Healthcare Analytics 2 (2022) 100078

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

A review on Natural Language Processing Models for COVID-19 research ☆

Karl Hall [a,*], Victor Chang [b,*], Chrisina Jayne [a]

[a] SCEDT, Teesside University, UK
[b] Operations Information Management, ABS, Aston University, UK

ARTICLE INFO

Keywords:
Natural Language Processing
COVID-19
Machine learning
Transformer models
Sentiment analysis

ABSTRACT

This survey paper reviews Natural Language Processing Models and their use in COVID-19 research in two main areas. Firstly, a range of transformer-based biomedical pretrained language models are evaluated using the BLURB benchmark. Secondly, models used in sentiment analysis surrounding COVID-19 vaccination are evaluated. We filtered literature curated from various repositories such as PubMed and Scopus and reviewed 27 papers. When evaluated using the BLURB benchmark, the novel T-BPLM BioLinkBERT gives groundbreaking results by incorporating document link knowledge and hyperlinking into its pretraining. Sentiment analysis of COVID-19 vaccination through various Twitter API tools has shown the public's sentiment towards vaccination to be mostly positive. Finally, we outline some limitations and potential solutions to drive the research community to improve the models used for NLP tasks.

Contents

☆ Acknowledgement
Prof Chang's work is partly supported by VC Research, UK (VCR 0000183).

* Corresponding authors.
E-mail addresses: K.Hall@tees.ac.uk (K. Hall), v.chang1@aston.ac.uk (V. Chang).

https://doi.org/10.1016/j.health.2022.100078
Received 22 May 2022; Received in revised form 8 July 2022; Accepted 12 July 2022

1. Introduction

Since the outbreak of the Alpha variant of the novel SARS-CoV-2 (COVID-19) virus in December 2019, many different variants of the virus have emerged, such as Delta and Omicron. To cope with the rapidly mutating viruses and their severe impacts, clinical medical research must be accelerated. In the case of the SARS-CoV-2 variant B.1.617.1, classified by the World Health Organization as the Delta variant, which emerged in India in October 2020, more than 15 million cases have been diagnosed in the UK alone. The speed at which the virus is developing has undoubtedly made it more difficult and stressful for researchers to study. Not only do clinicians need to spend time collating information on previous variants, but they also must deal with the clinical manifestations of new variants in patients who have other underlying health conditions themselves [1].

In the age of Big Data, machine learning (ML) is changing the way we approach these challenges. The increasing amount of electronic health records (EHRs) available in tandem with modern data analysis methods has seen the popularity of machine learning in this regard substantially increase in recent times. EHRs often consist of both qualitative and quantitative information, allowing for a mixed-methods approach. In the past, EHR data was more commonly seen as a byproduct of the healthcare delivery process, and only recently has its full potential been realised as a tool for data analysis. When guided by questions posed by healthcare professionals, the use of machine learning techniques to analyse EHR data has proven to be a successful tool for supporting clinical decisions [2]. This approach is currently also being used to fight the global COVID-19 pandemic.

The advancement of ML techniques, most commonly in deep learning for qualitative data and natural language processing for quantitative data, has put them at the forefront of COVID-19 EHR data analysis. Approaches such as this have become commonplace in many other branches of healthcare, including radiology and pathology [3], cancer prognosis and diagnosis prediction [4], mental illness diagnosis [5], diabetes diagnosis and management [6], and more recently, COVID-19 data analysis [7]. Since the outbreak of the pandemic, COVID-19 analysis in particular has been at the forefront of medical research. This paper reviews recent research in this area, focusing more narrowly on transformer-based machine learning models and their performance when used for solving both natural language processing and classification problems relating to the COVID-19 pandemic.

1.1. Natural language processing

Natural language processing (NLP) is a branch of ML concerned with the processing and analysis of language and semantics. Using computational linguistics, ML, and statistical analysis, NLP algorithms provide the ability to process and understand text in a more human-like manner than other models are typically capable of. NLP techniques have previously been successfully employed for use in applications such as voice recognition software for IoT environments [8], chatbots for healthcare assistance [9], colonoscopy procedures [10], and the processing of EHRs in cancer patients [11].

Data used by NLP models is pre-processed using techniques such as lemmatisation and stemming – reducing words to a single root for uniformity, keyword extraction – extracting important phrases and keywords, named entity recognition – classification of named entities into predefined categories, and tokenisation – breaking down text into smaller chunks. NLP algorithms perform language processing tasks such as natural language inference, entity extraction, relation extraction, text classification, question answering and text summarisation [12]. Such tasks are explained further in Section 3, and are used as performance indicators of the models compared and discussed in this review.

Examples of NLP models initially used in healthcare include TF-IDF [13], Word2Vec [14], LSI/LDA [15] and PorterStemmer [16]. Then, trends shifted towards incorporating deep learning methods, such as recurrent neural networks (RNNs) into NLP algorithms. However, RNNs have shown to have some limitations, namely their slow training time and long-range sequence dependencies. The Long Short-Term Memory (LSTM) model, an evolution of RNNs, were developed as a response to these limitations and were somewhat able to overcome them. More recently, in 2017, a new family of related novel state-of-the-art NLP architectures have emerged at the forefront of NLP development, known as transformer models.

1.2. Transformer models

Transformers are the first transduction NLP models relying on self-attention without the use of RNNs or convolutional neural networks (CNNs). In this context, transduction refers to the transformation of input sequences to output sequences within the model. Transformers are designed in such a way that they can handle the dependencies between the inputs and outputs without the use of any attention or recurrence. The architecture of transformer-based models is shown in Fig. 1 and is based on the groundbreaking work outlining the initial proposal of transformer models. Examples of transformer-based NLP

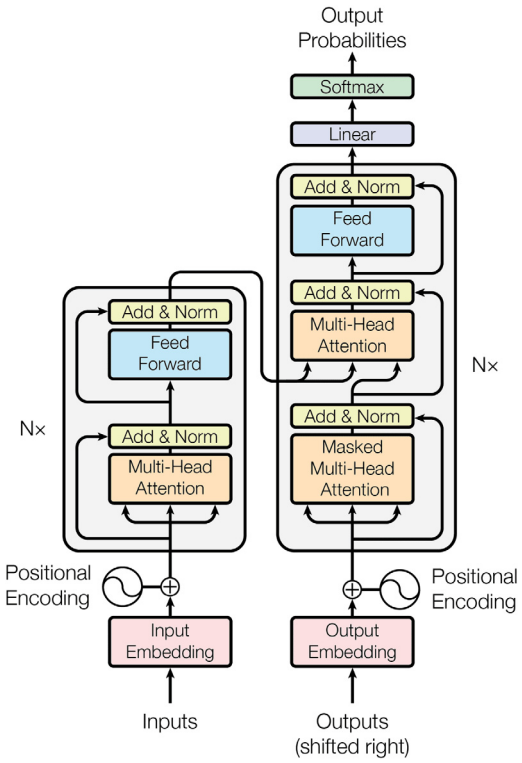

Fig. 1. Transformer model architecture [19].

pipelines discussed in this paper include Google's Bidirectional Encoder Representations from Transformers (BERT) [17] and Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT) [18], a more specialised version of BERT designed for use in biomedical text mining.

1.3. Ensemble models for COVID-19 diagnosis

In ML, ensemble models refer to algorithms where features of more than one model are integrated simultaneously in order to implement a hybrid model approach. Examples of such models include CatBoost, Random Forest, XGBoost and LightGMB. Such models are often much more effective and achieve better results than individual models. This approach of using ensemble transformer models was proposed for image analysis [20].

