applied sciences

MDPI

*Article*

# Survey of BERT-Base Models for Scientific Text Classification: COVID-19 Case Study

**Mayara Khadhraoui [1,2,*], Hatem Bellaaj [2], Mehdi Ben Ammar [3,4], Habib Hamam [4,5,6,7] and Mohamed Jmaiel [2]**

[1] National Engineering School of Sfax (ENIS), University of Sfax, Sfax 3038, Tunisia
[2] ReDCAD Laboratory, Department of Computer Engineering and Applied Mathematics, University of Sfax, Sfax 3029, Tunisia; hatem.bellaaj@redcad.org (H.B.); mohamed.jmaiel@redcad.org (M.J.)
[3] Solutions Galore Inc., Moncton, NB E1C 5Y1, Canada; mehdi.benammar@gmail.com
[4] Faculty of Engineering, Université de Moncton, Moncton, NB E1A 3E9, Canada; habib.hamam@umoncton.ca
[5] International Institute of Technology and Management, Commune d'Akanda, Libreville BP 1989, Gabon
[6] Spectrum of Knowledge Production & Skills Development, Sfax 3027, Tunisia
[7] Department of Electrical and Electronic Engineering Science, School of Electrical Engineering, University of Johannesburg, Johannesburg 2006, South Africa
* Correspondence: khadhraouimayara@gmail.com

**Abstract:** On 30 January 2020, the World Health Organization announced a new coronavirus, which later turned out to be very dangerous. Since that date, COVID-19 has spread to become a pandemic that has now affected practically all regions in the world. Since then, many researchers in medicine have contributed to fighting COVID-19. In this context and given the great growth of scientific publications related to this global pandemic, manual text and data retrieval has become a challenging task. To remedy this challenge, we are proposing CovBERT, a pre-trained language model based on the BERT model to automate the literature review process. CovBERT relies on prior training on a large corpus of scientific publications in the biomedical domain and related to COVID-19 to increase its performance on the literature review task. We evaluate CovBERT on the classification of short text based on our scientific dataset of biomedical articles on COVID-19 entitled COV-Dat-20. We demonstrate statistically significant improvements by using BERT.

**Keywords:** BERT; COVID-19; scientific text classification; transfer learning; scientific publications; deep learning

check for updates

**Citation:** Khadhraoui, M.; Bellaaj, H.; Ammar, M.B.; Hamam, H.; Jmaiel, M. Survey of BERT-Base Models for Scientific Text Classification: COVID-19 Case Study. *Appl. Sci.* **2022**, *12*, 2891. https://doi.org/10.3390/app12062891

Academic Editors: Stefano Silvestri and Francesco Gargiulo

Received: 14 December 2021
Accepted: 7 March 2022
Published: 11 March 2022

## 1. Introduction

Since December 2019 and possibly before, the world has faced one of the most serious dangers in its history: the new coronavirus pandemic [1]. By the end of February 2022, the world had recorded an affected population of more than 430 million people and around 6 million deaths [2]. The entire world is on high alert to find a radical solution to this pandemic and to minimize the cases of death.

### 1.1. Context

We collaborated with epidemiologists to help medical researchers accelerate research on COVID-19. The researchers appreciated having relevant data exposed and classified. Research laboratories in epidemiology constantly need to collect relevant and pertinent data to be able to analyze an epidemic (or pandemic), predict the evolution of contamination in the population in question, determine the shape of the virus, discover its genome sequencing, produce a vaccine, develop a drug, conceive a screening method or implement a medical device dedicated to the epidemic (intensive care, etc.). We are aware of the time and effort put in by medical researchers to develop their epidemiological studies. The manual pre-screening by researchers of all the information available on the web, in specialized datasets and other research, can be overwhelming as well as time and energy consuming. Search engines provide an astronomical amount of information; many are unclassified and irrelevant to the researcher's query. Therefore, we aimed at providing epidemiology researchers with a configurable open-source platform for automatic pertinent data retrieval and classification from scientific abstracts and full papers on COVID-19 by using various search engines, in particular PubMed, Google Scholar, Science Direct, etc. To achieve our goal, we advanced Deep Learning (DL) techniques to classify unclassified biomedical abstracts. In the context of scientific research, DL, part of the large field of artificial intelligence, allows machines to learn and perform decisions in an automatic way. Given the heterogeneity of the data available in the scientific field, the search for information seems to be a difficult task and can generate difficulties for researchers. For our work, we were interested in the literature review component in the field of epidemiological research, which represents a challenging task. Particular attention was given to this major challenge. To overcome it, DL techniques and algorithms were investigated and adapted to our context of text classification related to fighting COVID-19. In this context, we exploited algorithms to facilitate the learning process based on sourced data in order to better manipulate target data. The main approach is called transfer learning, which enables information and knowledge from a past task in order to improve the next task [3,4]. Transfer learning is based on a recent approach called "Universal Embeddings", which is essentially pre-trained embeddings obtained from the training models of DL on a large corpus. In fact, "Universal Embeddings" enables using the pre-trained embeddings in various NLP tasks, including scenarios with constraints such as heterogeneity of unlabeled data.

The aim of this article is twofold:

1. First, to produce an appropriate dataset for training. Our research is oriented to multiclass classification rather than to multilabel classification. When we began our research in January 2020, we could not find any datasets on COVID-19 with short text that is classified according to categories. For this reason, we were required to build a new dataset on COVID-19 by using the PubMed search engine, and it was validated by experts from Community and Preventive Medicine at the Faculty of Medicine, Department of Epidemiology. The validation process took over 6 months. We advance a method to build the required dataset and address our needs.

The proposed dataset, entitled Cov-Dat-20, contains 4304 papers distributed in an equitable manner according to four categories, namely: COVID-19, Virology, Public Health and Mental Health.

2. Based on the literature, we considered concerns and issues that can now be addressed through natural language processing (NLP) based on different pre-trained language models including Glove [5], ELMo [6], OpenAI GPT model [7] and Bidirectional Encoder Representations from Transformers (abbreviated as BERT) [8]. Compared to the cited models, BERT provides better results for many use cases and without necessarily requiring a large amount of labelled data thanks to a "pre-training" phase without labels, allowing it to acquire a more detailed knowledge of the language. In addition, the BERT model uses a specific manner to handle several limits such as the reduced size of input text and the lack of vocabulary as was our case when we deal with the summary of scientific articles. Bearing in mind the several benefits of BERT, we propose the CovBERT model to help medical research epidemiologists fight against COVID-19. We have released the CovBERT model as a new resource to enhance the performance of NLP tasks in the scientific domain. Our proposed model is a pre-trained language model based on BERT and trained on our large corpus of scientific text, Cov-Dat-20. After training, we fine-tuned the CovBERT model with a specific subject related to COVID-19. Finally, we evaluated the CovBERT model on different numerously studied text classification datasets.

### 1.2. Traditional ML versus DL Models

Since 2005, the outlook of artificial intelligence has changed dramatically with machine learning (ML) and the emergence of DL, which draws inspiration from neuroscience. In fact, traditional learning, or ML, as part of artificial intelligence is using techniques (such as DL) which allow machines to learn from their experiences in order to improve the way they perform their tasks. In traditional learning, the learning process is based on several steps:

- Feed an algorithm with data;
- Use this data to train a model;
- Test and deploy the model;
- Use the deployed model to perform an automated predictive task.

DL is one of the main ML and artificial intelligence technologies that are based on neural networks. The learning process is qualified as deep because the structure of artificial neural networks consists of several input, output and hidden layers. Each layer contains units that turn the input data into information that the next layer can use for a specific predictive task. Based on this structure, a machine can learn through its own data processing.

