Original Paper

Targeting COVID-19 and Human Resources for Health News Information Extraction: Algorithm Development and Validation

Mathieu Ravaut[1], MSc; Ruochen Zhao[1], MSc; Duy Phung[1], MSc; Vicky Mengqi Qin[1], PhD; Dusan Milovanovic[2], BSc; Anita Pienkowska[1], PhD; Iva Bojic[1], PhD; Josip Car[3], PhD; Shafiq Joty[1,4], PhD

[1]Nanyang Technological University, Singapore, Singapore

[2]Episteme Systems, Geneva, Switzerland

[3]King's College London, London, United Kingdom

[4]Salesforce Research, San Francisco, CA, United States

Corresponding Author:

Mathieu Ravaut, MSc
Nanyang Technological University
50 Nanyang Avenue
Singapore, 639798
Singapore
Phone: 65 88947729
Email: mathieuj001@e.ntu.edu.sg

Abstract

Background: Global pandemics like COVID-19 put a high amount of strain on health care systems and health workers worldwide. These crises generate a vast amount of news information published online across the globe. This extensive corpus of articles has the potential to provide valuable insights into the nature of ongoing events and guide interventions and policies. However, the sheer volume of information is beyond the capacity of human experts to process and analyze effectively.

Objective: The aim of this study was to explore how natural language processing (NLP) can be leveraged to build a system that allows for quick analysis of a high volume of news articles. Along with this, the objective was to create a workflow comprising human-computer symbiosis to derive valuable insights to support health workforce strategic policy dialogue, advocacy, and decision-making.

Methods: We conducted a review of open-source news coverage from January 2020 to June 2022 on COVID-19 and its impacts on the health workforce from the World Health Organization (WHO) Epidemic Intelligence from Open Sources (EIOS) by synergizing NLP models, including classification and extractive summarization, and human-generated analyses. Our DeepCovid system was trained on 2.8 million news articles in English from more than 3000 internet sources across hundreds of jurisdictions.

Results: Rules-based classification with hand-designed rules narrowed the data set to 8508 articles with high relevancy confirmed in the human-led evaluation. DeepCovid's automated information targeting component reached a very strong binary classification performance of 98.98 for the area under the receiver operating characteristic curve (ROC-AUC) and 47.21 for the area under the precision recall curve (PR-AUC). Its information extraction component attained good performance in automatic extractive summarization with a mean Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score of 47.76. DeepCovid's final summaries were used by human experts to write reports on the COVID-19 pandemic.

Conclusions: It is feasible to synergize high-performing NLP models and human-generated analyses to benefit open-source health workforce intelligence. The DeepCovid approach can contribute to an agile and timely global view, providing complementary information to scientific literature.

(JMIR AI 2024;3:e55059) doi: 10.2196/55059

KEYWORDS

COVID-19; SARS-CoV-2; summary; summarize; news articles; deep learning; classification; summarization; machine learning; extract; extraction; news; media; NLP; natural language processing

Introduction

The unprecedented outbreak and rapid spread of COVID-19 have led to detrimental impacts on almost the whole population worldwide. Early detection of such an outbreak or its impact on the population can help policymakers identify intervention points and set priorities and policies [1,2]. This detection, also called public health surveillance (PHS), is defined as “the continuous, systematic collection, analysis, and interpretation of health-related data needed for the planning, implementation, and evaluation of public health practice” [3,4]. Traditional PHS, which is mostly passively conducted, is often limited by data quality and timeliness, restricting the accurate and quick or even instantaneous identification of outbreaks and subsequent impacts and adoption of an effective intervention [2,5]. PHS has evolved over time as technological advances provide an opportunity for more accurate and timely information collection and analysis [6].

Data-driven artificial intelligence is one of the innovative technologies that can address the limitation of traditional PHS [7]. The open-source textual data from publicly available sources that is of high frequency, high volume, and relatively low effort to collect provide a great potential for the application of natural language processing (NLP), a subset of artificial intelligence, to process and analyze large amounts of natural language data [8,9]. Moreover, deep learning NLP models can be further fine-tuned on a large variety of tasks that could reach performance on par or if not better than humans [10,11]. One of the most popular data sources used for NLP is social media, such as Twitter [12], Facebook [13], Sina Weibo, and Yahoo!, and online forums like Reddit [14], to name a few.

