]	Bidirectional Long short-term memory (LSTM) and Conditional random field (CRF) architecture for NER.	
BiLSTM-CNN-Char [23]	A hybrid LSTM and Convolutional Neural Network (CNN) architecture that learns both character-level and word-level features for the NER task.	
BiLSTM-CRF-MTL [24]	A multi-task learning (MTL) framework with a BiLSTM-CRF model to collectively use the training data of different types of entities.	
Att-BiLSTM-CRF [25],	Attention (Att) based BiLSTM model with a CRF layer for chemical NER task.	
Doc-Att-BiLSTM-CRF [26]	Document (Doc)-level Attention (Att)-based BiLSTM-CRF network for disease NER task.	
MCNN [27]	A multiple (M) label CNN-based network for disease NER from biomedical literature.	
CollaboNet [28]	A collaboration of deep neural networks, i.e., BiLSTM-CRF with a single task model trained for each specific entity type.	
SciBERT [29]	A pre-trained language model based on Bidirectional Encoder Representations from Transformers (BERT) pretrained on a large multi-domain corpus of scientific publications to improve performance on downstream scientific tasks including NER.	
BioBERT [30]	A pre-trained biomedical language representation model based on BERT for biomedical text mining	

According to the Healthy People 2030 initiative, SDoHs related to population health [31] have a major impact on people's health, well-being, and quality of life, and are related to health outcomes; this is a rather underexplored area of research in biomedicine and clinical research. In this work, we mention some SDOH in our dataset.

In a 2016 survey, nearly 95% of eligible hospitals in the United States use EHRs [32], with that figure expected to rise in these years. The standard EHRs contain 18 categories of critical private information about patients (e.g., name, age, and address), which must be de-identified before they are made public, as required by HIPPA [33]. For the de-identification purpose, researchers used a variety of methods, including rule-based, machine learning-based, and hybrid [34]. The CRF models [35], and Structured Support Vector (SVM) [36] are some of the commonly used models for NER and de-identification tasks. Deep learning models based on recurrent neural networks (RNN) and CNN models are also used for the de-identification of clinical notes [37]. The BioBERT [9], SciBERT [29], and recent Transformer-based models are also used to identify the named entities from biomedical texts.

In this work, we also use deep learning-based methods to build a pipeline for the biomedical NER and de-identification tasks. We identify many biomedical named entities including SDOH from COVID-19 texts.

3. Materials and Methods

3.1. Data Cohort

We have collected the scientific articles and clinical case reports from different journals (Lancet, BMJ, AMJ, Clinical Medicine and related) through LitCOVID [4] API, a resource of scholarly articles. The inclusion and exclusion criteria for data collection are given below:

- We specify the timeline between November 2021 and March 2022 for data collection.
- We specify English as the language to get the publications.
- We exclude many early-pandemic scientific articles, the intuition being that the disease symptoms and diagnosis, drugs and vaccination information were not clear during that time.
- We specify the population groups in adults: 19–44 years, middle-aged: 45–64 years, aged: 65+ years, during data collection.

After obtaining the scientific articles from these sources, we use the Spark OCR [38] library to automatically extract content from the PDF files and convert them into dataframes, where each row corresponds to one document (publication). After all these steps and filtration criteria, we acquired around 15 k scientific articles. Because we specify limited age groups in the population setting, English as the only language, and a time period of 5 months, the number of articles obtained here is lower than those obtained in the actual repository (LitCOVID) during that time period.

Gold-standard dataset: We annotated around 200 scientific articles from our collected dataset using the JohnSnowLabs annotation lab [39], and prepare a gold-standard dataset. A gold-standard dataset [40] means a corpus of text or a set of documents, annotated or tagged with the desired labels by expert annotators. We use the application of active learning [41] to re-annotate a larger portion of the data, where we specified the gold-standard data as the seed. By the end of this step, we acquired around 500 articles that were annotated. According to research [42], this amount of data is sufficient to begin training an NLP model. We used the following named entities, shown in Table 2, as the gold labels. We saved this data in CONLL [14] format.

Table 2. Biomedical entities used in this study.

Entity Type	Entities
Clinical name entities	Admission (patient admission status), oncology (tumor/cancer), blood pressure, respiration (e.g., shortness of breath), dosage (amount of medicine/drug taken), vital signs, symptoms, kidney disease, temperature (body), diabetes, vaccine, time (days, weeks or so), obesity (status), BMI, height (of patient), heart disease, pulse, hypertension, drug name, cerebrovascular disease, disease, treatment, clinical department, weight (of patient), admission/discharge (from hospital), modifier (modifies the current state), external body part, test, strength, route, test result.
Non-clinical entities	Name (of patient), location, date, relative date, duration, relationship status, social status, family history (family members, alone, with family, homeless), employment status, race/ethnicity, gender, social history, sexual orientation, diet (food type, nutrients, minerals), alcohol, smoking.

3.2. Biomedical Named Entity Recognition Pipeline Structure

In this study, we propose a trainable ML pipeline that includes a pre-processor, tokenizer, embedding component, a deep neural network based on BiLSTM, CNN and CRF models, and a de-identifier. The novelty of this approach lies in the subtle integration of different components that are stacked together to train the pipeline. We build this pipeline following the Spark ML pipeline [43], which provides a default scalable solution without requiring much computation power [44]. The workflow of this pipeline is shown in Figure 1.

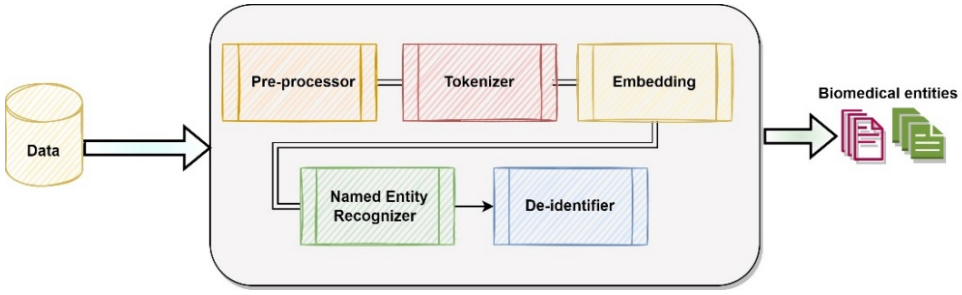

Figure 1. Biomedical Pipeline.