Ensemble models that combine deep learning with transformer-based models have been used to aid the diagnosis of COVID-19, most commonly by processing images of chest x-rays. For example, a model integrating both Vision Transformers and deep learning algorithms was used to obtain a 97.6% accuracy when diagnosing COVID-19 by analysing chest x-ray images [21].

1.4. COVID-19 sentiment analysis

Sentiment analysis (SNA), also known as opinion mining, is a natural language processing technique concerned with gauging people's opinions on a particular topic to determine whether their overall sentiment is positive, negative, or neutral, but can go beyond this to identify specific emotions towards the topic, such as relief or anger. This is done through the use of a combination of lexicon banks and machine learning techniques. Sentiment analysis is most commonly performed on users' social media posts by evaluating the phrasing of their online discussions. It is often used by businesses to understand consumer opinions of their products. This allows the interested party to tailor their products or policies to better meet customers' needs.

The current landscape of sentiment analysis for COVID-19 mostly revolves around understanding people's thoughts on government policies in various countries surrounding the COVID-19 pandemic and their impact on combating the COVID-19 pandemic. The public's opinions on policies such as those involving vaccinations, self-isolation, business lockdowns and the wearing of face coverings can be better understood by utilising SNA techniques to guide policy making. Furthermore, the general outlook of whether the COVID-19 situation is improving or worsening at any given time can be understood by analysing how sentiments have changed within particular time periods.

Another way in which sentiment analysis can be used lies in combating the constant stream of misinformation being published surrounding the COVID-19 pandemic. Most of this misinformation is published and spread online, as many online platforms provide anonymity, and therefore a lack of accountability when spreading misinformation online. The proliferation of fake news can be damaging to individuals, groups, or society as a whole.

1.5. Research objectives

The objective of this study is to systematically review the current literature on the use of modern transformer-based NLP techniques in COVID-19 research. In particular, the effectiveness of various algorithms are evaluated, compared and discussed when applied to diagnosis, bioinformatics and opinion mining. Furthermore, current NLP trends are discussed, along with open problems and challenges regarding the use of NLP in COVID-19 research.

2. Literature selection

2.1. Collection and criteria

The papers in this review were searched from various publicly available scientific literature databases, namely PubMed, Google Scholar, Crossref and Scopus, with the aim of finding relevant abstracts related to transformers, NLP and COVID-19 research. The CORD-19 Open Research database, curated by semantic scientists at the Allen Institute in collaboration with Microsoft Research, IBM researchers, Kaggle and others, was additionally accessed as a more specialised resource. The CORD-19 database aggregates metadata and full-text data from COVID-19 publications and is updated on a regular basis. For the former repositories, they were accessed through the use of Publish or Perish, a literature screening tool that has the ability to search various literature databases, including the ones used in this study. The search terms used for the screening process were "COVID-19 AND Deep learning OR Natural language processing OR Transformers OR Neural networks". To access the articles within the CORD-19 database, a tool for search functionality was built using Python, which ranks the relevance of the articles returned from the search terms, using publicly available NLP libraries for Python. The searches were limited to English language publications with no publication date restrictions, as any work relating to COVID-19 naturally has a date restriction beginning around early 2020.

2.2. Identification and eligibility screening

Initial literature searches returned 157 records from PubMed, 203 records from Google Scholar, 120 records from Crossref, 144 records from Scopus, and 384 records from CORD-19, for a total of 1008 articles. 295 duplicate papers were removed, leaving a total of 713 abstracts to be screened for eligibility. Of these 713 abstracts, 657 were excluded for not meeting the criteria of being written in English, being a survey paper, not being a full-text article, or not focusing on transformer-based NLP algorithms and their use in COVID-19 research, leaving 56 remaining articles for full-text analysis. Of these, 27 were selected for use in the systematic review. Reasons for exclusion at the

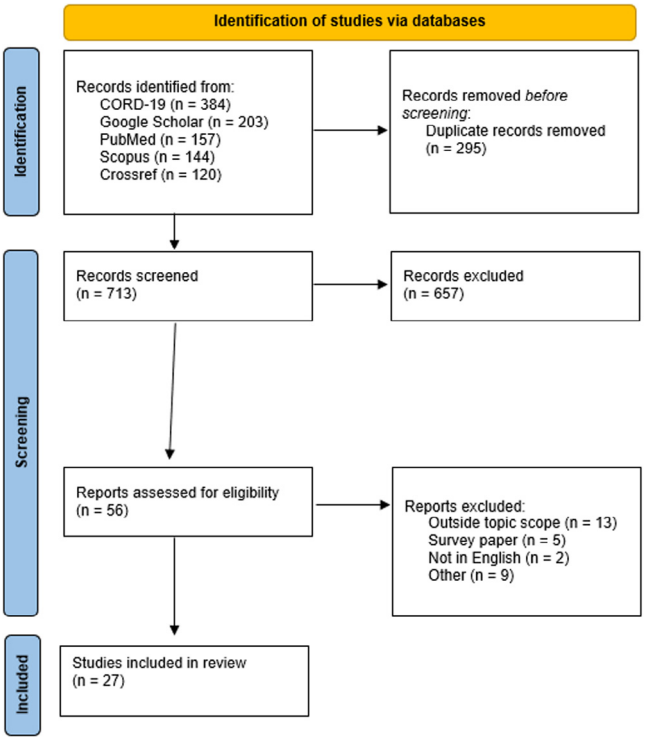

Fig. 2. PRISMA flow diagram of the literature selection process.

final stage included the text not being in English ($n = 2$), not being focused on COVID-19 ($n = 4$) and not being focused on the application of NLP algorithms ($n = 9$). The PRISMA workflow of the selection process and the inclusion criteria is shown in Fig. 2. Critical information from the papers that remained were entered into a tracking spreadsheet to facilitate data extraction for the systematic review. Such information included authors, year of publication, study objectives, methodology, results, and the models used in the study.

3. Transformer design and architecture

3.1. System overview

The main components of transformer-based pretrained language models (T-BPLMs) are embedding and transformer encoder layers. The typical architecture of such models is shown in Fig. 3. The embedding layer returns corresponding vectors from each of the input tokens. The embedding layer consists of at least three layers, providing embedding vectors for each of the input tokens. The vectors of each embedding type are summed to produce the final input vector for each token. By encoding global contextual information using the self-attention mechanism within the model, each input token vector is enhanced. These models have the capability to encode complex language information in the input token vectors by applying a sequence of these transformer encoder layers.

3.2. Embedding layers

The embedding layer within T-BPLMs consists of minimum three sub-layers which apply various embedding types. Depending on the model, there can be more than this. BEHRT, a transformer model for processing electronic health records [23], for example, contains five such embeddings: code, gender, age, position and segment embeddings. The first sub-layer produces sequences of vectors by converting input tokens, and can be based on differing architectures. For example, BERT

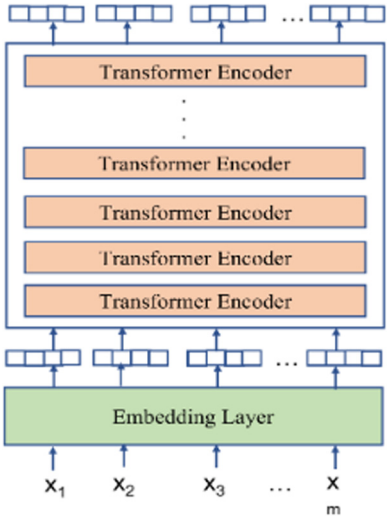

Fig. 3. T-BPLM architecture [22].