The remainder of the paper is organized in four sections in addition to the introduction. The present introduction includes a first subsection illustrating the context of our work. Traditional versus DL models are then discussed in a separate subsection. Section 2 is devoted to presenting related works. In Section 3, we will provide our methodology. Section 4 discusses the presented experimental results. In Section 5, we will have our concluding remarks and include our strategies for future research.

## 2. Related Works

This section is devoted to past and current research in biomedical text classification [9,10]. Text classification is a fundamental task of natural language processing (NLP), with the aim of assigning a text to one or more categories. The applications of text classification include sentiment analysis [11], question classification [12] and topic classification [13], the latter of which we are interested in for this work. We investigated several proposed approaches related to text-classification tasks based on deep neural networks and pre-trained language models.

Today, deep neural networks have several techniques and models that prove/demonstrate new state-of-the-art results on fully examined text-classification datasets. There are some models, such as convolutional neural networks (CNN) [14], recurrent neural networks (RNN) [15] and artificial neural networks (ANN) [16], as well as some more complex networks such as C-LSTM [17], CNN-LSTM [18] and DRNN [19].

Nevertheless, DL models mention additive advantages over traditional ML models based on the backpropagation algorithm. In fact, related to [20], the backpropagation algorithm is an optimization algorithm that adjusts the parameters of a network of multilayer neurons to match inputs and outputs referenced in a learning dataset. According to reference [21], the usage of DL for text classification requires entering the text into a deep network to obtain the text representation then entering it into the Softmax() function and obtaining the probability of each category. Yao et al. [22] proposed an improvement of distributed document representations by adding descriptions of medical concepts for the classification task of the clinical files for the benefit of traditional Chinese medicine. The active learning technique [23] was applied in the clinical domain, which exploits untagged corpora to enhance the clinical text-classification process. An ordinary approach is to first map the narrative text to concepts of different knowledge sources such as the Unified Medical Language System (UMLS), then train the classifiers on document representations that include the unique concept identifiers of the UMLS—Concept Unique Identifiers (CUIs)—as functionalities [24]. In [25], the authors are interested in the acute kidney injury (AKI) prediction based on DL models. They used knowledge-guided CNNs to merge word features with UMLS CUI features. They used pre-trained word embeddings and CUI embeddings of clinical notes as the input.

Nevertheless, the neural networks proved their effectiveness until the advent of the transfer learning approach. In 2018, that main approach appeared and proved its effectiveness. It consists of training a complete model to perform a task with many data. Then, the pre-trained model can be used to complete other tasks, building on previous learning. This is called transfer learning.

By definition, a pre-trained model is a recorded network that has already been trained on a large dataset. Generally, we use them on large-scale text-classification tasks. A pre-trained model is ready to be used as is, or, based on the transfer learning technique, the model can become personalized for a given task. The intuition behind transfer learning for text classification is that if a model is trained on a sufficiently large and general dataset, that model will effectively serve as a generic model. The model is efficient at taking advantage of the data learned without the necessity of starting from scratch by training a large model on a large dataset.

In general, NLP projects rely on pre-trained word embedding on large volumes of unlabeled data by means of algorithms such as word2vec [26] and GloVe [5]. They aim at initializing the first layer of a neural network. Then, the obtained model is trained on specific data for a particular task. That said, many current models for supervised NLP tasks are pre-trained as models in language modeling (which is an unsupervised task) and then fine-tuned (which is a supervised task) with tagged task-specific data. Recent advances in ULMFiT [27], ELMo [6], OpenAI Transformer [28] and BERT [8] present a quintessential shift, in paradigmatic terms, from the simple initialization of the first layer of models to pre-training the entire model with hierarchical representations to improve the natural language processing process, including text classification. All these approaches enable pre-training an unsupervised language model on a large dataset such as Wikipedia and then fine-tuning these pre-trained models on specific tasks.

Instead of associating a static embedding vector with each word, pre-trained models build richer representations that consider the semantic and syntactic context of each word.

Related to [27], ULMFiT (Universal Language Model Fine-Tuning) is a recent generic method used to build efficient text-classification systems, setting a new state of the art on several benchmarks in NLP tasks. The present method has proven its efficiency in terms of not requiring a huge amount of data to train the model.

In addition, ELMo (Embeddings from Language Models) examines the entire sentence before assigning and embedding each word it contains instead of using a fixed embedding for each word [6]. It uses a bidirectional LSTM trained on a particular task to be able to create these embeddings. ELMo can be trained on a massive dataset. ELMo is trained to reveal the next word in each sentence. This is convenient because of the large amount of textual data.

In [7], the authors present the OpenAI GPT, short for Generative Pre-Training Transformer, which is a multi-layered unidirectional transformer decoder. The proposed model was trained on a huge corpus and aims to perform various NLP tasks based on precise adjustments. To start, the transformer language model was trained in an unsupervised manner. The training process is based on a few thousand books from the Google Books corpus. From there, the pre-trained model will be adapted to the supervised target task.

In the same context, recently, the BERT model has achieved state-of-the-art results in a broad range of NLP tasks [8]. It is a variation of transfer learning. The main operating mode of BERT corresponds to a transfer by fine-tuning that is like the one used by ULMFiT. Furthermore, BERT can also be used in the transfer mode by extracting features like ELMo. The BERT model uses transformer architecture, which is a recent and powerful alternative to RNNs to achieve deep bidirectional pre-training. In addition, the use of two new tasks for pre-training, one at the word level and the other at the sentence level, defines the main innovation of BERT.

In our work, we revolve around the classification of scientific text in the biomedical field, and we intersect with the summary of the scientific article. The main challenge encountered is a reduced size of text and the lack of vocabulary. Based on BERT's advantages, the main model is instantiated in different general and specialized domains (biomedical and scientific domains, for example). In fact, BERT's contextualization introduces different advantages related to the performance of the model, the learning time, the learning cost, the quantity of data requested for the learning, the length of the input text, etc. In our context, we looked at BERT-base models in three different fields: multiple domains (the general case), the scientific domain (the scientific articles) and the biomedical domain (the COVID-19 case study) presented in Table 1.

**Table 1.** A comparative study of pre-trained Bert-base models.

| Models | Summary | NLP Tasks | Datasets | Learning Type | Mono/Multi-Class | Accuracy Model/Bert | Domain |
|---|---|---|---|---|---|---|---|
| BoostingBERT | The model integrates multi-class boosting into the BERT model. The boosting technique is demonstrated to be able to be used to enhance the performance of BERT, instead of other techniques such as bagging or stacking. Based on the experimental results, BoostingBERT outperforms the bagging BERT constantly. Two approaches are compared, making use of the base transformer classifier in the BoostingBERT model: weights privacy vs. weights sharing, and the former one constantly outperforms the latter one. | Multiple NLP tasks | GLUE dataset | Ensemble learning | Multi-class | 82.93%/80.72% with CoLa dataset<br>93.35/92.55 with SST-2 dataset | Multi-domain |
| EduBERT | The use of pre-trained models has proven a great advance in learning analytics. They apply the BERT approach to the three LAK tasks previously explored on the MOOC forum data: detection of confusion, urgent intervention by teachers and classification of sentimentality. The experimental results have proven an improvement in performance beyond the state of the art. | Sentiment analysis (SA), named entity recognition (NER) and question answering (QA) | Stanford MOOCPosts dataset | Supervised/unsupervised | Not mentioned | 89.78%/89.47% | Multi-domain |
| ALBERT | ALBERT introduces two optimizations to reduce model size: a factorization of the embedding layer and parameter sharing across the hidden layers of the network. The result of combining these two approaches results in a baseline model with only 12M parameters, compared to BERT's 108M, with an accuracy of 80.1% on several NLP benchmarks compared with BERT's 82.3% average. | Question answering | GLUE<br>SQuAD<br>RACE | Supervised/unsupervised | Not mentioned | 88.7%/85.2% | Multi-domain |
| FinBERT | FinBERT is a language model based on BERT for financial NLP tasks. FinBERT is evaluated on two financial sentiment analysis datasets. The authors achieve the state of the art on FiQA sentiment scoring and Financial PhraseBank. They implement two other pre-trained language models, ULMFit and ELMo, for financial sentiment analysis to compare with FinBERT. Experiments are conducted to investigate the effects of further pre-training on the financial corpus, training strategies to prevent catastrophic forgetting and fine-tuning only a small subset of model layers for decreasing training time without a significant drop in performance. | Sentiment analysis and text classification | Financial sentiment analysis datasets | Supervised/unsupervised | Not Mentioned | 86% | Scientific domain |
| SciBERT | The authors release SCIBERT, a pre-trained language model based on BERT to address the scientific data. SCIBERT leverages unsupervised pre-training on a large multi-domain corpus of scientific publications to improve performance on downstream scientific NLP tasks. SCIBERT largely outperforms BERT and previous state-of-the-art models in a variety of biomedical text-mining tasks including sequence tagging, sentence classification and dependency parsing, with datasets from a variety of scientific domains. SCIBERT makes improvements over BERT. | Named entity recognition (NER), PICO extraction (PICO), text classification (CLS), relation classification (REL), dependency parsing (DEP) | Corpus of scientific text | Supervised/unsupervised | Not mentioned | 99.01%/88.85% | Scientific domain |