Next, we explain each component of this pipeline.

Data Collection: The input to the pipeline can be any raw textual data. We provided the data from our data cohort for this purpose.

Pre-processor: The preprocessor takes the text data as an input that comes from the data collection phase, pre-processes it, and detects the sentence boundaries in each record (document). Then, it transforms the data into a format that is readable by the next stage in the pipeline. The output from the pre-processor is the set of records that are pre-processed.

Tokenizer: The tokenizer takes the pre-processed data from the pre-processor as input. Tokenization is the process of breaking the input text into smaller chunks (words, or sentences) called tokens [45]. These tokens aid in comprehending the context and in developing the NLP model. The output from the tokenizer is transformed data, containing the tokens (words) corresponding to each document (scientific article, case report and so on).

Embedding: The tokenized data from the tokenizer goes into the embedding component, which maps tokens to vectors. We have fine-tuned the pre-trained BlueBERT model [46] that is trained on PubMed abstracts and MIMIC-III [47] on our gold-data to provide task-specific embeddings.

Named Entity Recognizer: This component identifies biomedical entities in the text. This is an algorithm based on the BiLSTM-CNN-CRF [48] model. We modify the vanilla BiLSTM-CNN-CRF for the task-specific embeddings and make our modifications. We introduce our NER model in Figure 2 and explain its working below.

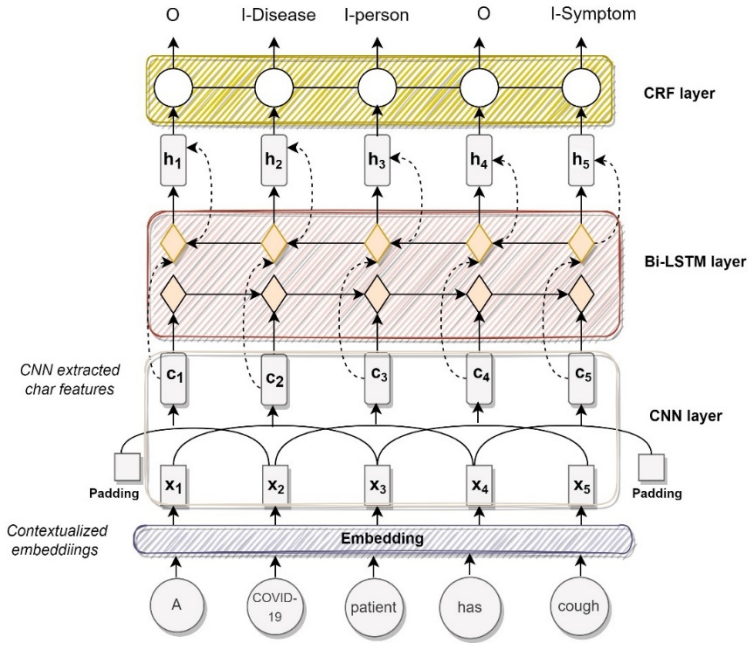

Figure 2. Named Entity Recognition algorithm.

As shown in Figure 2, the algorithm takes as input the sequence of words or a sentence. This sequence is represented as $s = [w_1, w_2, \ldots, w_N]$, where N is the sentence length and $w_i \in R^V$ is the i_{th} token in the sequence. This input goes to the embedding layer.

The embedding layer is the first layer in the model that converts a sentence from a sequence of tokens into a sequence of dense vectors. In this work, we use our fine-tuned transformer model for the embeddings. The output of this layer is a sequence of vectors $x = [x_1, x_2, \ldots, x_N]$, where $x_i = \mathrm{E}w_i \in R^D$, E is for embedding and x_i is the dense vector representation of word w_i.

The second layer in this model is a CNN layer that is used to capture local information within given words in a biomedical context. The CNN is just for char embeddings to represent letters. The main feature is word embeddings coming from the embedding layer (BERT-based embeddings). The output of the CNN layer is $c = [c_1, c_2, \ldots, c_N]$, where $c_i \in R^M$, M is the number of filters. The contextual representation c_i of the i_{th} character is the concatenation of the outputs of all filters at this position

The third layer in the model is the Bi-LSTM network, which is used to learn hidden representations of characters or tokens in a sequence using all of the previous contexts (in both directions). The output of the Bi-LSTM layer is $h = [h_1, h_2, \ldots, h_N]$, where $h_i = R^{2S}$ and S is the dimension of hidden states in LSTM.

The fourth layer on the top of the Bi-LSTM network is the CRF layer [49]. The input to the CRF layer is the hidden representations of characters $h = [h_1, h_2, \ldots, h_N]$ generated by the Bi-LSTM layer. To ensure that the predicted labels are valid, the CRF layer captures the dependency relationship between the named tags and constrains them to the final predicted labels [22]. The output of the CRF layer is $y = [y_1, y_2, \ldots, y_N]$, which is a label sequence of sentence s, where $y_i \in R^L$ is the one-hot representation of the i_{th} character's label and L is the number of labels. In this work, the biomedical entities are the labels. A tanh layer on top of the BiLSTM layer is added to predict the confidence scores (CS) for the word with each of the possible labels as the output score of the network.

De-identifier: We use the data obfuscation technique, which is a process that obscures (masks) the meaning of data [50]. For example, to replace identified names with different fake names or to mask some data, value <02-02-2022> with <DATE> is used. This component provides HIPAA [13] compliance when dealing with text documents containing any protected health information. We use the pre-trained de-identification model from Johnsnowlabs [51] and embed it inside the pipeline to de-identify the personal records of the patients.