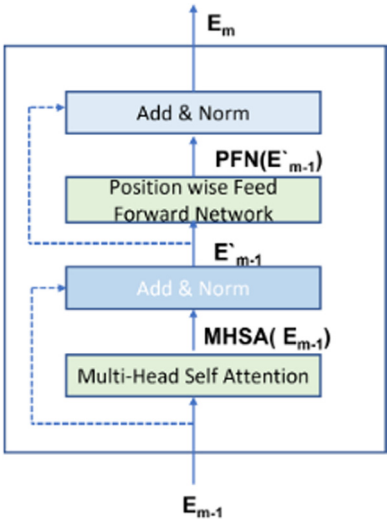

Fig. 4. Transformer encoder layer architecture.

uses WordPiece embeddings [24] and CharacterBERT uses CNN embeddings [25], while MedBERT [26], BERT-EHR [27] and BEHRT [23] use code embeddings.

3.3. Transformer encoders

The architecture of the transformer encoder layer is comprised of Multi-Head Self Attention (MSHA) layers, Position-wise Feed Forward Networks (PFN), and Add and Norm layers (Fig. 4). MSHA iteratively applies self-attention to independently relate the input sequence tokens. The token vectors are subjected to PFN to build non-linear hierarchy features. Finally, Add and Norm layers are added on top of the PFN and MSHA layers and represents residual and layer normalisation to avoid exploding and vanishing gradients.

3.3.1. Self-attention

The self-attention mechanism used by T-BPLMs is an improvement over the convolution and recurrent layers used to encode contextual information in CNNs and RNNs respectively. Self-attention updates input token vectors by encoding contextual information. That is, each token vector is expressed as a weighted sum of all token vectors,

and calculated using attention scores. The final input vector X is transformed into Query (1), Key (2) and Value (3) vectors.

$$Q \in R^{(n\times q)} \tag{1}$$

$$K \in R^{(n\times k)} \tag{2}$$

$$V \in R^{(n\times v)} \tag{3}$$

The output of the self-attention layer is computed by applying the following steps:

1. Computing the similarity matrix $S \in R^{(n\times n)}$
2. Obtaining stable gradients by scaling similarity matrix values and using softmax functions in order to convert similarity scores into probability quantifiers, in order to compute the matrix $P \in R^{(n\times n)}$
3. Computing the value matrix.

3.3.2. Multi-head self attention (MHSA)

There are limitations to only applying one self-attention layer. For example, when processing a particular word, the definition of a word can sometimes only be expressed by using word itself. To get around this, self-attention is applied several times simultaneously using matrices of different weights. Therefore, this process provides the transformer the ability to process multiple positions while carrying out word encodings. The value matrix is then calculated by combining these matrices.

3.3.3. Position-wise feed forward network (PFN)

The PFN consists of two layers. The PFN process is applied to each of the vectors. Different models use different activation functions for the PFN. For example, BERT uses the Gelu [28] activation function.

3.3.4. Add and norm

Add refers to residual connection in this case, while Norm refers to the normalisation of layers. Add and Norm is used for the NHSA and PFN processes within the encoder to avoid exploding and vanishing gradients. T-BPLMs are comprised of sequential transformer encoder layers in addition to the embedding layer. The transformer encoder layers updates the input token vectors by encoding contextual information. The amount of language information encoded is maximised within the model by updating the vectors in sequence.

4. NLP task benchmarking for COVID-19 literature extraction

Researchers have already conducted numerous studies on different variants of COVID-19, including genetic and pharmacological correlations between coronaviruses and other chronic diseases. Although the extent of the links between COVID-19 and other diseases are still at an early stage of development, the field is accumulating a large body of academic research that suggests a complex multidimensional influence between COVID-19 and other diseases. The curation of such large repositories of research articles relating to COVID-19 have fuelled the need for researchers to have the ability to efficiently search for the information they need, given that some of these repositories contain upwards of half a million articles.

To facilitate this, NLP techniques can be implemented through the use of T-BPLMs in order to search for genes and drugs related to other diseases in the COVID-19 literature repositories, allowing research scientists to effectively analyse the rapidly incoming data. T-BPLMs have the capability of performing well across various NLP tasks. In this section, the NLP tasks relevant to analysing COVID-19 literature are explained. Additionally, the performance of these tasks is used to compare the effectiveness of each T-BPLM reviewed in this paper.

4.1. BLURB benchmark

BLURB (Biomedical Language Understanding & Reasoning Benchmark) [29] is the most comprehensive benchmark for assessing the NLP capabilities of Biomedical NLP models. It is comprised of a set of six Biomedical NLP tasks carried out on a standardised set of publicly available datasets. Table 1 shows an overview of the datasets used in BLURB benchmarking. The NLP tasks evaluated with BLURB benchmarking are Named Entity Recognition, Population Interventions Comparator and Outcomes (PICO), Question Answering, Document Classification, Sentence Similarity and Relation Extraction. The BLURB score is the mean performance across these tasks.

4.2. Named entity recognition

Named entities (NEs) carry important information serving as useful targets for NLP algorithms [42]. Named Entity Recognition (NER) has evolved since its inception. Initially, it could only classify general terms such as locations or names of people. Modern T-BPLMs geared towards processing biomedical data can use NER to identify domain-specific technical information, such as drug and proteins relevant to COVID-19 treatment. NE annotations have been proven to provide increased performance on other NLP related tasks, such as Question Answering [43].

4.3. PICO

Population/Problem, Intervention, Comparison, and Outcome (PICO) is a widely adopted framework used for evidence retrieval through the use of clinical question formulation. The PICO framework is specialised in such a way that it can deconstruct the need for evidence into searchable keywords and also provide research question formulation [44]. It has been shown that PICO adoption within T-PBLMs can improve evidence search against popular biomedical literature databases, such as PubMed [45].

4.4. Entity and relation extraction

Entity Extraction (EE) is a text analysis technique applied to unstructured text to extract specific data, and classifies it according to predefined categories. The extraction of biomedical entities such as drugs, or genetic information such as RNA proteins can be used to identify knowledge within biomedical literature related to COVID-19.

Previously, BERT-based models produce contextual delineations and apply some combination of Softmax, BiLSTM and CRF layers. It was found that combining BERT-based models with BiLSTM layers does little to increase entity extraction performance [46], since the transformer encoding layers in BERT-based models essentially do the same thing as the token representations in BiLSTM layering. The best approach to take in this regard lies in first pre-training BERT-based models on task-related data sets before applying them in a more general sense. In [47], a novel EE approach was taken by combining intermediate fine-tuning with semi-supervised learning. More recently, the task of EE was reformulated as question answering [48] and it was found that combining BioBERT with Question Answering outperformed the combination of BioBERT and Softmax and BiLSTM-CRF layers on six different data sets.

As a natural progression from EE, Relation Extraction (RE) is a task concerned with understanding the semantic relationships between entities within the text. The process of performing EE as the precursor to RE allows for the conversion of unstructured text into structured data which can be useful for increasing performance in other NLP tasks. The best performance was achieved using MIMIC-BERT in combination with Softmax layers [49]. It was also shown that combining SciBERT with Softmax layers produces better outcomes than combining BERT with Softmax layers when processing biomedical RE data sets [50].

Table 1
Datasets used in the BLURB benchmark.