**Table 1.** *Cont.*

| Models | Summary | NLP Tasks | Datasets | Learning Type | Mono/Multi-Class | Accuracy Model/Bert | Domain |
|---|---|---|---|---|---|---|---|
| KnowBert | KnowBert represents a general method to embed multiple knowledge bases (KBs) into large-scale models. The proposed model aims to enhance scientific data representations with structured, human-curated knowledge. For each KB, the retrieve of the relevant entity is based on an integrated entity linker; then, the contextual word representations are updated via a form of word-to-entity attention. After integrating WordNet and a subset of Wikipedia into BERT, KnowBert demonstrates improved perplexity, ability to recall facts as measured in a probing task and downstream performance on relationship extraction, entity typing and word sense disambiguation. KnowBert's runtime is comparable to BERT's, and it scales to large KBs. | Relation extraction, entity typing, word sense disambiguation | Wikipedia | Supervised/unsupervised | Not mentioned | 89.01%/89% | Scientific domain |
| ClinicalBert | The authors are exploring and releasing BERT models for clinical text: one for generic clinical text and another for discharge summaries specifically. The main approach demonstrates that using a domain-specific model enhances performance improvements on three common clinical NLP tasks compared with nonspecific embeddings. These domain-specific models are not as performant on two clinical de-identification tasks, and the authors argue that this is a natural consequence of the differences between de-identified source text and synthetically non-de-identified task text. | Readmission prediction, diagnosis predictions, mortality risk estimation | MIMIC-III dataset | Supervised/unsupervised | Not mentioned | 80.8%/77.6% | Biomedical domain |
| BlueBERT | The proposed approach is a BERT-base model pre-trained on PubMed abstracts and MIMIC-III clinical notes. Based on the BERT model, BlueBERT is specialized by an extra linear layer on top of the existing model to transform the output into 10 classes, one for each ICD-9 code. The authors implemented the BCEWithLogits loss for multi-class classification. Related to the BERT model's architecture, BlueBERT introduces three small architectural variations: (1) adding three linear layers with ReLU non-linearity instead of just one linear layer, (2) freezing the BlueBERT weights from the first variant so that only the linear layer weights would be tuned and (3) adding a dropout layer after the BERT layer from the second variant. | Text classification | PubMed abstracts and MIMIC-III datasets | Supervised/unsupervised | Multi-class | 89.2%/86.9% | Biomedical domain |
| BioBERT | The BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) model is a domain-specific language representation model pre-trained on large-scale biomedical corpora. Based on experimental results, BioBERT largely outperforms BERT and previous state-of-the-art models in a variety of biomedical text-mining tasks when pre-trained on biomedical corpora. BioBERT outperforms on the following three representative biomedical text-mining tasks: biomedical named entity recognition, biomedical relation extraction and biomedical question answering. The analysis results show that pre-training BERT on biomedical corpora helps it to understand complex biomedical texts. | NER<br>Biomedical relation extraction<br>Bio question answering | PubMed abstracts/4,5B<br>PMC full papers/13,5B | Supervised/ unsupervised | Not mentioned | 89.04% /88.30% | Biomedical Domain |

## 3. Methodology

We are interested in offering relevant information when searching available media (search engines, datasets). In this context, text classification [29–31], defined as the process of associating a category with a text of various length based on the information it contains, is an important element of information-retrieval systems. We face text-classification challenges and accuracy issues. For each new entry, the main challenge consists of determining the category to which this entry belongs. The text annotation process is time-consuming and is generally performed manually because of the language complexity. Therefore, the automation of this process has become a priority for the scientific community to be efficient. For our work, we aim to collaborate with epidemiologists to help medical research fight COVID-19. Our main goal is to not only classify scientific texts but to predict unseen data based on the pre-trained model.

### *3.1. COV-Dat-20 Dataset Creation*

We propose a dataset containing data on COVID-19 extracted from summaries of PubMed. It is made up of 4304 articles distributed in an equitable manner according to four categories, namely: COVID-19, Virology, Public Health and Mental Health.

For the methodology, amongst the articles that were proposed by PubMed, some could have been classified under more than one category. For our work, our research is oriented to the multiclass classification type rather than to multilabel classification. In this case, our experts in the Department of Epidemiology recommended that we classify the articles based on the highest percentage proposed by PubMed.

The Cov-Dat-20 dataset is elaborated based on advanced SQL queries from PubMed. For example, we present an advanced SQL query in PubMed: (((“public health” [MeSH Major Topic]) OR “preventive medicine” [MeSH Major Topic])) AND “COVID-19” [Supplementary Concept]; (health knowledge, attitudes, practice [MeSH Terms]) AND COVID-19 [Supplementary Concept]; (((COVID-19 [Title/Abstract]) OR COVID-19 [Supplementary Concept])) AND (((epidemiological assessment [Title/Abstract]) OR public health interventions [Title/Abstract]) OR epidemiology [Title/Abstract]).

Figure 1 presents the scientific paper distribution in the dataset according to the categories mentioned above.

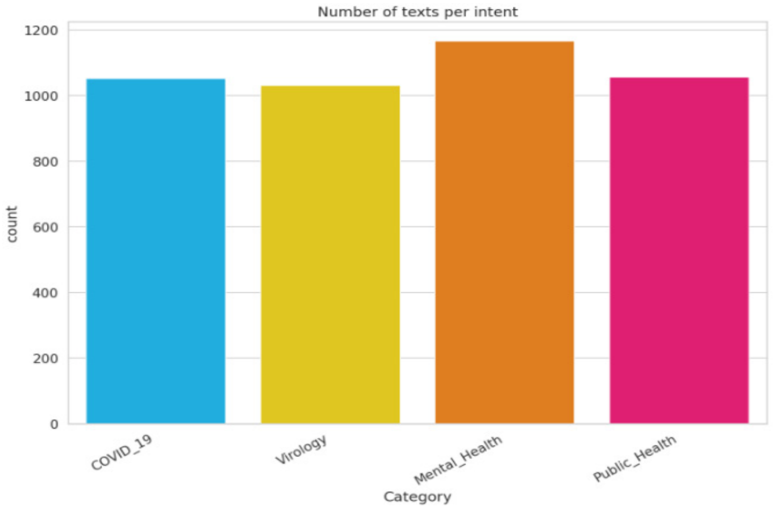


**Figure 1.** Data repartition in the Cov-Dat-20 dataset.

For more details, 4304 rows and 3 columns, 1,013,901 words, 43,123 unique words and 4304 sentences characterize the Cov-Dat-20 dataset.