Biomedical Named Entities: The output of the pipeline is the biomedical entities, shown in Table 2.

3.3. Evaluation

We adopted a two-fold evaluation technique: (1) to evaluated the accuracy of the proposed approach, and (2) to analyze the results of our approach for pandemic surveillance. To evaluate the accuracy of the proposed approach, we considered a number of baseline methods and benchmark datasets including our test set. To evaluate the pandemic surveillance, we analyzed the results of our model and summarized the key findings.

Benchmark datasets: We used the JNLPBA [19] for chemical entities, NCBI-Disease [15] for disease entities, BC5CDR [16] dataset for chemical and disease mentions, BC2GM [18] for genes, and i2b2-Clinical [20] for clinical entities. From here, we obtained datasets that were already available in CoNLL-2003 format [52]. We performed further processing to convert them into IOB (Inside-Outside-Before) [53] scheme. All the datasets were divided into training, validation, and test sets, with a 70:15:15 ratio for all experiments. The Stratified 5-Folds cross-validation (CV) strategy was used for train/test split if original datasets did not have an official train/test split. We also set aside 30% of our gold dataset as a test set.

Baseline Methods: We compared the performance of our approach against the following state-of-the-art baseline methods: BiLSTM-CRF [54], BiLSTM-CRF-MTL [24], CT-BERT [55], SciBERT [29], and BioBERT [9] (v1.0, v1.1, v1.2). All of the baselines were trained

on the aforementioned datasets. Each baseline was tuned to its optimal hyperparameter setting and the best results were reported for each method.

Training environment: All the experiments were run on Google Colab Pro (NVIDIA P100 or T4, 24 GB RAM, 2 x vCPU). The grid search was used to get optimal values for the hyperparameters and early stopping was performed to overcome overfitting. We specified the following hyperparameters as shown in Table 3.

Table 3. Hyperparameters used—optimal parameter (range of values).

Hyperparameter	Optimal Value (Values Used)
Learning rate	1×10^{-3} (1×10^{-2}, 1×10^{-3}, 1×10^{-5}, 2×10^{-5}, 5×10^{-5}, 3×10^{-4})
Batch size	64 (8, 16, 32, 64, 128)
Epochs	30 ({2, 3, . . . , 30})
LSTM state size	200 (200, 250)
Dropout rate	0.5 ({0.3, 0.35, . . . , 0.7})
Optimizer	Adam
CNN filters	2 (2, 3, 4, 5)
Hidden Size	768
Embedding Size	128
Max Seq Length	512
Warmup Steps	3000

Evaluation metrics: Following the standard practice [41,56] to evaluate NER tasks, we used the following metrics:

- Micro-average F1 to measures the F1-score of aggregated contributions of all classes.
- Macro-average F1that adds all the measures (Precision, Recall, or F-Measure) and divides with the number of labels, which is more like an average.

4. Results

The results and analysis are given below.

4.1. Comparison with Baseline Methods

We show the performance of our approach for accuracy in Table 4.

Table 4. Test results using macro-average F1 (macro) and micro-average F1 (micro) scores on all datasets using different methods. The best scores are in bold and the second-best in italic.

Methods/ Dataset	Metric	NCBI	BC5CDR	BC2GM	JNLPBA	i2b2-Clinical	Our Dataset
BiLSTM-CRF	micro	85.80	84.22	78.46	74.29	83.66	87.10
	macro	86.12	85.09	80.01	75.10	84.01	*88.01*
BiLSTM-CRF-MTL	micro	86.46	84.94	80.34	77.03	82.38	88.39
	macro	88.01	85.00	81.12	77.14	83.96	88.97
CT-BERT	micro	77.50	76.85	74.10	68.00	77.07	78.10
	macro	78.50	77.96	75.37	68.98	78.01	78.98
SciBERT	micro	82.88	82.94	84.08	75.77	78.19	80.95
	macro	83.32	83.13	85.84	77.01	79.10	81.14
BioBERT-Base v1.0	micro	84.01	86.56	78.68	86.28	85.87	84.01
	macro	79.10	78.90	79.00	78.13	72.18	79.10
BioBERT-Base v1.1	micro	88.52	87.15	79.39	76.16	86.27	88.52
	macro	85.89	87.10	87.18	*75.45*	*87.78*	85.89
BioBERT-Base v1.2	micro	*89.12*	*87.81*	*83.34*	*76.45*	*86.88*	*89.12*
	macro	*86.78*	*87.89*	*86.07*	75.15	86.98	86.78
Our approach	micro	**90.58**	**89.90**	**89.15**	**79.92**	**89.10**	**94.78**
	macro	**91.83**	**90.34**	**90.38**	**80.94**	**90.48**	**95.37**

Overall, these results show that our approach achieved state-of-the-art performance on five public biomedical benchmarks, as well as on our dataset designed specifically for biomedical named entities. This demonstrates the generalizability of our methodology across different domains.

Our approach achieved the best micro F1 score of 94.78 on our dataset (52 entities), 90.58 on NCBI Disease (disease entity), 89.90 on BC5CDR (chemicals), 89.15 on BC2GM (gene/proteins), 79.92 on JNLBPA (chemical) and 89.10 on the i2b2 (clinical) dataset. We see similar patterns and higher performances in our pipeline for macro F1 scores.

The BioBERT model shows competitive performance in these results. Among the variants of the BioBERT, we see overall better performance of BioBERT v1.2 than its other variants, except for a few places, where BioBERT v1.1 marginally outperforms BioBERT v1.2. The better performance of BioBERT v1.2 is attributed to its training method, which is the same method as BioBERT v1.1 but includes an LM head [57]. Among the BERT-based models (BioBERT, SciBERT, CT-BERT), BioBERT performs best. The BioBERT is quite generalizable compared to other BERT-based methods, the SciBERT is initially trained on scientific data (not clinical) [29], and, CT-BERT is pre-trained on social media data, so they perform differently with different entity types.