Dataset	Task	Train	Dev	Test	Evaluation metrics
BC5-chem [30]	Named Entity Recognition	5,203	5,347	5,385	F1 entity-level
BC5-disease [30]	Named Entity Recognition	4,182	4,244	4,424	F1 entity-level
NCBI-disease [31]	Named Entity Recognition	5,134	787	960	F1 entity-level
BC2GM [32]	Named Entity Recognition	15,197	3,061	6,325	F1 entity-level
JNLPBA [33]	Named Entity Recognition	46,750	4,551	8,662	F1 entity-level
EBM PICO [34]	PICO	339,167	85,321	16,364	Macro F1 word-level
ChemProt [35]	Relation Extraction	18,035	11,268	16,364	Micro F1
DDI [36]	Relation Extraction	25,296	2,496	5,716	Micro F1
GAD [37]	Relation Extraction	4,261	535	534	Micro F1
BIOSSES [38]	Sentence Similarity	64	16	20	Pearson
HoC [39]	Document Classification	1,295	186	371	Micro F1
PubMedQA [40]	Question Answering	450	50	500	Accuracy
BioASQ [41]	Question Answering	670	75	140	Accuracy

Following on from this, a BioBERT model was used with extra attention layering to achieve even better results by concatenating CLS vectors and weighted sum vectors [51]. They also concluded that this additional attention layer achieves better results than using an LSTM approach.

4.5. Sentence similarity

Semantic Textual Similarity (SS) is a quantifier concerned with measuring the semantic similarity of sentence pairs. While they are both concerned with sentence-level semantics, SS is slightly different to Natural Language Inference (NLI), as SS outputs a numeric quantifier that measures the similarity degree, whereas NLI takes a classification-based approach. SS has proven useful to solve tasks such as question answering [52], duplicate text detection [53] and topic relatedness [54]. It was also proven that training T-BPLMs on SS data sets increases models' ability to learn sentence-level semantics [55]. Like with NLI, BERT-based models can process both sentences within the sentence pair simultaneously.

Similarly to EE data sets, SS data sets are typically small, so the best approach appears to be pre-training models on SS data sets before fine-tuning on more generalised clinical data sets [56]. A Pearson correlation score of 0.83 was also achieved by fine-tuning Clinical BERT using combination of SS and clinical data sets [53].

4.6. Document classification

Document classification is a process of assigning categories or classes to documents to make them easier to manage, search, filter, or analyse. This is usually achieved through the use of a T-BPLM encoder and a Softmax classification is utilised to calculate the corresponding label probabilities. Information has been processed about medication prescriptions and achieved good results by using a variety of different T-BPLMs [57], such as BERT, RoBERTa, ALBERT and DistillBERT. Good results were also shown when using PubMedBERT and BioBERT models by using rule-based NLP algorithms [58]. Like with RE, it was shown that using the base BERT model and providing it with additional layering improves its performance [59].

4.7. Question answering

Question Answering (QA) is concerned with the processing of semantic questions or queries and produce corresponding answers. Question Answering can save a lot of time when processing biomedical literature of clinical notes, which are often highly complex. However, due to the complexity involved, developing substantial QA data sets can be problematic. The approach of Biomedical Entity Masking was introduced [60], following a similar approach to those solving the optimisation of other NLP tasks. Biomedical models' performance can be improved by first pre-training on more specialised data sets. NER has been used as an auxiliary task while training in order to build a model based on BioBERT that specialised in QA, delivering state-of-the-art performance levels [61].

4.8. Evaluation

Using performance benchmarks is an effective way to standardise comparisons between NLP models. The first such proposed benchmark standard was GLUE [69], and other popular benchmarks have since emerged, such as NAS-Bench-NLP [70], ERASER [71], and BLURB. We use the BLURB benchmark in this regard since it has become the most widely used, and comparisons of the T-BPLMs performance across the BLURB metrics are shown in Table 2.

From this, we can see that BioLinkBERT outperforms other T-BPLMs in all but one of the performance indicators and significantly outperforms in general. In particular, it performs much better at QA tasks. This can be attributed to its novel approach of incorporating document link knowledge, such as hyperlinking between information by pretraining on Wikipedia hyperlinks and PubMed with citation links.

5. Sentiment analysis

The proliferation of social media in the past decade has allowed information to spread much more easily. This poses some challenges, particularly regarding the spread of misinformation surrounding contentious issues, including issues surrounding the COVID-19 pandemic. This can mostly be attributed to social media platforms not holding users to the same journalistic integrity as traditional journalism. As a result of this, "fake news" or similar unfounded rumours, can easily be spread exponentially on these platforms, which can often be harmful to its users. This phenomenon has been studied extensively since the start of the pandemic.

One major way in which SNA has been applied in this regard is to measure the public's opinions towards vaccinations and associated government rules and policies. As public opinions have a lot of power, this information is valuable to governments and healthcare providers, providing them access to accurate insights and consistent analysis.

Unlike most other social media platforms, Twitter provides public access to its API platform. Because of this, it has become the platform of choice for researchers conducting SNA research. Typically, SNA results return a classification as either positive, negative, or neutral towards a particular topic. In this section, we focus on reviewing the literature surrounding SNA of COVID-19 vaccination hesitancy.

5.1. Tools

The RTweet tool [72] was developed to provide accessible web scraping functionality for the R language and is commonly used [73–75]. The "Twint" scraping tool bypasses Twitter API restrictions [76–

Table 2
Evaluation of T-BPLM performance on NLP tasks as defined in the BLURB benchmark.

Model	NER	PICO	RE	SS	DC	QA	BLURB score
BioLinkBERT-Large [62]	**87.01**	74.19	**82.74**	**93.63**	**84.87**	**83.50**	**84.30**
BioLinkBERT-Base [62]	86.40	73.97	81.56	93.25	84.35	80.82	83.39
PubMedBERT-Large [63]	86.28	73.61	81.77	92.73	82.70	80.37	82.91
BioELECTRA [64]	86.74	**74.26**	81.56	92.49	83.50	76.30	82.48
PubMedELECTRA-Large [63]	85.98	74.02	80.52	92.69	82.37	79.08	82.44
PubMedBERT-Base [29]	86.08	73.38	81.19	92.30	82.32	71.70	81.16
BioBERT [18]	85.81	73.18	79.79	89.52	81.54	72.19	80.34
PubMedELECTRA-Base [63]	85.65	73.70	80.17	80.24	81.45	77.68	79.81
SciBERT [65]	85.43	73.12	79.56	86.25	80.66	68.12	78.86
ClinicalBERT [66]	83.99	72.06	76.91	91.23	80.74	58.79	77.29
RoBERTa [67]	83.09	73.02	77.71	81.25	79.66	64.02	76.46
BlueBERT [68]	84.50	72.54	76.13	85.38	80.48	58.58	76.27
BERT [17]	82.99	72.34	77.44	82.68	80.20	60.99	76.11

Table 3
Classification model summary.

Research	Classification model
Paul and Gokhale (2020)	Random Forest
Paul and Gokhale (2020) Garcia and Berton (2021) Nurdeni et al. (2021)	SVM
Paul and Gokhale (2020)	Multilayer Perceptron
Garcia and Berton (2021) Nurdeni et al. (2021) Ritonga et al. (2021) Pano and Kashef (2020)	Naive Bayes
Paul and Gokhale (2020) To et al. (2021)	LSTM
Garcia and Berton (2021) Muller et al. (2020) Muller and Salathé (2020) To et al. (2021)	BERT

78]. Netlytic is a tool that queries the Twitter REST API [79,80]. Crowdbreak is a digital platform utilising the stream filter API and was used in COVID-Twitter-BERT [81,82]. Tweepy [83], is a popular Python package used to communicate with the Twitter API [84–86]. Rapid Miner is a tool that can also be used for clustering and classification [87,88].