Table 2 presents the label encoding related to our dataset. In our work, we were interested in the multi-class concept in the classification process. For each category, we offer a selective and detailed list of keywords.

**Table 2.** Label encoding.

| Category | Description |
|---|---|
| COVID-19 | It is interested in scientific papers of probable treatment and the different symptoms related to COVID-19. |
| Virology | It deals with scientific papers about the study of viruses, genome sequencing, etc. |
| Public Health | It focuses on scientific papers about the study, prevention, control, in particular through vaccination, and epidemiological data against COVID-19. |
| Mental Health | It spotlights several scientific papers about the impact of COVID-19 on mental health. |

- COVID-19: focus on the (1) COVID-19 treatment and (2) COVID-19 symptoms: (1) treatment, chloroquine, hydroxychloroquine, interferon beta-1a, remdesivir, lopinavir/ ritonavir; (2) fever, headache, cough, chills, shortness of breath or difficulty breathing, muscle pain, repeated shaking with chills, new loss of taste or smell;
- Virology: genome sequencing, phylogenetic analysis, SARS-CoV-2, MERS-CoV, nomenclature, virus composition, virus layers;
- Public Health: COVID-19, interventions, awareness, behavior, behavioral change, coronavirus, pandemic, public health protection, public health measures;
- Mental Health: COVID-19, mental health disorders, SARS-CoV-2, neural taxonomies, personalized medicine, precision psychiatry, social connection, mental health, psychiatry.

The Cov-Dat-20 is available on https://www.kaggle.com/mayarakh/Cov-Dat-20 accessed on 6 March 2022.

### *3.2. Data Pre-Processing*

Raw data needs to be transformed into an understandable format. To do this, we opted for applying data pre-processing techniques to build a DL classifier. In fact, data pre-processing techniques eliminate characteristics of less important data and improve accuracy.

For our case, we used a common preprocessing approach that can integrate with various NLP (natural language processing) tasks using NLTK (Natural Language Toolkit) [32]. Before tackling the learning process, the text in the dataset goes through some stages, namely, the elimination of punctuation, putting all the text in lower case, tokenization, cleaning and lemmatization.

- Lowercasing is a widespread approach to reduce all the text to lower case for simplicity.
- Tokenization: text pre-processing step, which involves splitting the text into tokens (words, sentences, etc.)
- Cleaning is a form of pre-processing to filter out useless data such as punctuation removal and stop-word removal (a stop word is a commonly used word in text and stop-word removal is a form pre-processing to filter out useless data).
- Lemmatization is an alternative approach to stemming for removing inflection.

At the end of this step, data are ready to move to the step of decomposing the dataset into a part for "Learning" and a part reserved for the "Test" phase. We chose to carry out this decomposition by reserving 80% for the training set and 20% for testing.

### *3.3. Exploration of the BERT Model*

As mentioned before, we focused on the scientific-text-classification task. In the same context, Google's BERT [8], having received deep bidirectional training using the transformer, gave state-of-the-art results for many NLP tasks, more precisely, in the text-classification task. In addition, our decision to explore the BERT model is justified by its several advantages compared to similar models. BERT aims at improving the understanding of users' requests in order to provide more relevant results, especially for requests formulated in a natural way.

### 3.3.1. BERT-Base Characteristics

BERT is a neural network that can treat a wide variety of NLP (natural language processing) tasks [8]. To do so, the learning phase is broken down into two phases. First, we proceed with the pre-training phase, which is very time and computation consuming. Once this phase is performed, a network is created that has a certain general understanding of the language. Then, the second phase is called the adjustment phase, which trains the network on a specific task. Moreover, BERT uses a part of the transformer network architecture. The advantage of this architecture is that it treats the relationships between distant words better than recurrent networks (LSTM/GRU) [33]. On the other hand, the network cannot process sequences of any length but has a finite input dimension to learn in a reasonable time. At the scale of this work, we use the basic model of BERT-base with fixed characteristics: 12-layer, 768-hidden, 12-heads, 110 M parameters.

### 3.3.2. BERT-Base Operation

The Bert model is a bidirectional model. Unlike its predecessors, which were unidirectional and so read the text in a particular direction (e.g., left to right), the main model of BERT goes through the entire text in both directions simultaneously, which presents the property of "bidirectionality". Technically speaking, BERT consists of multiple layers forming a "Transformer", which learns contextual relationships between the different words composing the text. The transformers aim at analyzing the words of a complex query to relate them in order to comprehend the semantics of the sentence and to better understand its overall meaning. TPU Clouds are integrated circuits that accelerate the workload of transformers to make them faster and more efficient.

The BERT architecture in our proposal is illustrated in Figure 2. We took the case of the category Virology and an input text composed of two sentences: "The COVID-19 genome is decrypted. The virus composition is . . . ". The algorithm will go in both directions, from sentence 1 to sentence 2 until the end of the text (Abstract, full article, . . . ) but also from sentence 2 to sentence 1 until the beginning of the text as depicted in Figure 2.

The core of the architecture is mainly decomposed into two components. It uses an encoder to read the input text and thus generates a vector representation of the words. In addition, BERT uses a specific decoder to perform the expected prediction task.

BERT-base offers a vocabulary of 30,522 words. The vocabulary is built in a way that is based on the tokenization process. Indeed, the main process consists in dividing the input text into a list of words, called tokens, that are available in the constructed vocabulary. To process words that are out-vocabulary, BERT-base uses a technique, called BPE-based WordPiece tokenization. Regarding non-vocabulary words, this approach proposes to divide them into sub-words. Each word is then represented by a group of sub-words. For each sub-word, BERT-base provides contextual representations. Therefore, the context of the word is merely the combination of the sub-words' contexts. In this work, we adopted the BERT idea to epidemiology by optimizing the word keys and subkeys.