Among the BiLSTM-based models (BiLSTM-CRF, BiLSTM-CRF-MTL), we observe the good performance of the BiLSTM-CRF model in identifying many diseases, chemicals, and gene/protein entities in these experiments. Our algorithm (BiLSTM-CNN-CRF) performs better than the BiLSTM-CRF baseline, probably because we are using biomedical embeddings on top of char-level embeddings. The fine-tuned transformer model's embeddings enhance the performance of our model.

Although we fine-tuned each baseline method to its optimal hyperparameter settings, we anticipate that the relatively low scores of these baselines on our dataset can be attributed to the following: (i) the absence of an annotated dataset for training new biomedical entities, and (ii) different training/test set splits used in previous works that were unavailable.

Ablation Study: We performed an ablation experiment in which we evaluated the component of our pipeline. This component was based on our modified BiLSTM-CNN-CRF model. We replaced the standard BiLSTM-CNN-CRF in the sequence labeling architecture (Figure 2) with a direct feedforward map with and without a CRF decoder. We used a simple linear map over the embeddings to determine their direct information content. The results of this ablation study on our test set, based on macro average F1-score, are shown in Table 5.

Table 5. Ablation study of the model. Bold shows best macro-average F1 score.

Model	Macro
BiLSTM-CNN-CRF	**94.18 ± 0.12**
BiLSTM-CNN	87.37 ± 0.02
Map-CNN-CRF	80.55 ± 0.03
Map-CNN	69.25 ± 0.04

The results, in Table 5, show that the effect of removing the BiLSTM layer is far more than removing the CRF layer from BiLSTM-CNN-CRF. This is shown with a dropped macro F1 of more than 15% when we remove the BiLSTM layer, compared to removing only the CRF layer. The most impacted performance is seen with Map-CNN where we removed these two layers (BiLSTM and CRF). With all these results, we find that our default settings are best in this setup.

4.2. Pandemic Surveillance

In this section, we demonstrate the effectiveness of our approach in demonstrating the key findings on pandemic surveillance. First, we show the most common entity types predicted by our approach after parsing 500 case reports, and show the performance of the model in terms of precision, recall, F1-score (F1), micro-average and macro-average in

Table 6. The formulae for these performance metrics are based on true positives (TP), false positives (FP) and false negatives (FN).

Table 6. Performance of most used entity from random 500 case reports.

Entity	TP	FP	FN	Prec	Recall	F1
Disease	818	98	112	0.89	0.88	0.89
Gender	390	78	101	0.83	0.79	0.81
Employment	234	29	132	0.89	0.64	0.74
Race_Ethnicity	334	65	96	0.84	0.78	0.81
Smoking	309	24	97	0.93	0.76	0.84
Psychological_Condition	218	29	58	0.88	0.79	0.83
Death_Entity	387	34	103	0.92	0.79	0.85
BMI	146	12	29	0.92	0.83	0.88
Diabetes	157	10	28	0.94	0.85	0.89
Macro-average	2993	379	756	0.89	0.79	0.84
Micro-average	2993	379	756	0.89	0.80	0.84

As seen in Table 6, we can accurately predict a large number of entities with quite a high score. We also show the prevalence of the most common symptoms observed in our data in Figure 3.

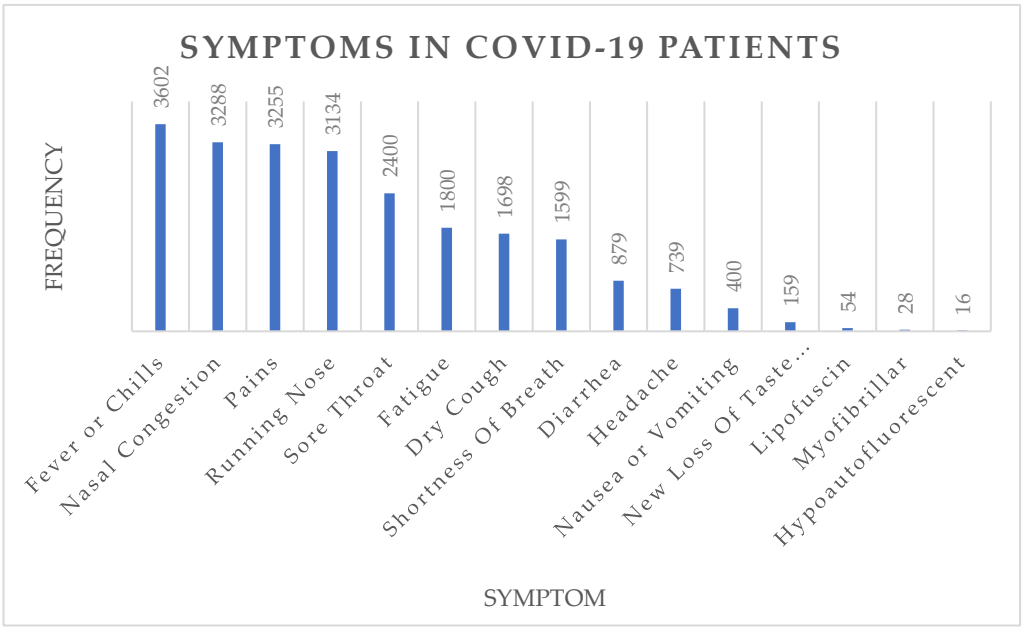

Figure 3. Most common symptoms of COVID-19 patients. Number at the top of each bar represents the number of times the symptoms were mentioned in test set.

The results in Figure 3 show that fever, nasal congestion, pains, a running nose, and sore throat are among the most common COVID-19 symptoms. Next, we show the most occurring named entities (occurrence > 70%) under the prominent entity types (drugs, vaccines, treatments) and show the results in Table 7.

Table 7. Most prevalent named entities under entity types (drugs, vaccine, treatments).