5.2. Classification models

Approaches taken varied, but either consisted of classical ML models, such as SVM and Logistic Regression, DL models such as LSTM and Bi-LSTM, or BERT-based models. The general consensus was that BERT-based transformer models were the most effective for SNA as they use non-sequential processing, multi-head self-attention, and positional embeddings instead of recurrence to provide improved understanding of relationships between words. Other models, such as LSTM-based or other DL models, use sequential processing (sentences are processed in sequence, word by word). Therefore, in these models, each state is dependent on the previously processed information, resulting in worse performance than BERT-based models. Table 3 shows a summary of the classification models used by the reviewed literature.

5.2.1. Random forest

Random Forest [89] is an ensemble ML algorithm consisting of multiple decision trees. RF looks to correct the well-known problem that Decision Trees have — overfitting to the data used for training. Modern RF models employ a technique known as bootstrap aggregating, or bagging, in order to mitigate this, hence improving performance. RF is most commonly used for classification and regression tasks, including classification-based sentiment analysis.

5.2.2. Support vector machine

Support Vector Machines [90] is a well established classical supervised ML model used for both linear and non-linear classification. Hyperplanes are used to maximise the distance between classifications and is oriented in a way such that its margin is maximised, by minimising the equation shown in (4).

$$\left[\frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y_i(w^T x_i - b))\right] + \lambda\|w\|^2. \tag{4}$$

subject to $y_i(w^T x_i - b) \geq 1 - \zeta_i$, where $\zeta_i = max(0, 1 - y_i(w^T x_i - b))$ and $\zeta_i \geq 0$ for all i.

5.2.3. Multilayer perceptron

Multilayer perceptron [91] is a type deep learning model, more specifically, part of the feed-forward neural network family of algorithms. Consisting of multiple layers of perceptrons, MLPs are particularly effective at binary classification, but can also be used for multiclass classification. MLP utilises a linear activation function, most commonly the sigmoid activation functions which can be described as in (5) and (6):

$$y(v_i) = tanh(v_i) \tag{5}$$

$$y(v_i) = (1 + e^{-}v_i)^{-1} \tag{6}$$

Like other neural networks, MLP models typically consist of multiple layers, where the nodes are interconnected by a weight value w_{i_j} to the nodes in the next layer. The value of these weights is updated as the data is passed through by the use of backpropagation.

5.2.4. Naive Bayes

Naive Bayes [92] are a family of probability-based classifiers based on Bayes' theorem, which is stated as the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{7}$$

where A and B are probabilistic events, P(A) and P(B) are the probabilities of A and B independently occurring, and $P(B) \neq 0$.

NB is considered "naive" as it assumes that each of the data features are independent of one another. These models are popular due to them being highly scalable on large data sets and are commonly used for document classification NLP tasks.

5.2.5. Long short-term memory

Long Short-Term Memory [93] models are an artificial neural network similar to MLP but with additional features. While MLP models have a more simplistic feedforward architecture, LSTM utilises feedback mechanisms, and is therefore classified as a recurrent neural network. LSTM uses gates (input, output and forget and state candidate) which regulate the direction of information through the model, while making use of time backpropagation to alter the weight values. The LSTM model architecture is shown in Fig. 5.

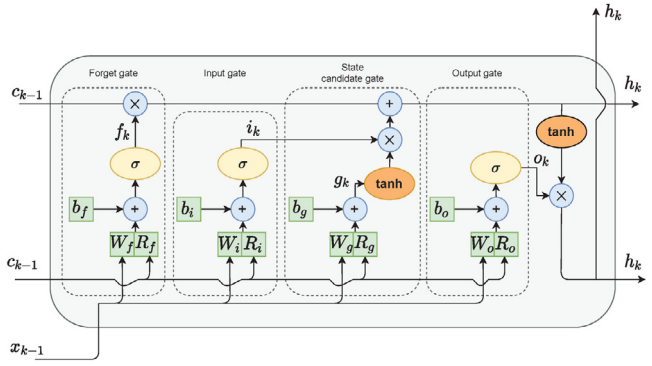

Fig. 5. LSTM model architecture [94].

5.3. Results

All of the studies classified the sentiment analysis results using a positive or negative polarity classification rating, commonly denoted as f_s. There are a various number of ways of calculating sentiment score, but the main goal is still the same. That is, to quantify whether people feel negatively or positively towards COVID-19. Some of the more popular methods used in the reviewed studies are outlined below. All of the SNA methods follow a lexicon-based prediction approach by referencing a pre-labelled sentiment dataset. In this case, $\sum w_p$ and $\sum w_n$ refer to all positive and negative encoded sentiment words respectively, and $\sum w_a$ refers to all words combined.

5.3.1. Absolute proportional difference

The Absolute Proportional Difference (APD) function results in an f_n score with boundaries between 0 and 1, with 0 being the most negative and 1 being the most positive. The sum of all negative words is subtracted from the sum of all positive words and divided by the total number of words as shown in (8). Because of this, f_s can be influenced by words not used to calculate the sentiment score and is therefore not an ideal function to use. However, it has been shown to perform well on smaller passages of text, which explains its popularity when analysing Twitter posts, which have a 280 character limit.

$$f_s = \frac{\sum w_p - \sum w_n}{\sum w_a} \tag{8}$$

5.3.2. Relative proportional difference

The Relative Proportional Difference (RPD) function results in an f_n score with boundaries between −1 and 1, with −1 being the most negative and 1 being the most positive. The main appeal of RPD, is that unlike APD, RPD does not use non-sentiment words in its calculation (9). However, this leaves RPD prone to uneven distribution and clustering around the extremes of the classification scale.

$$f_s = \frac{\sum w_p - \sum w_n}{\sum w_p + \sum w_n} \tag{9}$$

5.3.3. Logit scale

The Logit Scale has no predefined numerical boundaries or endpoints, allowing for the sentiment score to essentially range from -∞ to +∞. The function is adapted from political science [95], and was initially used for measuring the political spectrum of left and right. rooted in exponents though, making it increasingly difficult to achieve higher or lower sentiment scores. This is calculated using the empirical logit formula as shown in (10).

$$f_s = log(P + 0.5) - log(N + 0.5) \tag{10}$$

Logit-based approaches are considered more balanced and are symmetrical around 0. Since log(0) is considered undefined, the function is modified to prevent this from occurring by using β as a fixed coefficient as shown in (11).

$$f_s = log(\sum w_p + \beta) - log(\sum w_n + \beta) \tag{11}$$

Table 4
SNA results summary.

Research	Sentiment result
Dubey (2021)	
Melton et al. (2021)	
Paul and Gokhale (2020)	
Kwok et al. (2021)	
Rahul et al. (2021)	
Sv et al. (2021)	
Muller et al. (2020)	
To et al. (2021)	
Nurdeni et al. (2021)	
Pano and Kashef (2020)	Positive
Garcia and Berton (2021)	
Ritonga et al. (2021)	Negative

5.4. Study comparison

From the results shown in Table 4, it can be deduced that the sentiment towards COVID-19 vaccination uptake is mostly positive, with 84.6% of studies coming to this conclusion.

6. Future directions

6.1. Domain cost

The most common approach when pretraining T-BPLMs is Mixed-Domain Pretraining (MDPT). MDPT involves pretraining on massive amounts of domain-specific data and requires high end GPUs or Tensor Processing Units. This process can take days or even weeks. While this approach is successful at developing effective T-BPLMs, they require high energy and computing costs. Therefore, there is a need for ways to adopt lower cost domains. One such way currently being explored is through the use of Task Adaptive Pretraining (TAPT). TAPT provides the ability for T-BPLMS to learn both in-domain knowledge by pretraining on small datasets relative to MDPT. Another way of reducing the cost associated with developing T-BPLMs is by taking generic T-BPLMs and further refining their embedding layer with additional in-domain text [96].