There is a growing volume of literature adopting NLP techniques to extract and analyze social media data for PHS including monitoring public sentiments and health behaviors, predicting a pandemic, and detecting misinformation [1,14-18]. However, there could be potential bias from using social media data due to selected data sets that could overlook underrepresented population groups (generalizability) or contain misinformation (validity) [19-21]. On the other hand, Open Source Intelligence, which includes published and broadcasted news reports, may play a central role in national security, including regarding health emergencies, which often are highly covered. However, such news sources have been less leveraged in the existing models and literature [19,22]. Varol et al [22] published one of the few pieces of literature analyzing news coverage of CNN and the Guardian by using clinical and biomedical NLP models from the Spark NLP for Healthcare library to understand adverse reactions to drugs and vaccines that are used to combat the virus [22].

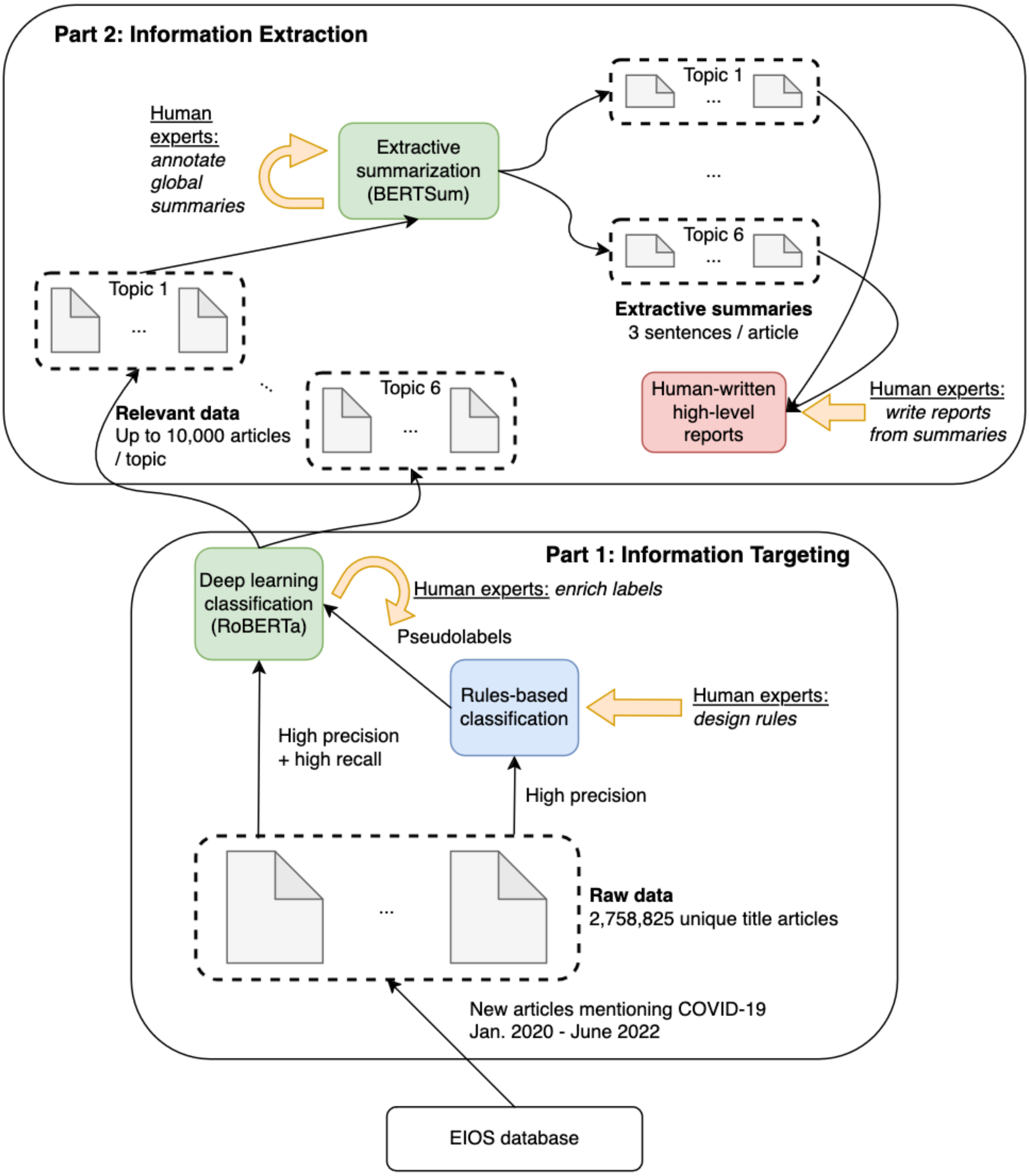
Although most PHS studies applying NLP on open-source data from publicly available sources focused on the population in the community [1,6,23], less is understood about the frontline health workers who are essential for the provision of health care services yet most directly affected by the pandemic. Compared with the general population, health workers were more susceptible to infections due to frequent contact with infected patients [24]. In addition to higher rates of infection and death, health workers also faced challenges from discontinued education or training, financial hardship, and impaired health and wellness due to the pandemic, which could further negatively affect the quality of services and patient outcomes [25]. Hence, it is necessary to have a timely understanding of the different impacts of COVID-19 on health workers in order to construct a targeted intervention.