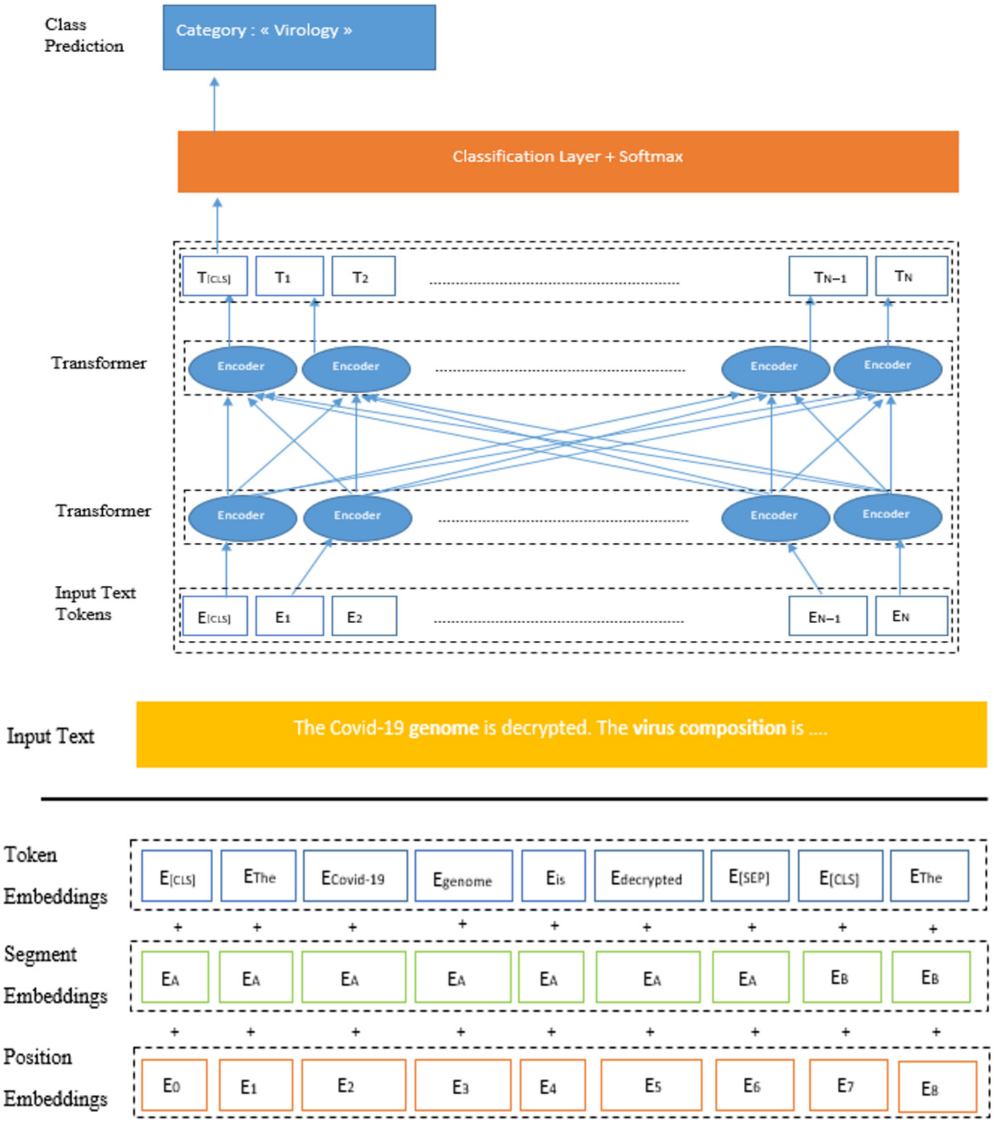


**Figure 2.** BERT-base fine-tuning model architecture: COVID-19 case study.

### 3.3.3. Contextual Embeddings in Biomedical and Scientific Domains

Considering the literature, the word-embedding technique shows its effectiveness in traditional word-level vector representations [25] GloVe [5] and fastText [34]. However, this technique faces some limitations by expressing all possible meanings of a word as a single vector representation. In addition, it cannot disambiguate the word senses based on the surrounding context. To overcome these limitations, ELMo and BERT present efficient solutions by providing contextualized word representations. For example, ELMo creates a context-sensitive embedding for each word in each sentence by pre-training it on a large text corpus as a language model. Compared to ELMo, BERT goes deeper and involves much more parameters for contextualized word representations. It can be fine-tuned to accomplish a specific task in several domains such as the biomedical domain. In this context, we present BioBERT and ClinicalBERT models. In [35], BioBERT is a BERT-base model finetuned over a corpus of biomedical research articles from PubMed. BioBERT focused on several NLP tasks presented in Table 2. In the same context, we present the ClinicalBERT model [36]. The main model is based on BERT and then pre-trained on clinical notes from the MIMICIII dataset. In addition, several works utilize the BERT-base model and then perform fine-tuning in scientific domains, such as SciBERT [37]. SciBERT is concerned with named entity recognition, relation extraction and text classification as pointed out in Table 2. Furthermore, we used the BERT model pre-trained in the English language to classify scientific papers from PubMed, and we aim to fine-tune it in the field of COVID-19 for biomedical- and scientific-text-classification tasks.

In Table 3, we produce a comparative study of BERT and its variants in terms of NLP tasks. The dataset and the main size, the special domain, the several hyper parameters, the length period and the different methods used are presented in Table 3.

**Table 3.** Comparative study of BERT variations.

| Model | NLP Tasks | Dataset /Size | Characteristics | Hyperparameters | Learning Period | Methods |
|---|---|---|---|---|---|---|
| BioBERT [35] | NER<br>Biomedical relation extraction<br>Bio question answering | PubMed Abstracts/4,5B<br>PMC Full Papers/13,5B | Biomedical domain | Sentence length: 128–512 tokens | 23 days<br>8 NVIDIA V100 (32GB) GPUs | Word piece tokenization<br>Pre-training BERT on biomedical corpora: Naver Smart ML<br>Fine-tuning BioBERT |
| ClinicalBERT [36] | Readmission prediction<br>Diagnosis predictions<br>Mortality risk estimation | MIMIC-III | Clinical domain | Sequence length: 128–512 tokens | Amazon Web Services using a single K80 GPU | Subword embeddings<br>Self-attention mechanism |
| SciBERT [37] | NER<br>Text classification<br>Relation classification<br>Dependency parsing | Semantic Scolar/1.14M | Scientific domain | Sentence length: 128–512 tokens | 5 days + 2 days<br>TPU v3 with 8 cores | Finetuning BERT: Frozen BERT embeddings<br>Contextualize word embeddings |
| KnowBERT [38] | Relation extraction<br>Entity typing<br>Word sense disambiguation | Wikipedia | Knowledge domain | - | - | Mention-span representations<br>Retrieval of relevant entity embeddings<br>Recontextualization of entity span embeddings |

### 3.3.4. Self-Attention Mechanism

This section is devoted to detailing the self-attention mechanism. In fact, the functioning of our cerebral cortex freely inspires the attention mechanism. For example, when analyzing an image to describe it, our attention is instinctively focused on a few areas containing important information without looking at every part of the image equally. This mechanism, therefore, resembles a means of saving processing resources in the face of complex data for analysis. Similarly, when an interpreter translates a text from a source language into a target language, it focuses, based on several experiences, on which words in a source sentence are associated with a certain term in the translated sentence. This attention mechanism is now an integral part of most modern semantic analysis solutions [39]. The attention mechanism is formulated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where the parameters $Q$, $K$ and $V$ stand for three vectors, which are query, key and value, generated through input embedding, and $d_k$ designates the size of key vectors. $K^T$ stands for transposed vector of $K$. For example, let us consider that we have four queries, which means:

$$Q = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{bmatrix}\text{, and three keys } (d_k = 3)\text{, that is } = \begin{bmatrix} k_1 \\ k_2 \\ k_3 \end{bmatrix}\text{, then}$$

$$\frac{QK^T}{\sqrt{d_k}} = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{bmatrix} \cdot \begin{bmatrix} k_1 & k_2 & k_3 \end{bmatrix} = \frac{1}{\sqrt{d_k}} \begin{bmatrix} q_1{\cdot}k_1 & q_1{\cdot}k_2 & q_1{\cdot}k_3 \\ q_2{\cdot}k_1 & q_2{\cdot}k_2 & q_2{\cdot}k_3 \\ q_3{\cdot}k_1 & q_3{\cdot}k_2 & q_3{\cdot}k_3 \\ q_4{\cdot}k_1 & q_4{\cdot}k_2 & q_4{\cdot}k_3 \end{bmatrix} \tag{2}$$

where, for example, $q_3{\cdot}k_2$ means the second key, $k_2$, applied on the third query, $q_3$. Then, the maximum of the matrix is calculated according to the function Softmax(), also called soft argmax or normalized exponential function.

## 4. Pre-Training BERT-Base on COV-Dat-20

To demonstrate the pertinence of Cov-Dat-20 for language model pre-training, we trained BERT-base on 4304 abstracts on several topics such as COVID-19 treatment, COVID-19 symptoms, virology, public health and mental health. Therefore, CovBERT is a BERT-base model trained on multiple domains of scientific abstracts. In fact, we chose to focus on abstracts only from the complete scientific papers. Our choice is justified by a comparative study [40] between the experimental results of a scientific-text-classification approach based on (1) the full article and (2) the abstract only. Based on the experimental results, we observed that the abstract classification approach is more efficient than the full article approach in terms of learning time, the model size and complexity.