6.2. Dataset size

Pretraining on larger amounts of biomedical text is the key to improving task-specific performance. Many of the specialist datasets used for pretraining T-BPLMs are very small, particularly those relating to Sentence Similarity and Question Answering. However, it is not always possible to get a large biomedical datasets. Larger but less related datasets can be used to mitigate this somewhat. Other ways to approach this problem include using semi-supervised learning with pseudo-labelled data to allow for larger datasets [47], using backtranslation to train the similarity models [97], and using intermediate [98] or multi-task [99] fine tuning to gain additional domain specific knowledge.

6.3. Benchmarking

While there are many emerging biomedical NLP benchmarking frameworks, they are designed for evaluating the performance of literature-based datasets. That is to say, they are not particularly useful at analysing performance relating to EHR data or sentiment analysis. There is a need for benchmark frameworks for these areas, similar to the BLURB benchmark used to evaluate the literature-based datasets. Additionally, the currently existing benchmarks are not capable of evaluating some important characteristics of T-BPLMs, such as compactness and robustness.

6.4. Model efficiency

As pretraining ideally requires substantial computing power, time, and huge datasets to pretrain on, novel models that can reduce both the pretraining time and size of the pretraining corpus required are needed. Models such as ConvBERT [100] and DeBERTa [101] have been proposed to try and address both of these problems. For example. ConvBERT can outperform other T-BPLMs while only using 25% or less of their pretraining costs.

6.5. Quantitative and time-sensitive SNA

COVID-19 SNA approaches typically take a classification approach (positive, negative, neutral). There is a large issue with this approach in that it does not quantify the sentiments. Quantitative SNA has been conducted in other areas, allowing for a more detailed understanding. Furthermore, it can be possible to use SNA to get snapshots of time-sensitive public sentiment. This can be used to analyse how sentiments change over time, or in response to important events.

6.6. Bias mitigation

As T-BPLMs are used in sensitive real-world applications, such as medicine and law, it is crucial that these models do not develop systematic bias against certain groups of people. It has already been shown to be the case in some instances, due to bias in the datasets used for pretraining. These biases can manifest in many ways, such as gender bias or racial bias. It was shown, for example, that SciBERT has ethnicity bias as a result of this [65], and that its performance varies for different protected attributes. Regarding the identification and reduction of gender bias, a data augmentation approach has been proposed [102] to mitigate this. An A-INLP framework [103] was also proposed to identify bias-sensitive tokens in order to ensure fairness.

7. Conclusion

In this survey, We introduce transformer-based models, and explain why and how they are state-of-the-art, and why they are the future of NLP.

We summarise various models and their related studies that used NLP models for COVID-19 research in the two key areas of literature extraction and sentiment analysis.

We explain the BLURB benchmarking framework and use it to assess the performance of various transformer-based models at performing NLP tasks related to COVID-19 literature extraction. We also review sentiment analysis surrounding COVID-19 vaccination hesitancy. Finally, we discuss some of the challenges and directions for future research which we hope will improve the use of T-BPLMs going forward.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] F. Caramelo, N. Ferreira, B. Oliveiros, Estimation of risk factors for COVID-19 mortality-preliminary results, 2020, MedRxiv.

[2] T.B. Murdoch, A.S. Detsky, The inevitable application of big data to health care, JAMA 309 (13) (2013) 1351–1352.

[3] S. Jha, E.J. Topol, Adapting to artificial intelligence: radiologists and pathologists as information specialists, JAMA 316 (22) (2016) 2353–2354.

[4] K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Comput. Struct. Biotechnol. J. 13 (2015) 8–17.

[5] G. Cho, J. Yim, Y. Choi, J. Ko, S.-H. Lee, Review of machine learning algorithms for diagnosing mental illness, Psychiatry Investig. 16 (4) (2019) 262.

[6] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, I. Chouvarda, Machine learning and data mining methods in diabetes research, Comput. Struct. Biotechnol. J. 15 (2017) 104–116.

[7] A. Alimadadi, S. Aryal, I. Manandhar, P.B. Munroe, B. Joe, X. Cheng, Artificial intelligence and machine learning to fight COVID-19, Physiol. Genomics 52 (4) (2020) 200–202.

[8] P.J. Rani, J. Bakthakumar, B.P. Kumaar, U.P. Kumaar, S. Kumar, Voice controlled home automation system using natural language processing (NLP) and internet of things (IoT), in: 2017 Third International Conference on Science Technology Engineering & Management, ICONSTEM, IEEE, 2017, pp. 368–373.

[9] S. Ayanouz, B.A. Abdelhakim, M. Benhmed, A smart chatbot architecture based NLP and machine learning for health care assistance, in: Proceedings of the 3rd International Conference on Networking, Information Systems & Security, 2020, pp. 1–6.

[10] O.V. Patterson, T.B. Forbush, S.D. Saini, S.E. Moser, S.L. DuVall, Classifying the indication for colonoscopy procedures: a comparison of NLP approaches in a diverse national healthcare system, in: EHealth-Enabled Health, MEDINFO 2015, IOS Press, 2015, pp. 614–618.

[11] S. Datta, E.V. Bernstam, K. Roberts, A frame semantic overview of NLP-based information extraction for cancer-related EHR notes, J. Biomed. Inform. 100 (2019) 103301.

[12] R. Zhu, X. Tu, J.X. Huang, Utilizing BERT for biomedical and clinical text mining, in: Data Analytics in Biomedical Engineering and Healthcare, Elsevier, 2021, pp. 73–103.

[13] M. Alodadi, V.P. Janeja, Similarity in patient support forums using TF-IDF and cosine similarity metrics, in: 2015 International Conference on Healthcare Informatics, IEEE, 2015, pp. 521–522.

[14] O. Jacobson, H. Dalianis, Applying deep learning on electronic health records in Swedish to predict healthcare-associated infections, in: Proceedings of the 15th Workshop on Biomedical Natural Language Processing, 2016, pp. 191–195.

[15] A. Alibasic, M.C.E. Simsekler, T. Kurfess, W.L. Woon, M.A. Omar, Utilizing data science techniques to analyze skill and demand changes in healthcare occupations: case study on USA and UAE healthcare sector, Soft Comput. 24 (7) (2020) 4959–4976.

[16] A.R. Kulkarni, S.D. Mundhe, An application of porters stemming algorithm for text mining in healthcare, Int. J. Manag. IT Eng. 7 (11) (2019) 223–228.

[17] I. Tenney, D. Das, E. Pavlick, BERT rediscovers the classical NLP pipeline, 2019, arXiv preprint arXiv:1905.05950.

[18] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[21] K.S. Krishnan, K.S. Krishnan, Vision transformer based COVID-19 detection using chest X-rays, in: 2021 6th International Conference on Signal Processing, Computing and Control, ISPCC, IEEE, 2021, pp. 644–648.

[22] K.S. Kalyan, A. Rajasekharan, S. Sangeetha, AMMU: a survey of transformer-based biomedical pretrained language models, J. Biomed. Inform. (2021) 103982.

[23] Y. Li, S. Rao, J.R.A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, G. Salimi-Khorshidi, BEHRT: transformer for electronic health records, Sci. Rep. 10 (1) (2020) 1–12.

[24] Y. Wu, M. Schuster, Z. Chen, Q.V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al., Google's neural machine translation system: Bridging the gap between human and machine translation, 2016, arXiv preprint arXiv:1609.08144.