Drugs	Vaccine	Non-Medical Treatments
Hydroxychloroquine	Pfizer-BioNTech	Isolation
Paxlovid	Moderna	Wear masks
Actemra	AstraZeneca	Vaccination
Immunomodulators	CoronaVac	Oxygen support
Steroid	BBIBP-CorV	Medication
Amoxicillin	Janssen	Hand sanitization

We also gave a snippet from a COVID-19 related case report to our pipeline and show the confidence score for the predicted entities. The results are shown in Table 8.

Table 8. Test Results on all Datasets using different Methods.

Sentence	Begin	End	Chunks	Biomedical Entity	Confidence
0	2	12	73-year-old	Age	1.00
0	14	18	woman	Gender	1.00
0	32	43	Fever Clinic	Clinical Department	0.98
0	52	65	First Hospital	Clinical Department	0.51
0	109	134	Fever, temperature	Symptom	0.80
0	156	160	Cough	Symptom	0.99
0	163	175	Expectoration	Symptom	1.00
0	178	196	Shortness of breath	Symptom	0.39
0	203	218	General weakness	Symptom	0.77
0	233	244	Prior 5 days	Relative Date	0.42
1	247	249	She	Gender	1.00
1	261	264	Mild	Modifier	0.90
1	266	273	Diarrhea	Symptom	1.00
1	280	289	Stools/day	Symptom	0.85
1	292	303	2 days prior	Relative Date	0.68
1	322	329	Hospital	Clinical Department	1.00
1	386	402	COVID-19 positive	Disease Syndrome	0.90
1	436	454	Healthcare provider	Employment	0.94
2	486	494	Cirrhosis	Disease Syndrome	0.96
2	500	514	Type 2 diabetes	Diabetes	0.95
2	535	541	Smoking	Smoking	1.00
2	546	553	Drinking	Alcohol	0.93

The result in Table 8 shows that our model can predict many named entities with a high level of confidence score.

We take the nominal race groups [58] and report the results where the race group accounts for more than 5% of the population. This finding shown in Figure 4 is based on a subset of available data from a specific time period, so it may not be an accurate representation of racial groups as a whole during the COVID-19 outbreak.

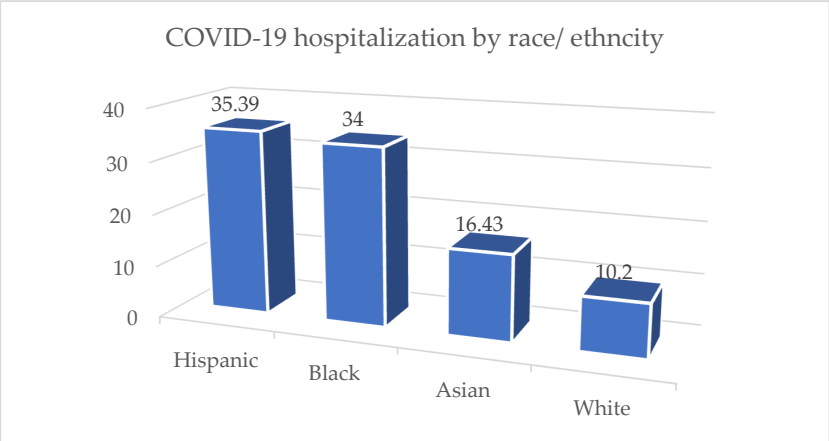

Figure 4. COVID-19 hospitalization by race and ethnicity.

We show a sample prediction of our model on a case report [59] in Figure 5, where we can see that many clinical and SDOH are being detected.

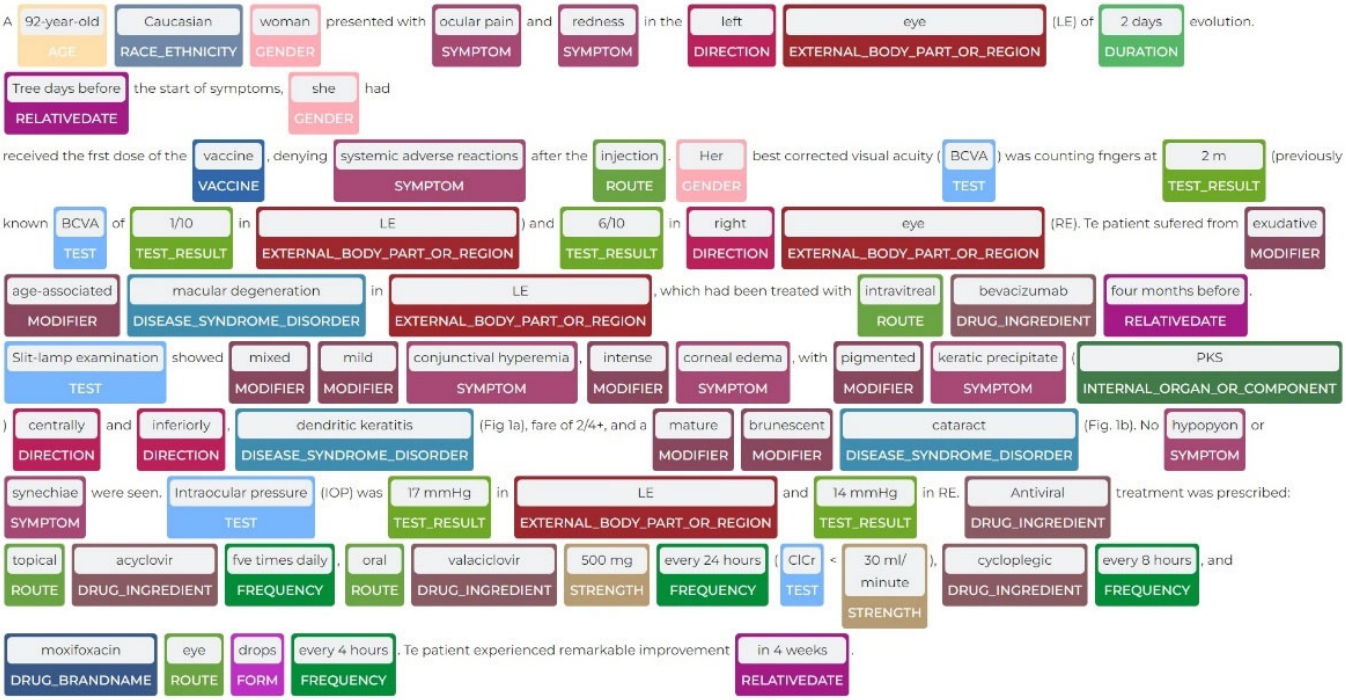

Figure 5. Biomedical entities recognized by proposed pipeline.