Furthermore, the tokenization step is essential in the BERT fine-tuning phase. To feed our text into BERT, we divided it into tokens, and then these tokens were mapped to their index in the tokenizer vocabulary. Related to BERT-base model, the maximum sentence length is 512 tokens. Applying pre-trained BERT requires us to use the tokenizer provided by the model. In fact, the BERT-base generates a specific vocabulary of a fixed size. Added to that, BERT's tokenizer uses a specific manner to handle out-of-vocabulary words.

### *4.1. Importing Libraires*

In order to adjust the BERT-base model to our needs, we imported several necessary libraries related to the text-classification task, such as tensorflow, pandas, numpy, transformers, etc.

### *4.2. Needed Parameters for Training*

In order to obtain a high performance of our model, we followed the pre-training hyper-parameters used in BERT [8]. For fine-tuning, most hyper-parameters are the same as pre-training, except for batch size, learning rate and number of training epochs.

- Max Length: 64;
- Batch size: 32;
- Learning rate (Adam): 2e-5;
- Number of epochs: 4;
- Seed val.: 42.

### *4.3. Model Characteristics*

In order to be suitable to our classification task, we modified the pre-trained BERT model and we trained it on our dataset. The CovBERT model has several layers and output types designed to accommodate our specific NLP task.

### 4.4. Evaluation Metrics

In the main subsection, we present several indicators measuring the quality of the model. To measure the performance of this classifier, we introduce four types of elements classified for the desired class, namely, *TP*, *FP*, *TN* and *FN*:

- *TP*: the positive class correctly predicted by the models;
- *FP*: the positive class incorrectly predicted by the models;
- *TN*: the false class correctly predicted by the models;
- *FN*: the false class incorrectly predicted by the models.

In what follows, we present the evaluation metrics adopted to measure the performance of the different DL models used. Indeed, our assessment is based on four different measures, including: *Accuracy*, *Precision*, *Recall*, F1-*Score*. The evaluation metrics are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

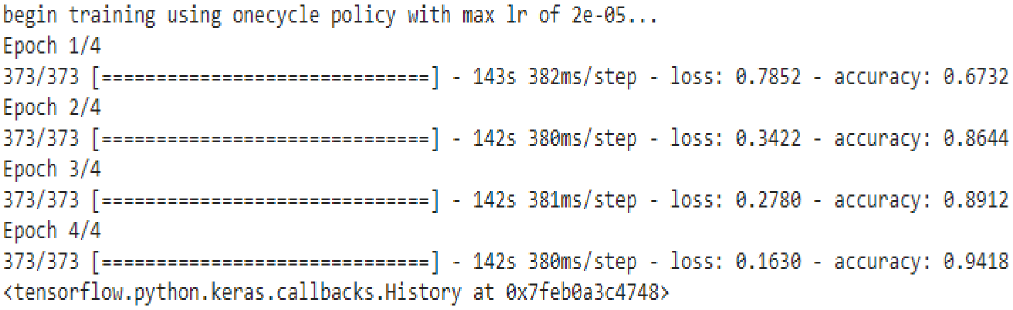
$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{6}$$

We tackled the text-classification task based on the BERT-base model after exploring DL and experimenting with its limitations. Let us begin by describing our experiments. We started with the automatic collection of scientific papers related to COVID-19 from PubMed. We then created our dataset made up of 4304 scientific papers distributed in an equitable way on four different categories: COVID-19, Virology, Public Health and Mental Health. In order to validate the data classification process in our dataset, we contacted epidemiologists to carry out this task. Based on the manual verification, we concluded that some papers were misclassified. Based on our previous experiments and considering some related works, we opted for performing the scientific paper classification with the PubMed engine. We converged towards a DL solution to tackle the pandemic of COVID-19. We named our model CovBERT.

To adapt the existing pre-trained BERT model to our needs, we applied some modifications, and we trained it on our dataset Cov-Dat-20. We then explored the modified BERT-base model with the graphics processing unit (GPU) to obtain a better performance in terms of learning cost. In addition, we fixed the training hyper-parameters such as the max. length, the batch size, the learning rate, the epoch's number, and the seed val. We monitored the validation loss and kept the best model on the validation set.

Figure 3 presents the CovBERT accuracy with four epochs. We notice, with CovBERT, an accuracy of 94% at epoch four. Beyond four epochs, we risk falling into overfitting.



```
begin training using onecycle policy with max lr of 2e-05...
Epoch 1/4
373/373 [==============================] - 143s 382ms/step - loss: 0.7852 - accuracy: 0.6732
Epoch 2/4
373/373 [==============================] - 142s 380ms/step - loss: 0.3422 - accuracy: 0.8644
Epoch 3/4
373/373 [==============================] - 142s 381ms/step - loss: 0.2780 - accuracy: 0.8912
Epoch 4/4
373/373 [==============================] - 142s 380ms/step - loss: 0.1630 - accuracy: 0.9418
<tensorflow.python.keras.callbacks.History at 0x7feb0a3c4748>
```

**Figure 3.** CovBERT model accuracy.

The related confusion matrix of the proposed model is presented in Figure 4. We noticed that the high precision was maintained for the Public Health category with 94%. It was also maintained for the Mental Health and the Virology categories with 79% being the least precision value.

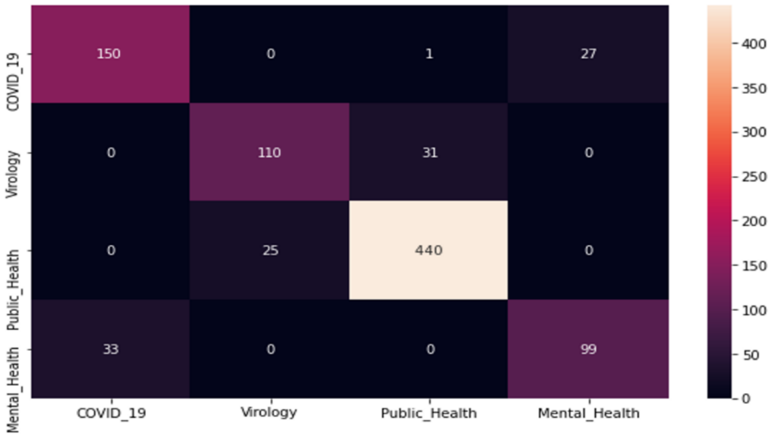


**Figure 4.** Model confusion matrix of our proposed CovBERT model.

The error related to the training set (training loss) of the proposed dataset is shown in Figure 5. The training loss value starts with 0.8. Then, it gradually falls until 0.2 in the third epoch. Figure 6 shows the accuracy evolution over time. We noticed that the accuracy increases over time from 84% in epoch 1 to 94% in epoch 4. We concluded that, from one epoch to another, the model acquired more knowledge and proved its effectiveness.