[25] H.E. Boukkouri, O. Ferret, T. Lavergne, H. Noji, P. Zweigenbaum, J. Tsujii, CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters, 2020, arXiv preprint arXiv:2010.10392.

[26] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, D. Zhi, Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction, NPJ Digit. Med. 4 (1) (2021) 1–13.

[27] Y. Meng, W. Speier, M.K. Ong, C.W. Arnold, Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression, IEEE J. Biomed. Health Inf. 25 (8) (2021) 3121–3129.

[28] D. Hendrycks, K. Gimpel, Gaussian error linear units (gelus), 2016, arXiv preprint arXiv:1606.08415.

[29] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Trans. Comput. Healthc. (HEALTH) 3 (1) (2021) 1–23.

[30] J. Li, Y. Sun, R.J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A.P. Davis, C.J. Mattingly, T.C. Wiegers, Z. Lu, BioCreative V CDR task corpus: a resource for chemical disease relation extraction, Database 2016 (2016).

[31] R.I. Doğan, R. Leaman, Z. Lu, NCBI disease corpus: a resource for disease name recognition and concept normalization, J. Biomed. Inform. 47 (2014) 1–10.

[32] L. Smith, L.K. Tanabe, C.-J. Kuo, I. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C.M. Friedrich, K. Ganchev, M. Torii, et al., Overview of BioCreative II gene mention recognition, Genome Biol. 9 (2) (2008) 1–19.

[33] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, N. Collier, 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, Citeseer, 2004, pp. 70–75.

[34] B. Nye, J.J. Li, R. Patel, Y. Yang, I.J. Marshall, A. Nenkova, B.C. Wallace, A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting, 2018, NIH Public Access, 2018, p. 197.

[35] M. Krallinger, O. Rabal, S.A. Akhondi, M.P. Pérez, J. Santamaría, G.P. Rodríguez, G. Tsatsaronis, A. Intxaurrondo, J.A. López, U. Nandal, et al., Overview of the BioCreative VI chemical-protein interaction track, in: Proceedings of the Sixth BioCreative Challenge Evaluation Workshop, 1, 2017, pp. 141–146.

[36] M. Herrero-Zazo, I. Segura-Bedmar, P. Martínez, T. Declerck, The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions, J. Biomed. Inform. 46 (5) (2013) 914–920.

[37] A. Bravo, J. Piñero, N. Queralt-Rosinach, M. Rautschka, L.I. Furlong, Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research, BMC Bioinformatics 16 (1) (2015) 1–17.

[38] G. Soğancıoğlu, H. Öztürk, A. Özgür, BIOSSES: a semantic sentence similarity estimation system for the biomedical domain, Bioinformatics 33 (14) (2017) i49–i58.

[39] S. Baker, I. Silins, Y. Guo, I. Ali, J. Högberg, U. Stenius, A. Korhonen, Automatic semantic classification of scientific literature according to the hallmarks of cancer, Bioinformatics 32 (3) (2016) 432–440.

[40] Q. Jin, B. Dhingra, Z. Liu, W.W. Cohen, X. Lu, PubMedQA: A dataset for biomedical research question answering, 2019, arXiv preprint arXiv:1909.06146.

[41] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M.R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, et al., An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition, BMC Bioinformatics 16 (1) (2015) 1–28.

[42] B. Mohit, Named entity recognition, in: Natural Language Processing of Semitic Languages, Springer, 2014, pp. 221–245.

[43] A. Toral, E. Noguera, F. Llopis, R. Munoz, Improving question answering using named entity recognition, in: International Conference on Application of Natural Language to Information Systems, Springer, 2005, pp. 181–191.

[44] W.S. Richardson, M.C. Wilson, J. Nishikawa, R.S. Hayward, et al., The well-built clinical question: a key to evidence-based decisions, ACP J. Club 123 (3) (1995) A12–A13.

[45] C. Schardt, M. Adams, T. Owens, S. Keitz, P. Fontelo, BMC medical informatics and utilization of the PICO framework to improve searching PubMed for clinical questions. 6, 1–6, 2007.

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

[47] S. Gao, O. Kotevska, A. Sorokine, J.B. Christian, A pre-training and self-training approach for biomedical named entity recognition, PLoS One 16 (2) (2021) e0246310.

[48] C. Sun, Z. Yang, L. Wang, Y. Zhang, H. Lin, J. Wang, Biomedical named entity recognition using BERT in the machine reading comprehension framework, J. Biomed. Inform. 118 (2021) 103799.

[49] Q. Wei, Z. Ji, Y. Si, J. Du, J. Wang, F. Tiryaki, S. Wu, C. Tao, K. Roberts, H. Xu, Relation extraction from clinical narratives using pre-trained language models, in: AMIA Annual Symposium Proceedings, 2019, American Medical Informatics Association, 2019, p. 1236.

[50] X. Liu, J. Fan, S. Dong, et al., Document-level biomedical relation extraction leveraging pretrained self-attention structure and entity replacement: Algorithm and pretreatment method validation study, JMIR Med. Inf. 8 (5) (2020) e17644.

[51] P. Su, K. Vijay-Shanker, Investigation of bert model on biomedical relation extraction based on revised fine-tuning mechanism, in: 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2020, pp. 2522–2529.

[52] D. Hoogeveen, A. Bennett, Y. Li, K.M. Verspoor, T. Baldwin, Detecting misflagged duplicate questions in community question-answering archives, in: Twelfth International AAAI Conference on Web and Social Media, 2018.

[53] F.W. Mutinda, S. Nigo, D. Shibata, S. Yada, S. Wakamiya, E. Aramaki, Detecting redundancy in electronic medical records using clinical bert, 2020.

[54] K.S. Kalyan, S. Sangeetha, A hybrid approach to measure semantic relatedness in biomedical concepts, 2021, arXiv preprint arXiv:2101.10196.

[55] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, 2019, arXiv preprint arXiv:1908.10084.

[56] D. Mahajan, A. Poddar, J.J. Liang, Y.-T. Lin, J.M. Prager, P. Suryanarayanan, P. Raghavan, C.-H. Tsou, et al., Identification of semantically similar sentences in clinical notes: Iterative intermediate training using multi-task learning, JMIR Med. Inf. 8 (11) (2020) e22508.

[57] M.A. Al-Garadi, Y.-C. Yang, H. Cai, Y. Ruan, K. O'Connor, G.-H. Graciela, J. Perrone, A. Sarker, Text classification models for the automatic detection of nonmedical prescription medication use from social media, BMC Med. Inf. Decis. Making 21 (1) (2021) 1–13.

[58] Z. Shen, Y. Yi, A. Bompelli, F. Yu, Y. Wang, R. Zhang, Extracting lifestyle factors for alzheimer's disease from clinical notes using deep learning with weak supervision, 2021, arXiv preprint arXiv:2101.09244.

[59] M. Tang, P. Gandhi, M.A. Kabir, C. Zou, J. Blakey, X. Luo, Progress notes classification and keyword extraction using attention-based deep learning models with BERT, 2019, arXiv preprint arXiv:1910.05786.

[60] G. Pergola, E. Kochkina, L. Gui, M. Liakata, Y. He, Boosting low-resource biomedical qa via entity-aware masking strategies, 2021, arXiv preprint arXiv:2102.08366.

[61] A. Akdemir, T. Shibuya, Transfer learning for biomedical question answering, in: CLEF (Working Notes), 2020.

[62] M. Yasunaga, J. Leskovec, P. Liang, LinkBERT: pretraining language models with document links, 2022, arXiv preprint arXiv:2203.15827.