5. Discussion

5.1. Implications in Healthcare

There are many different ways that this pipeline can be used in healthcare settings. These biomedical entity types can assist physicians, nurses, and other healthcare professionals in matching symptoms to a diagnosis, a course of treatment, and follow-up. Health disparities can be decreased by tracking social determinants [60]. The clinical data can be converted into knowledge, evidence, and clinical impact using this research as well. This pipeline emphasizes best practices, openness, reproducibility, automation, and the capacity to recognize complex named entities from biomedical texts. With little to no code modification, this pipeline can also be applied to any other domain.

5.2. Transfer Learning

The advantages of transfer learning in detecting COVID-19-named entities become clearer because of this work. The proposed approach (combining BiLSTM-CRF-CNN with Transformer-based embeddings) achieves a performance comparable to pure Transformer-based models (BioBERT), and performs at least 1 to 5% better compared to conventional BiLSTM models. In the future, it would be beneficial to have our own pre-trained embeddings that can be used to study a large number of clinical and non-clinical entities.

5.3. Limitations

Although the BiLSTM-CNN-CRF model that we used for this approach showed good results and outperformed the current state-of-the-art solutions, there is still room for improvement, and the following points are what we would consider implementing in the future: first, we plan to increase the number of layers in this deep neural network. We intend to pre-train a transformer-based model. In this regard, one approach would be to first prepare more data for annotation and then pre-train the model on the annotated data.

So far, we have annotated a portion of the dataset, which suffices for the purpose of model training. In the future, we strongly encourage the inclusion of medical professionals

in the annotation guideline. We also plan to annotate a large number of documents for this type of study.

We also plan to test the model on additional benchmark datasets. Furthermore, we intend to curate more clinical data; in particular, getting real-time access to EHRs would be helpful. Since we are already providing a de-identifier to de-identify patients' personal information through this pipeline, we hope to gain access to such a dataset soon while adhering to HIPAA guidelines. Lastly, due to the black-box nature of most deep neural networks, we also plan to handle bias or systematic error in research methods, which may influence disease associations and predictions.

6. Conclusions

In conclusion, this paper presents a pipeline that consists of a number of ML components stacked together. We used an approach to train models for the biomedical named entities using the BiLSTM-CNN-CRF model plus BERT-based embeddings. This paper shows that using contextualized word embedding, pre-trained on biomedical corpora, significantly improves the results of biomedical NER tasks. We evaluated the performance of this approach on benchmark datasets and our own test set, and our approach achieved the state-of-the-art results compared to the baselines. This pipeline can be used in different health science settings, provided that the annotated data to train the model and the pipeline is available.

Author Contributions: Conceptualization, S.R.B., S.R. and V.K.; methodology, S.R.B., S.R. and V.K.; software, V.K., Johnsnowlabs; validation, S.R.K, S.R. and V.K.; formal analysis, S.R.B.; investigation, S.R.; resources, S.R.B. and S.R.; data curation, S.R.; writing—original draft preparation, S.R.B. and S.R.; writing—review and editing, S.R.B. and S.R.; visualization, S.R. and V.K.; supervision, S.R.; project administration, S.R.B.; funding acquisition, S.R.B. and U.Q. performed additional experiments, proofread the document and validated the analysis. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data can be made available upon request from corresponding author.

Acknowledgments: The earlier version of this work is published in Proceedings of Machine Learning Research [12].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chen, Q.; Leaman, R.; Allot, A.; Luo, L.; Wei, C.-H.; Yan, S.; Lu, Z. Artificial Intelligence in Action: Addressing the COVID-19 Pandemic with Natural Language Processing. *Annu. Rev. Biomed. Data Sci.* **2021**, *4*, 313–339. [CrossRef] [PubMed]
2. Raza, S.; Schwartz, B.; Rosella, L.C. CoQUAD: A COVID-19 Question Answering Dataset System, Facilitating Research, Benchmarking, and Practice. *BMC Bioinform.* **2022**, *23*, 210. [CrossRef] [PubMed]
3. Allen Institute. COVID-19 Open Research Dataset Challenge (CORD-19). 2020. Available online: https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge (accessed on 27 November 2022).
4. Chen, Q.; Allot, A.; Lu, Z. LitCovid: An Open Database of COVID-19 Literature. *Nucleic Acids Res.* **2021**, *49*, D1534–D1540. [CrossRef] [PubMed]
5. Wang, L.L.; Lo, K. Text Mining Approaches for Dealing with the Rapidly Expanding Literature on COVID-19. *Brief. Bioinform.* **2021**, *22*, 781–799. [CrossRef]
6. Reeves, R.M.; Christensen, L.; Brown, J.R.; Conway, M.; Levis, M.; Gobbel, G.T.; Shah, R.U.; Goodrich, C.; Rickets, I.; Minter, F.; et al. Adaptation of an NLP System to a New Healthcare Environment to Identify Social Determinants of Health. *J. Biomed. Inform.* **2021**, *120*, 103851. [CrossRef]
7. Nadeau, D.; Sekine, S. A Survey of Named Entity Recognition and Classification. *Lingvisticae Investig.* **2007**, *30*, 3–26. [CrossRef]
8. Boudjellal, N.; Zhang, H.; Khan, A.; Ahmad, A.; Naseem, R.; Shang, J.; Dai, L. ABioNER: A BERT-Based Model for Arabic Biomedical Named-Entity Recognition. *Complexity* **2021**, *2021*, 6633213. [CrossRef]