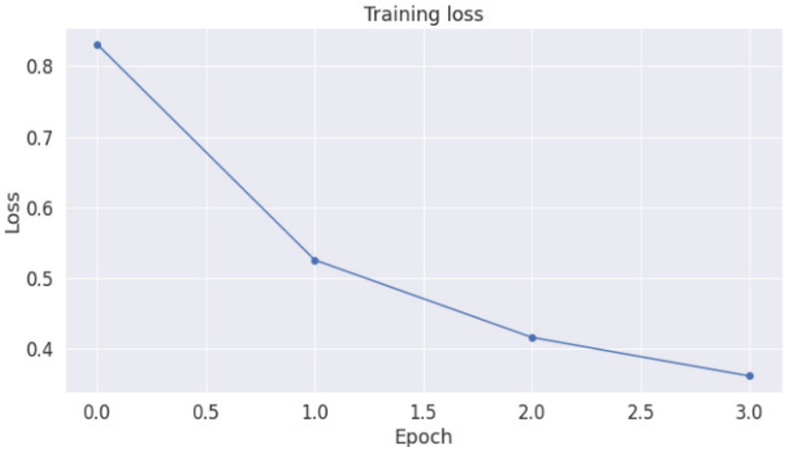


**Figure 5.** Training loss of our proposed CovBERT model.

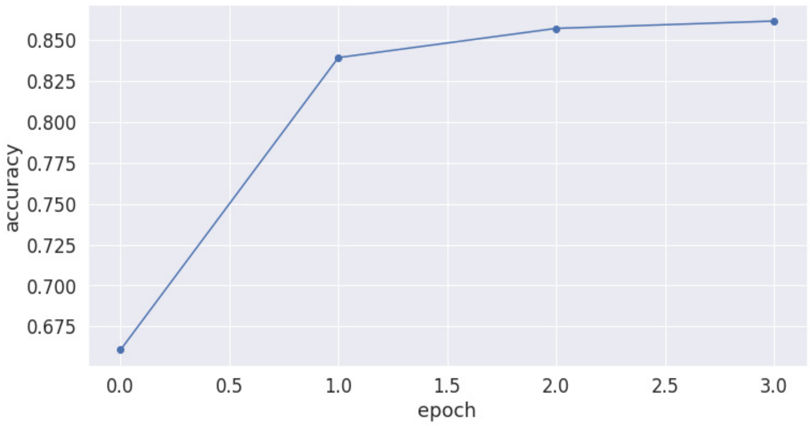


**Figure 6.** CovBERT model accuracy evolution over time with 4 epochs.

In summary, we proved the high performance of the modified BERT model. We noticed that this training model is better suited for the specific NLP task compared to other DL training models such as CNN and BiLSTM. Among the advantages of the modified BERT model, we noticed that it is easy to implement. The pre-trained modified BERT model weights already encode a large quantity of information on the English language. The fine-tuning step is based on a much smaller dataset for a specific task. Furthermore, we noticed that using the modified BERT model is more efficient in terms of cost and learning time, as well as in terms of model complexity and size.

## 5. Discussion

In this section, we present, in Table 4, BERT's variants, the relative domains and natural language processing (NLP) tasks. Then, we discuss the main results of our comparative study. Indeed, we used Huggingface's models in our approach resumed in Table 5.

**Table 4.** BERT's variants, relative domain and NLP tasks.

| Models | Domain | NLP Tasks |
|---|---|---|
| roberta-base | Multi-Domain | 1. Named Entity Recognition<br>2. Sequence Classification<br>3. Question Answering |
| albert-base-v1 | Multi-Domain | 1. Sequence Classification<br>2. Question Answering |
| allenai/scibert_scivocab_uncased | Scientific Domain | 1. Named Entity Recognition (NER)<br>2. PICO Extraction<br>3. Text Classification<br>4. Relation Classification (REL)<br>5. Dependency Parsing (DEP) |
| allenai/scibert_scivocab_cased | Scientific Domain | 1. Named Entity Recognition (NER)<br>2. PICO Extraction<br>3. Text Classification<br>4. Relation Classification (REL)<br>5. Dependency Parsing (DEP) |
| emilyalsentzer/Bio_ClinicalBERT | Biomedical Domain | 1 Biomedical Named Entity Recognition<br>2. Biomedical Relation Extraction<br>3. Biomedical Question Answering |
| dmis-lab/biobert-base-cased-v1.1 | Biomedical Domain | 1. Biomedical Named Entity Recognition<br>2. Biomedical Relation Extraction<br>3. Biomedical Question Answering |
| monologg/biobert_v1.1_pubmed | Biomedical Domain | 1. Biomedical Named Entity Recognition<br>2. Biomedical Relation Extraction<br>3. Biomedical Question Answering |
| dmis-lab/biobert-v1.1 | Biomedical Domain | 1. Biomedical Named Entity Recognition<br>2. Biomedical Relation Extraction<br>3. Biomedical Question Answering |
| gsarti/biobert-nli | Biomedical Domain | 1. Biomedical Named Entity Recognition<br>2. Biomedical Relation Extraction<br>3. Biomedical Question Answering |
| CovBERT | Biomedical and Scientific Domains | 1. Text Classification |

**Table 5.** Comparative study of BERT, its variants and our proposed CovBERT model.

| Models | Accuracy | Average Loss | Recall | Precision | F1 Metric |
|---|---|---|---|---|---|
| roberta-base | 83% | 29% | 51% | 68% | 57% |
| albert-base-v1 | 84% | 39% | 41% | 56% | 47% |
| allenai/scibert_scivocab_uncased | 84% | 33% | 70% | 71% | 69% |
| allenai/scibert_scivocab_cased | 84% | 30% | 74% | 74% | 73% |
| emilyalsentzer/Bio_ClinicalBERT | 87% | 25% | 83% | 82% | 82% |
| dmis-lab/biobert-base-cased-v1.1 | 87% | 14% | 68% | 68% | 66% |
| monologg/biobert_v1.1_pubmed | 87% | 17% | 77% | 79% | 76% |
| dmis-lab/biobert-v1.1 | 88% | 19% | 68% | 68% | 66% |
| gsarti/biobert-nli | 89% | 19% | 66% | 71% | 65% |
| **CovBERT** | **94%** | **18%** | **88%** | **86%** | **86%** |

Based on the Tables 4 and 5, we conclude that BERT-base models that are instantiated in the biomedical domain perform better in terms of accuracy than models that are instantiated in other domains.

Table 5 presents a comparative study of the initial BERT [8] and its variants such as SciBERT [37], BioBERT [35] and CovBERT. We noticed that the BERT model has proven its effectiveness in 11 NLP tasks. The BERT model is fine-tuned in several domains such as the scientific domain, SciBERT, for five NLP tasks, namely, NER, PICO extraction, text classification, relation classification and dependency parsing. In addition, BioBERT is another BERT variant for the biomedical domain. BioBERT focuses on the recognition of biomedical named entities, on biomedical relation extraction, as well as on biomedical question answering. Furthermore, we presented our proposed model, titled CovBERT, trained on PubMed abstracts in the scientific and biomedical domains.

Based on the comparative study, we concluded that BERT-base multi-domain models, such as ALBERT and Roberta, are less relevant than domain-specific models, with 84% and 83% accuracy, respectively. In addition, models pre-trained on biomedical domains are more accurate than models in the scientific domain with 89% accuracy. In addition, models in the biomedical field (specific field) are more relevant than pre-trained models in scientific domain (several fields). From there comes the effectiveness of our model, which concentrates the biomedical and scientific fields. We were able to show an accuracy improvement in the text-classification NLP task from 84% (SciBERT) in the scientific domain to 94% in the biomedical and scientific domains (CovBERT).

## 6. Conclusions

In this context of COVID-19, we advanced a new BERT-base pre-training model, referred to as CovBERT. The BERT model largely outperforms previous state-of-the-art models in a variety of NLP tasks. Furthermore, our choice is justified by the effectiveness of BERT to handle a lack of vocabulary, considering that, in our context, we deal with short texts (the summary of scientific articles). To assess the performance of our proposed model, we created a novel dataset, Cov-Dat-20, in the context of COVID-19, which contains several scientific papers collected from PubMed and classified into different categories according to COVID-19. Based on our experience, the CovBERT model outperforms the BERT-base model on text-classification tasks. The main approach is promising and presents an efficient increase based on the accuracy, precision, recall and F1 metrics. In future work, we intend to extend this study by enriching our dataset and developing our model in order to improve classification and prediction performance and to compare the new results to the present ones. Based on our promising results, we are inspired and aim to adapt our techniques to other subjects.