In this study, we leveraged millions of worldwide news articles in English from publicly available sources collected by the World Health Organization (WHO) Epidemic Intelligence from Open Sources (EIOS) database. We developed an NLP framework named DeepCovid that automatically finds then summarizes relevant articles from EIOS. DeepCovid was designed by a joint team of computer scientists, medical doctors, and population health experts. Beyond the COVID-19 use case, we present a framework that can be leveraged in other PHS applications and support health policymakers with strategic intelligence.

Methods

In this section, we describe each component of DeepCovid. The overall system, which aims to classify, arrange, and reduce a big volume of data comprising news articles, is pictured in Figure 1. The first part of the system aims to find relevant articles, and it trains and applies a deep learning classifier onto the database (information targeting). The resulting news articles move on to the next stage of information extraction, which aims to summarize relevant articles. An extractive summarization model summarizes each article into 3 sentences. Corresponding summaries are then analyzed by human experts to produce reports.

Figure 1. DeepCovid model architecture overview (read from bottom to top), with colored blocks corresponding to machine learning models and gold arrows indicating actions necessitated from human experts. BERT: Bidirectional Encoder Representations from Transformers; EIOS: Epidemic Intelligence from Open Sources; RoBERTa: Robustly Optimized BERT Pretraining Approach.


Part 2: Information Extraction
Human experts: annotate global summaries
Extractive summarization (BERTSum)
Topic 1
...
Topic 6
Extractive summaries
3 sentences / article
Topic 1
Topic 6
Relevant data
Up to 10,000 articles / topic
Human-written high-level reports
Human experts: write reports from summaries
Part 1: Information Targeting
Deep learning classification (RoBERTa)
Human experts: enrich labels
Pseudolabels
Rules-based classification
Human experts: design rules
High precision + high recall
High precision
Raw data
2,758,825 unique title articles
New articles mentioning COVID-19
Jan. 2020 - June 2022
EIOS database

Rules-Based Classification

After preliminary data cleaning, which included deduplication, we built inclusion rules for each of the 6 predetermined topics separately, validating choices through human assessment of precision. The end goal was to narrow the database to a set of relevant articles for each topic of interest: An article was kept if and only if it passed all rules for this topic. Our rules were independent from the lexical tagging already performed within EIOS comprising health care professions and a COVID-19 category.

Rules were designed on both the article title and body. Rules rely on sets of manually identified keywords listed by domain experts and the logical operators OR and AND. Rules can be inclusive, meaning the article is kept if some of the keywords are present, or exclusive, discarding the article if it contains some keywords. There could be multiple such operators nested to form a single rule, such as *one keyword among keywords_list_1 in the body OR one keyword among keywords_list_2 in the body AND two keywords among keywords_list_3 in TITLE*. When working on the article text body, some rules scan for at least one sentence being positive, in which case the entire article is considered to have passed the given rule. We list the set of rules for each topic in Multimedia Appendix 1.