9. Dmis-Lab. DMIS-Lab/Biobert: Bioinformatics'2020: BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. GitHub. 2020. Available online: https://github.com/dmis-lab/biobert (accessed on 27 November 2022).
10. Perera, N.; Dehmer, M.; Emmert-Streib, F. Named Entity Recognition and Relation Detection for Biomedical Information Extraction. *Front. Cell Dev. Biol.* **2020**, *8*, 673. [CrossRef]
11. Cho, H.; Lee, H. Biomedical Named Entity Recognition Using Deep Neural Networks with Contextual Information. *BMC Bioinform.* **2019**, *20*, 735. [CrossRef]
12. Raza, S.; Schwartz, B. Detecting Biomedical Named Entities in COVID-19 Texts. In Proceedings of the 1st Workshop on Healthcare AI and COVID-19, ICML 2022, Baltimore, MA, USA, 17–23 July 2022.
13. Nosowsky, R.; Giordano, T.J. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule: Implications for Clinical Research. *Annu. Rev. Med.* **2006**, *57*, 575–590. [CrossRef]
14. Sang, E.F.; De Meulder, F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv* **2003**, arXiv:cs/0306050.
15. Doğan, R.I.; Leaman, R.; Lu, Z. NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization. *J. Biomed. Inform.* **2014**, *47*, 1–10. [CrossRef] [PubMed]
16. Li, J.; Sun, Y.; Johnson, R.J.; Sciaky, D.; Wei, C.-H.; Leaman, R.; Davis, A.P.; Mattingly, C.J.; Wiegers, T.C.; Lu, Z. BioCreative V CDR Task Corpus: A Resource for Chemical Disease Relation Extraction. *Database* **2016**, *2016*, baw068. [CrossRef] [PubMed]
17. Krallinger, M.; Rabal, O.; Leitner, F.; Vazquez, M.; Salgado, D.; Lu, Z.; Leaman, R.; Lu, Y.; Ji, D.; Lowe, D.M.; et al. The CHEMDNER Corpus of Chemicals and Drugs and Its Annotation Principles. *J. Cheminform.* **2015**, *7*, S2. [CrossRef] [PubMed]
18. Smith, L.; Tanabe, L.K.; Kuo, C.-J.; Chung, I.; Hsu, C.-N.; Lin, Y.-S.; Klinger, R.; Friedrich, C.M.; Ganchev, K.; Torii, M.; et al. Overview of BioCreative II Gene Mention Recognition. *Genome Biol.* **2008**, *9*, S2. [CrossRef]
19. Collier, N.; Kim, J.-D. Introduction to the Bio-Entity Recognition Task at JNLPBA. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP), Geneva, Switzerland, 28–29 August 2004; pp. 73–78.
20. Uzuner, Ö.; South, B.R.; Shen, S.; DuVall, S.L. 2010 I2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text. *J. Am. Med. Inform. Assoc.* **2011**, *18*, 552–556. [CrossRef]
21. Sun, W.; Rumshisky, A.; Uzuner, O. Evaluating Temporal Relations in Clinical Text: 2012 I2b2 Challenge. *J. Am. Med. Inform. Assoc.* **2013**, *20*, 806–813. [CrossRef]
22. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural Architectures for Named Entity Recognition. *arXiv* **2016**, arXiv:1603.01360.
23. Chiu, J.P.C.; Nichols, E. Named Entity Recognition with Bidirectional LSTM-CNNs. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 357–370. [CrossRef]
24. Wang, X.; Zhang, Y.; Ren, X.; Zhang, Y.; Zitnik, M.; Shang, J.; Langlotz, C.; Han, J. Cross-Type Biomedical Named Entity Recognition with Deep Multi-Task Learning. *Bioinformatics* **2019**, *35*, 1745–1752. [CrossRef]
25. Luo, L.; Yang, Z.; Yang, P.; Zhang, Y.; Wang, L.; Lin, H.; Wang, J. An Attention-Based BiLSTM-CRF Approach to Document-Level Chemical Named Entity Recognition. *Bioinformatics* **2018**, *34*, 1381–1388. [CrossRef] [PubMed]
26. Xu, K.; Yang, Z.; Kang, P.; Wang, Q.; Liu, W. Document-Level Attention-Based BiLSTM-CRF Incorporating Disease Dictionary for Disease Named Entity Recognition. *Comput. Biol. Med.* **2019**, *108*, 122–132. [CrossRef] [PubMed]
27. Zhao, Z.; Yang, Z.; Luo, L.; Wang, L.; Zhang, Y.; Lin, H.; Wang, J. Disease Named Entity Recognition from Biomedical Literature Using a Novel Convolutional Neural Network. *BMC Med. Genom.* **2017**, *10*, 75–83. [CrossRef] [PubMed]
28. Yoon, W.; So, C.H.; Lee, J.; Kang, J. Collabonet: Collaboration of Deep Neural Networks for Biomedical Named Entity Recognition. *BMC Bioinform.* **2019**, *20*, 55–65. [CrossRef] [PubMed]
29. Beltagy, I.; Lo, K.; Cohan, A. SCIBERT: A Pretrained Language Model for Scientific Text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2020; pp. 3615–3620.
30. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef]
31. Office of Disease Prevention and Health Promotion Social Determinants of Health. Healthy People 2030. Available online: https://health.gov/healthypeople/objectives-and-data/social-determinants-health (accessed on 7 October 2021).
32. Toscano, F.; O'Donnell, E.; Unruh, M.A.; Golinelli, D.; Carullo, G.; Messina, G.; Casalino, L.P. Electronic Health Records Implementation: Can the European Union Learn from the United States? *Eur. J. Public Health* **2018**, *28*, cky213.401. [CrossRef]
33. Fernández-Calienes, R. Health Insurance Portability and Accountability Act of 1996. In *Encyclopedia of the Fourth Amendment*; CQ Press: Washington, DC, USA, 2013.
34. Meystre, S.M.; Friedlin, F.J.; South, B.R.; Shen, S.; Samore, M.H. Automatic De-Identification of Textual Documents in the Electronic Health Record: A Review of Recent Research. *BMC Med. Res. Methodol.* **2010**, *10*, 70. [CrossRef]
35. Lafferty, J.; Mccallum, A.; Pereira, F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data Abstract. In Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001), Williamstown, MA, USA, June 28–1 July 2001; pp. 282–289.
36. Tsochantaridis, I.; Joachims, T.; Hofmann, T.; Altun, Y.; Singer, Y. Large Margin Methods for Structured and Interdependent Output Variables. *J. Mach. Learn. Res.* **2005**, *6*, 1453–1484.