**Author Contributions:** Conceptualization: M.K. and M.B.A.; methodology, M.K., M.B.A., H.H., H.B. and M.J.; writing—original draft preparation, M.K.; writing—review and editing, M.K., M.B.A., H.H.,

H.B. and M.J.; supervision, M.J. and H.B. 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:** Not applicable.

**Acknowledgments:** The authors would like to thank Jihene Maatoug and Sihem Ben Fradj, members of the Epidemiology Department of CHU Farhat Hached, in Sousse, for their valuable cooperation and revision of our dataset. The authors also thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and New Brunswick Innovation Foundation (NBIF) for their financial support of the global project. These granting agencies did not contribute to the design of the study or the collection, analysis and interpretation of the data.

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

## References

1. Zu, Z.; Jiang, M.; Xu, P.; Chen, W.; Ni, Q.; Lu, G.; Zhang, L. Coronavirus disease 2019 (COVID-19): A perspective from china. *Radiology* **2020**, *296*, E15–E25. [CrossRef] [PubMed]
2. Worldometers for COVID-19. Available online: https://www.worldometers.info/ (accessed on 30 January 2020).
3. Pan, S.; Yang, Q. A survey on transfer learning. Knowledge and Data Engineering. *IEEE Trans.* **2010**, 1345–1359.
4. Yousaf, A.; Asif, R.M.; Shakir, M.; Rehman, A.U.; Alassery, F.; Hamam, H.; Cheikhrouhou, O. A Novel Machine Learning-Based Price Forecasting for Energy Management Systems. *Sustainability* **2021**, *13*, 12693. [CrossRef]
5. Pennington, J.; Socher, R.; Manning, C. Glove: Global vectors for word representation. In Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Volume 12, pp. 1532–1543.
6. Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; pp. 2227–2237.
7. Alt, C.; Hübner, M.; Hennig, L. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 1388–1398. [CrossRef]
8. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
9. Mirónczuk, M.; Protasiewicz, J. A recent overview of the state-of-the-art elements of text classification. *Expert Syst. Appl.* **2018**, *106*, 36–54. [CrossRef]
10. Holzinger, A.; Kieseberg, P.; Weippl, E.; Tjoa, A.M. Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI. In Proceedings of the International Cross-Domain Conference CD-MAKE 2018, Hamburg, Germany, 27–30 August 2018; pp. 1–8.
11. Maas, A.; Daly, R.; Pham, P.; Huang, D.; Ng, A.; Potts, C. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 142–150. [CrossRef]
12. Zhang, D. Question classification using support vector machines. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, ON, Canada, 28 July–1 August 2003; pp. 26–32. [CrossRef]
13. Wang, S.; Manning, C. Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, Jeju Island, Korea, 8–14 July 2012; Association for Computational Linguistics: Stroudsburg, PA, USA, 2012; pp. 90–94. [CrossRef]
14. Wang, X.; Yin, S.; Shafiq, M.; Laghari, A.A.; Karim, S.; Cheikhrouhou, O.; Alhakami, W.; Hamam, H. A New V-Net Convolutional Neural Network Based on Four-Dimensional Hyperchaotic System for Medical Image Encryption. *Secur. Commun. Netw.* **2022**, *2022*, 4260804. [CrossRef]
15. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modelling. IPS 2014 Workshop on Deep Learning. *arXiv* **2014**, arXiv:1412.3555.
16. Al-Shayea, Q. Artificial neural networks in medical diagnosis. *J. Appl. Biomed.* **2013**, *11*, 150–154. [CrossRef]
17. Shi, M.; Wang, K.; Li, C. A C-LSTM with Word Embedding Model for News Text Classification. In Proceedings of the IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), Beijing, China, 17–19 June 2019; pp. 253–257.
18. Xiao, Y.; Cho, K. Efficient character-level document classification by combining convolution and recurrent layers. *arXiv* **2016**, arXiv:1602.00367.

19. Wang, B. Disconnected recurrent neural networks for text categorization. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 2311–2320. [CrossRef]
20. Wang, X.; Lu, H.; Wei, X.; Wei, G.; Behbahani, S.S.; Iseley, T.T. Application of artificial neural network in tunnel engineering: A systematic review. *IEEE Access* **2020**, *8*, 119527–119543. [CrossRef]
21. Zheng, S.; Yang, M. A New Method of Improving BERT for Text Classification. In Proceedings of the Intelligence Science and Big Data Engineering. Big Data and Machine Learning, Nanjing, China, 17–20 October 2019; pp. 442–452. [CrossRef]
22. Yao, L.; Zhang, Y.; Wei, B.; Li, Z.; Huang, X. Traditional chinese medicine clinical records classification using knowledge-powered document embedding. In Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shenzhen, China, 15–18 December 2016. [CrossRef]
23. Figueroa, R.; Zeng-Treitler, Q.; Ngo, L.; Goryachev, S.; Wiechmann, E. Active learning for clinical text classification: Is it better than random sampling? *J. Am. Med. Inform. Assoc.: JAMIA* **2012**, *19*, 809–816. [CrossRef]
24. Garla, V.; Brandt, C. Knowledge-based biomedical word sense disambiguation: An evaluation and application to clinical document classification. *J. Am. Med. Inform. Assoc.* **2012**, *20*, 882–886. [CrossRef] [PubMed]
25. Yao, L.; Mao, C.; Luo, Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Med. Inform. Decis. Mak.* **2018**, *19*, 70–71. [CrossRef]
26. Asgari-Chenaghlu, M. Word Vector Representation, Word2vec, Glove, and Many More Explained. Ph.D. Thesis, University of Tabriz, Tabriz, Iran, 2017. [CrossRef]
27. Howard, J.; Ruder, S. Universal language model fine-tuning for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 328–339. [CrossRef]
28. Radford, A.; Karthik, N.; Tim, S.; Ilya, S. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 1 October 2021).
29. Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Inform. Process. Man.* **1988**, *24*, 513–523. [CrossRef]
30. Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of tricks for efficient text classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 3–7 April 2017; pp. 427–431. [CrossRef]
31. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
32. Bird, S. NLTK: The natural language toolkit. In Proceedings of the COLING/ACL on Interactive Presentation Sessions Association for Computational Linguistics 2006, Sydney, Australia, 17–18 July 2006; pp. 69–72.
33. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A review of recurrent neural networks: Lstm cells and network architectures. *Neural Comput.* **2019**, *31*, 1–36. [CrossRef] [PubMed]
34. Xu, J.; Du, Q. A deep investigation into fasttext. In Proceedings of the IEEE 21st International Conference on High Performance Computing and Communications, Zhangjiajie, China, 10–12 August 2019; pp. 1714–1719. [CrossRef]
35. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So CKang, J. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2019**, *36*, 1234–1240. [CrossRef] [PubMed]
36. Huang, K.; Altosaar, J.; Ranganath, R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv* **2019**, arXiv:1904.05342.
37. Beltagy, I.; Cohan, A.; Lo, K. Scibert: Pretrained contextualized embeddings 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) 2019, HongKong, China, 3–7 November 2019.
38. Matthew, E.P.; Mark, N.; Robert, L.; Roy, S.; Vidur, J.; Sameer, S.; Noah, A.S. Knowledge Enhanced Contextual Word Representations. 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), HongKong, China, 3–7 November 2019; pp. 43–54. [CrossRef]
39. Ashish, V.; Noam, S.; Niki, P.; Jakob, U.; Llion, J.; Aidan, N.G.; Lukasz, K.; Illia, P. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
40. Khadhraoui, M.; Bellaaj, H.; Ben Ammar, M.; Hamam, H.; Jmaiel, M. Machine Learning Classification Models with SPD/ED Dataset: Comparative Study of Abstract Versus Full Article Approach. In Proceedings of the ICOST 2020, Hammamet, Tunisia, 24–26 June 2020; pp. 24–26. [CrossRef]