After filtering news articles through rules, we also mapped each article to a unique country among its sets of countries tagged by EIOS. On average, each article has 2.07 such initial country tags from EIOS. Reducing to a single country tag reduces noise and enables the creation of pools of relevant articles per country, which allows further synthesis of key information. When country names are present in the title or first article sentence, we mapped the article to the most frequent such country. Otherwise, we used a deep learning embedding approach. Specifically, we collected all LOC (denoting location) and PERSON (denoting a person's name [eg, Barack Obama]) entities from the *spacy* library [26] in the article body, concatenated them, and encoded them with a Robustly Optimized Bidirectional Encoder Representations from Transformers (BERT) Pretraining Approach (RoBERTa) model [27]. RoBERTa CLS token embeddings then yield a representation with the desired behavior. We also encoded each country name with the same RoBERTa model and returned the country whose representation maximizes the cosine similarity with the article representation.

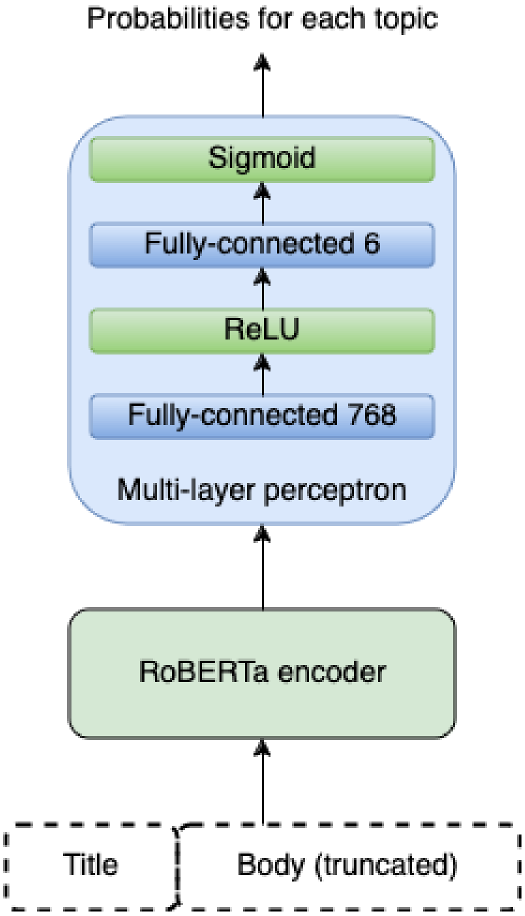
Deep Learning–Based Classification

Architecture

The aforementioned rules-based classification provides a hard assignment for each article to a predetermined topic: Either the article is marked as relevant for at least one topic, or it is not, and it is discarded. There are major limitations to such an algorithm: Articles found as positive may be irrelevant as the presence of key terms does not entice a core focus of the topic of interest (false positives), and many relevant articles might have been missed (false negatives), for instance if no keyword is found within the article. Designing a system that avoids false positives was, however, out of the scope for this study. To tackle false negatives and improve recall, we built a classifier based on a deep learning model: Such models learn a dense vector representation of a news article that can be used for further classification of the article without being limited by the specific choice of words in the text. The classifier assigns to each article a soft probability that it is positive for each of the topics of interest in a multilabel binary classification fashion.

Following the success of large pretraining language models for natural language understanding [27-32], we selected RoBERTa-base [27] as a backbone language model. As input to RoBERTa, we concatenated the title and article body and truncated the resulting string to the maximum input length of 512 tokens. We added 2 fully connected layers, one with a hidden size of 768, followed by rectified linear unit (ReLU) nonlinearity [33], and the last one with a hidden size of 6 (number of topics of interest), followed by Sigmoid as the final output layer. An overview of the classifier architecture is shown in Figure 2.

Figure 2. Deep learning classifier architecture. ReLU: rectified linear unit; RoBERTa: Robustly Optimized BERT Pretraining Approach.