37. Yang, X.; Lyu, T.; Li, Q.; Lee, C.Y.; Bian, J.; Hogan, W.R.; Wu, Y. A Study of Deep Learning Methods for De-Identification of Clinical Notes in Cross-Institute Settings. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 232. [CrossRef]
38. John Snow Labs. Spark OCR. Available online: https://nlp.johnsnowlabs.com/docs/en/ocr (accessed on 27 November 2022).
39. *Annotation Lab*; John Snow Labs: Lewes, DE, USA, 2022.
40. Ogren, P.V.; Savova, G.K.; Chute, C.G. Constructing Evaluation Corpora for Automated Clinical Named Entity Recognition. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 26 May–1 June 2008; pp. 3143–3150.
41. Chen, Y.; Lasko, T.A.; Mei, Q.; Denny, J.C.; Xu, H. A Study of Active Learning Methods for Named Entity Recognition in Clinical Text. *J. Biomed. Inform.* **2015**, *58*, 11–18. [CrossRef]
42. Snow, R.; O'Connor, B.; Jurafsky, D.; Ng, A.Y. Cheap and Fast—But Is It Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, USA, 25–27 October 2008; pp. 254–263.
43. ML Pipelines—Documentation. Available online: https://spark.apache.org/docs/latest/ml-pipeline.html (accessed on 27 November 2022).
44. Kocaman, V.; Talby, D. Spark NLP: Natural Language Understanding at Scale. *Softw. Impacts* **2021**, *8*, 100058. [CrossRef]
45. Webster, J.J.; Kit, C. Tokenization as the Initial Phase in NLP. In Proceedings of the COLING 1992. The 14th International Conference on Computational Linguistics, Nantes, France, 23–28 August 1992; Volume 4.
46. Peng, Y.; Yan, S.; Lu, Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. *arXiv* **2019**, arXiv:1906.05474.
47. Johnson, A.E.W.; Pollard, T.J.; Shen, L.; Lehman, L.W.H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; Mark, R.G. MIMIC-III, a Freely Accessible Critical Care Database. *Sci. Data* **2016**, *3*, 160035. [CrossRef] [PubMed]
48. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv* **2015**, arXiv:1508.01991.
49. Ma, X.; Hovy, E. End-to-End Sequence Labeling via Bi-Directional Lstm-Cnns-Crf. *arXiv* **2016**, arXiv:1603.01354.
50. Bakken, D.E.; Rarameswaran, R.; Blough, D.M.; Franz, A.A.; Palmer, T.J. Data Obfuscation: Anonymity and Desensitization of Usable Data Sets. *IEEE Secur. Priv.* **2004**, *2*, 34–41. [CrossRef]
51. *Medical Data De-Identification—John Snow Labs*; John Snow Labs: Lewes, DE, USA, 2022.
52. GitHub. ay94 NER-Datasets. 2022. Available online: https://github.com/ay94/NER-datasets (accessed on 27 November 2022).
53. Sexton, T. *IOB Format Intro*; Nestor: Gaithersburg, MA, USA, 2022.
54. Akbik, A.; Blythe, D.; Vollgraf, R. Contextual String Embeddings for Sequence Labeling. In Proceedings of the 27th international Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 1638–1649.
55. Müller, M.; Salathé, M.; Kummervold, P.E. Covid-Twitter-Bert: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. *arXiv* **2020**, arXiv:2005.07503.
56. Tsai, R.T.-H.; Wu, S.-H.; Chou, W.-C.; Lin, Y.-C.; He, D.; Hsiang, J.; Sung, T.-Y.; Hsu, W.-L. Various Criteria in the Evaluation of Biomedical Named Entity Recognition. *BMC Bioinform.* **2006**, *7*, 92. [CrossRef]
57. Perone, C.S.; Silveira, R.; Paula, T.S. Evaluation of Sentence Embeddings in Downstream and Linguistic Probing Tasks. *arXiv* **2018**, arXiv:1806.06259.
58. Abdi, S.; Bennett-AbuAyyash, C.; MacDonald, L.; Hohenadel, K.; Johnson, K.O.; Leece, P. Provincial Implementation Supports for Socio-Demographic Data Collection during COVID-19 in Ontario's Public Health System. *Can. J. Public Health* **2021**, *112*, 853–861. [CrossRef]
59. Ortiz-Egea, J.M.; Sánchez, C.G.; López-Jiménez, A.; Navarro, O.D. Herpetic Anterior Uveitis Following Pfizer–BioNTech Coronavirus Disease 2019 Vaccine: Two Case Reports. *J. Med. Case Rep.* **2022**, *16*, 127. [CrossRef] [PubMed]
60. Raza, S. A Machine Learning Model for Predicting, Diagnosing, and Mitigating Health Disparities in Hospital Readmission. *Healthc. Anal.* **2022**, *2*, 100100. [CrossRef]