[63] R. Tinn, H. Cheng, Y. Gu, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Fine-tuning large neural language models for biomedical natural language processing, 2021, arXiv preprint arXiv:2112.07869.

[64] K. Kanakarajan, B. Kundumani, M. Sankarasubbu, BioELECTRA: pretrained biomedical text encoder using discriminators, in: Proceedings of the 20th Workshop on Biomedical Language Processing, 2021, pp. 143–154.

[65] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, 2019, arXiv preprint arXiv:1903.10676.

[66] K. Huang, J. Altosaar, R. Ranganath, Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2019, arXiv preprint arXiv:1904.05342.

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

[68] Y. Peng, S. Yan, Z. Lu, Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets, 2019, arXiv preprint arXiv:1906.05474.

[69] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for NLP, in: International Conference on Machine Learning, PMLR, 2019, pp. 2790–2799.

[70] N. Klyuchnikov, I. Trofimov, E. Artemova, M. Salnikov, M. Fedorov, E. Burnaev, NAS-Bench-NLP: neural architecture search benchmark for natural language processing, 2020, arXiv preprint arXiv:2006.07116.

[71] J. DeYoung, S. Jain, N.F. Rajani, E. Lehman, C. Xiong, R. Socher, B.C. Wallace, ERASER: A benchmark to evaluate rationalized NLP models, 2019, arXiv preprint arXiv:1911.03429.

[72] M.W. Kearney, rtweet: Collecting Twitter data, 7, 2018, pp. 1–72, R Package Version 0.6.

[73] A.D. Dubey, Public sentiment analysis of COVID-19 vaccination drive in india, 2021, Available at SSRN 3772401.

[74] N. Paul, S.S. Gokhale, Analysis and classification of vaccine dialogue in the coronavirus era, in: 2020 IEEE International Conference on Big Data (Big Data), IEEE, 2020, pp. 3220–3227.

[75] S. Kwok, S. Vadde, G. Wang, Twitter speaks: an analysis of australian twitter users' topics and sentiments about COVID-19 vaccination using machine learning, J. Med. Internet Res. (2021).

[76] T. Nuzhath, S. Tasnim, R.K. Sanjwal, N.F. Trisha, M. Rahman, S.F. Mahmud, A. Arman, S. Chakraborty, M.M. Hossain, COVID-19 vaccination hesitancy, misinformation and conspiracy theories on social media: A content analysis of Twitter data, 2020, SocArXiv.

[77] K. Rahul, B.R. Jindal, K. Singh, P. Meel, Analysing public sentiments regarding COVID-19 vaccine on twitter, in: 2021 7th International Conference on Advanced Computing and Communication Systems, Vol. 1, ICACCS, IEEE, 2021, pp. 488–493.

[78] P. Sv, J. Tandon, H. Hinduja, et al., Indian citizen's perspective about side effects of COVID-19 vaccine–a machine learning study, Diabetes Metab. Syndr.: Clin. Res. Rev. 15 (4) (2021) 102172.

[79] K. Garcia, L. Berton, Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA, Appl. Soft Comput. 101 (2021) 107057.

[80] H. Dashtian, D. Murthy, Cml-covid: A large-scale covid-19 twitter dataset with latent topics, sentiment and location information, 2021, arXiv preprint arXiv:2101.12202.

[81] M. Müller, M. Salathé, P.E. Kummervold, Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter, 2020, arXiv preprint arXiv:2005.07503.

[82] M. Müller, M. Salathé, Addressing machine learning concept drift reveals declining vaccine sentiment during the COVID-19 pandemic, 2020, arXiv preprint arXiv:2012.02197.

[83] J. Roesslein, Tweepy documentation, 5, 2009, [Online] http://tweepy.readthedocs.io/en/v3.

[84] A.A. Chaudhri, S. Saranya, S. Dubey, Implementation paper on analyzing COVID-19 vaccines on twitter dataset using tweepy and text blob, Ann. Rom. Soc. Cell Biol. (2021) 8393–8396.
[85] Q.G. To, K.G. To, V.-A.N. Huynh, N.T. Nguyen, D.T. Ngo, S.J. Alley, A.N. Tran, A.N. Tran, N.T. Pham, T.X. Bui, et al., Applying machine learning to identify anti-vaccination tweets during the COVID-19 pandemic, Int. J. Environ. Res. Public Health 18 (8) (2021) 4069.
[86] D.A. Nurdeni, I. Budi, A.B. Santoso, Sentiment analysis on Covid19 vaccines in Indonesia: From the perspective of Sinovac and Pfizer, in: 2021 3rd East Indonesia Conference on Computer and Information Technology, EIConCIT, IEEE, 2021, pp. 122–127.
[87] M. Ritonga, M.A. Al Ihsan, A. Anjar, F.H. Rambe, et al., Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes lgorithm, IOP Conf. Ser.: Mater. Sci. Eng. 1088 (1) (2021) 012045.
[88] T. Pano, R. Kashef, A complete VADER-based sentiment analysis of bitcoin (BTC) tweets during the era of COVID-19, Big Data Cogn. Comput. 4 (4) (2020) 33.
[89] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[90] Y. Zhang, Support vector machine classification algorithm and its application, in: C. Liu, L. Wang, A. Yang (Eds.), Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14–16, 2012. Proceedings, Part II, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 179–186.
[91] M.W. Gardner, S. Dorling, Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences, Atmos. Environ. 32 (14–15) (1998) 2627–2636.
[92] G.I. Webb, E. Keogh, R. Miikkulainen, Naïve bayes, Encyclopedia Mach. Learn. 15 (2010) 713–714.
[93] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[94] K. Zarzycki, M. Ławryńczuk, LSTM and GRU neural networks as models of dynamical processes used in predictive control: A comparison of models developed for two chemical reactors, Sensors 21 (16) (2021) 5625.
[95] W. Lowe, K. Benoit, S. Mikhaylov, M. Laver, Scaling policy preferences from coded political texts, Legislative Stud. Q. 36 (1) (2011) 123–155.
[96] N. Poerner, U. Waltinger, H. Schütze, Inexpensive domain adaptation of pretrained language models: Case studies on biomedical NER and covid-19 QA, 2020, arXiv preprint arXiv:2004.03354.
[97] Y. Wang, F. Liu, K. Verspoor, T. Baldwin, Evaluating the utility of model configurations and data augmentation on clinical semantic textual similarity, in: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, 2020, pp. 105–111.
[98] W. Yoon, J. Lee, D. Kim, M. Jeong, J. Kang, Pre-trained language model for biomedical question answering, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2019, pp. 727–740.
[99] Y. Peng, Q. Chen, Z. Lu, An empirical study of multi-task learning on BERT for biomedical text mining, 2020, arXiv preprint arXiv:2005.02799.
[100] Z.-H. Jiang, W. Yu, D. Zhou, Y. Chen, J. Feng, S. Yan, Convbert: Improving bert with span-based dynamic convolution, Adv. Neural Inf. Process. Syst. 33 (2020) 12837–12848.
[101] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, 2020, arXiv preprint arXiv:2006.03654.
[102] J.R. Minot, N. Cheney, M. Maier, D.C. Elbers, C.M. Danforth, P.S. Dodds, Interpretable bias mitigation for textual data: Reducing gender bias in patient notes while maintaining classification performance, 2021, arXiv preprint arXiv:2103.05841.
[103] P.P. Liang, C. Wu, L.-P. Morency, R. Salakhutdinov, Towards understanding and mitigating social biases in language models, in: International Conference on Machine Learning, PMLR, 2021, pp. 6565–6576.