Deep Classification Labels Construction

We trained the deep classifier with the multilabel binary cross-entropy loss using the rules-based classification hard assignments as labels. Due to the very large volume of articles and a low expected fraction of relevant articles, it is unrealistic to collect human annotations for a training set for classification. We made a 90%-10% random training-validation split over the 28-month period from January 2020 to April 2022. The imbalanced nature of the classification problem was challenging: There were a few thousand positives but a few million negatives. Initial labels were provided by rules-based hard assignment. To ensure clean positive labels, volunteer human experts scanned all positive articles and discarded irrelevant ones. From the resulting labeling, in the training set, we kept all positive samples but randomly subsampled 100,000 negative articles. No class rebalancing was performed on the validation set.

After training the first version of the model, we made an inference on the entire 28-month data set and sorted down articles by decreasing the predicted probability for each topic. Human experts were asked to review articles initially flagged as negative but among the top 500 highest predicted scores, which significantly augmented the number of positive labels. After this labeling enrichment, we ended with 6512 positive articles in the training set, leaving 100,000 negative articles (positive ratio: 6.11%). The final validation set consisted of 270,324 articles, including 723 positives and 269,601 negatives (positive ratio: 0.27%). We fine-tuned the deep classifier again with these augmented sets of labels.

In both fine-tuning rounds, we trained for 5 epochs, with a learning rate of 1e^{-5} and the Adam optimizer [34]. We used a batch size of 4 and evaluated the model every 5000 optimization steps. We warmed up the learning rate linearly over the first 5% training steps, then linearly decreased it to 0 in the following 95% steps. We measured performance with the area under the receiver operating characteristic (ROC) of the area under the curve (AUC) metric and performed early stopping, saving a new checkpoint whenever the validation AUC improved.

For inference and real-time use of the system, we kept all articles with a predicted probability either high enough (>0.95) or within 3 times the number of articles flagged as positives by the rules-based model for each topic.

Extractive Summarization

Once the number of relevant articles has been narrowed through article-level classification, the goal of summarization is to give the user a high-level, concise summary of the key information present in the article. Despite recent progress in abstractive summarization, such models are known to be prone to hallucinations [35-37], a problem partly fueled by the fact that commonly used fine-tuning data sets themselves contain hallucinations [35,38]. Given the critical use case for DeepCovid, we decided to use an extractive summarization model [39,40]. In the following sections, we describe how we built 2 sets of extractive summarization labels to fine-tune DeepCovid.

Summarization Labels

Unlike the classification model, the summarization model operates on a manageable volume of news articles. Therefore, we decided to collect human annotations. We asked volunteer graduate students, all fluent English speakers, to label articles among the positives from the rules-based classification. Annotators were asked to highlight between 1 and 3 sentences forming a *global extractive summary* of the article. We obtained annotations for 4062 unique articles, with at least 300 annotations per topic.

To ensure human agreement, we collected labels from 3 different humans for each article for 1 of the topics. Human labels were lists of selected sentences, and we used Fleiss kappa [41] and Gwet AC1 [42] as metrics to measure agreement. The two are complementary, as Gwet AC1 does not account for chance, unlike Fleiss kappa. The Fleiss kappa was 34.23, and for this metric, random agreement stands at 0. The Gwet AC1 was 83.80, with a random agreement of 19.16 in our setup. These values were in line with reported results in extractive summarization research [43], and we concluded that the labelers agreed enough in this task for us to collect a single human annotation per data point. The distribution of sentence positions selected by the human annotators is shown in Multimedia Appendix 2.

On top of these human global summarization labels, we also made use of pseudolabels from the rules-based classification model to obtain topic-focused summarization labels. Indeed, all but Topic 4 rules make use of sentence-level inclusion rules (eg, the article is kept if at least 1 sentence contains 1 of the keywords). We treated such sentences as pseudolabels for extractive summarization and built a set of 7491 pseudolabels.

Summarization Fine-Tuning

We used BERTExt, a state-of-the-art extractive summarization model, as a sentence selection model [44]. Since our data had uppercase and lowercase letters, we used bert-base-cased as the backbone pre-trained BERT model in BERTExt, downloading it from the HuggingFace *transformers* library [45]. To fine-tune jointly for both sets of the aforementioned labels, we doubled the prediction head. This means that the model assigned 2 probabilities to each sentence of the article: 1 to predict if the sentence should be in the global summary and 1 to predict if the sentence should be in the topic-focused pseudosummary. Each prediction head gave us a ranking of sentences, sorted by decreasing predicted probabilities. We also summed both predicted probabilities and sorted sentences by decreasing sum. We output the first 3 sentences of this final ranking as the final predicted summaries. These summaries capture both a flavor of the global sense of the article and a flavor of the topic-specific information contained in the article.

Given the small volume of available labels from each label source, we fine-tuned BERTEx on the CNN-DailyMail (CNN/DM) data set first [46]. CNN/DM is arguably the most widely used data set in both extractive and abstractive summarization [46-48] and comprises more than 300,000 news articles with corresponding human-written highlights (bullet points) serving as abstractive summaries. Following prior work

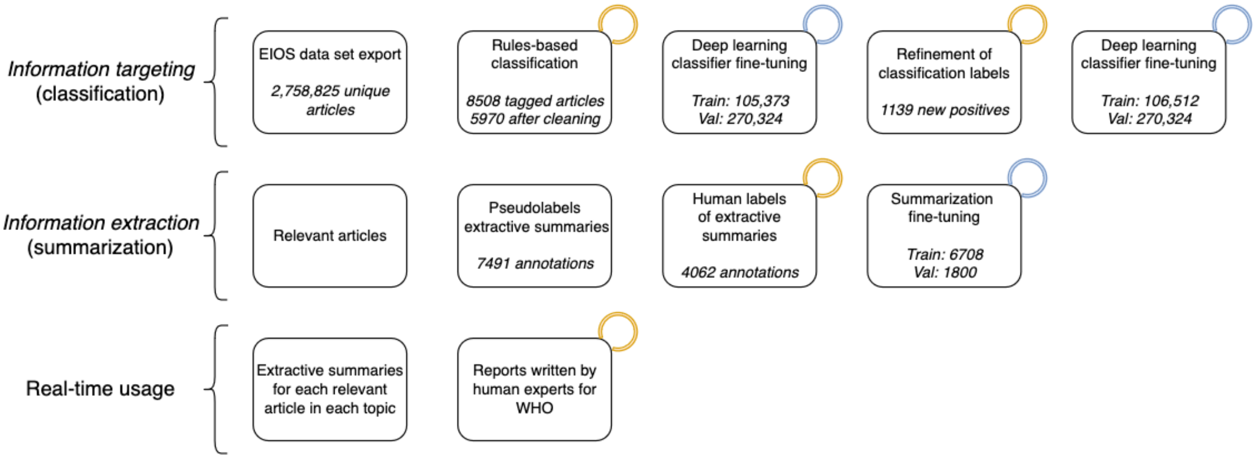
[49], we built extractive summary labels by greedily matching each bullet point summary sentence to the source sentence maximizing Recall-Oriented Understudy for Gisting Evaluation (ROUGE)-1 with it.

We randomly sampled 1800 (300 for each topic) articles to form a validation set, leaving the remaining 6708 articles marked positive by rules-based classification as a training set. We only included articles with human global summarization annotations in the validation set. Since articles in the training set may have lacked either the topic-specific or global summary, we only computed training loss on the available labels, masking out predictions in the case of missing labels. We trained the model for 10 epochs and evaluated the model every epoch. We used the mean of ROUGE-1, ROUGE-2, and ROUGE-L as metrics [50] and performed early stopping. We trained with the Adam optimizer with a learning rate of 1e^{-5} [34]. We warmed up the learning rate linearly over the first 10% training steps, then linearly decreased it to 0 in the following 90% steps. The same optimization procedure was used when performing the initial fine-tuning on the CNN/DM data set.

System Flow

In Figure 3, we show the interplay of each component of our DeepCovid system, with the corresponding data subset size. Our system automatically narrows down the raw data of 2.8 million articles to topic-focused short summaries of highly relevant articles.

Figure 3. Flowchart for DeepCovid showing the step-by-step process transforming a raw data set of 2.8 million news articles (top left) to high-level reports (bottom-right). Boxes with an orange top-right ring indicate the need for human annotation, while boxes with a blue ring correspond to training a deep learning model. EIOS: Epidemic Intelligence from Open Sources; Val: validation; WHO: World Health Organization.


Information targeting (classification)
EIOS data set export
2,758,825 unique articles
Rules-based classification
8508 tagged articles
5970 after cleaning
Deep learning classifier fine-tuning
Train: 105,373
Val: 270,324
Refinement of classification labels
1139 new positives
Deep learning classifier fine-tuning
Train: 106,512
Val: 270,324
Information extraction (summarization)
Relevant articles
Pseudolabels extractive summaries
7491 annotations
Human labels of extractive summaries
4062 annotations
Summarization fine-tuning
Train: 6708
Val: 1800
Real-time usage
Extractive summaries for each relevant article in each topic
Reports written by human experts for WHO

Results

Data Set

We used data from the EIOS database ranging from January 1, 2020, to June 30, 2022 (totaling 30 months). EIOS tracks news articles on the web from more than 12,000 publicly available news outlets in more than 200 countries and territories. Data were filtered for the English language and with keywords relevant to the health workforce (Multimedia Appendix 3). Each article in the resulting data set was tagged by EIOS in-house lexical classification patterns with at least 1 matching keyword (there could be more). Verification using the *langid* package confirmed that more than 99.8% of the articles were indeed in English [51]. The initial data set contained 3,235,657 news articles from 3472 different unique sources and tagged with 243 different locations. After removing duplicate articles based on the title, our final working data set contained 2,758,825 unique news articles. Further statistics on the working data set can be found in Multimedia Appendix 4.

Information Targeting Through Article-Level Classification

The information targeting component of DeepCovid serves the purpose of reducing noise in the data set to narrow it down to only the relevant articles for each of the 6 topics of interest from WHO. Namely, these topics of interest are (1) policy regarding management of and investments in the health workforce, (2) education of health workers, (3) vaccination of health workers, (4) strikes and industrial actions by health workers, (5) mental health issues of health workers, and (6) health worker infections and deaths.

We first created a rules-based classification, and the outputs were used to train the deep learning–based classification component of DeepCovid. Rules are lexical matches, with inclusion and exclusion criteria, and are defined at both the title level and article body level. The detailed list of rules for each topic can be found in Multimedia Appendix 1. This rules-based classification component was built to improve the precision of EIOS-retrieved articles and reduce the volume of irrelevant articles. We assessed the performance of the rules-based classification using human evaluation. Among articles marked as positive by the rules-based system, we subsampled 50 articles randomly for each topic and asked a human domain expert to label them as relevant with regards to the topic or not. Three human experts volunteered, and each human rater was assigned 2 different topics.

Rules-based classification number of positives (N) per topic, relevancy rate (precision), and overlap between topics are shown in Table 1. Overall, the rules-based classification identified a very small fraction of articles (8508 in total, 0.053% on average across the 6 topics) with a high fraction of them (258/300, 86%) being marked relevant by humans, proving its high precision. However, we highlight that this high precision is achieved after 2 rounds of article selection through lexical rules (the rules in EIOS and the subsequent proposed rules by us), and it is therefore not the "true" precision that would be achieved on a large random sample of articles crawled from the web. We also acknowledge the inherent subjectivity in human assessment of relevancy, and judgments may vary from one human to another [52,53]. Besides, as seen in the confusion matrix, the overlap between topics is small: for instance, of 1125 articles identified for Topic 1, 9 (0.8%) of them also belong to Topic 2.

Table 1. Rules-based classifier for the 28 million-article data set from January 2020 to April 2022.

Topic	Positive rate, n (%)[a]	Relevant, n (%)[b]	Overlap between topics[c], n					
			Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
Topic 1	1125 (0.041)	45 (90)	—[d]	9	0	5	9	1
Topic 2	1706 (0.062)	45 (90)	9	—	3	6	21	5
Topic 3	2077 (0.075)	44 (88)	0	3	—	14	22	81
Topic 4	1102 (0.040)	44 (88)	5	6	14	—	56	4
Topic 5	1444 (0.052)	36 (72)	9	21	22	56	—	41
Topic 6	1331 (0.048)	44 (88)	1	5	81	4	41	—

[a] Marked relevant by the rules-based system; overall mean: 1464/28,000,000, 0.053%.

[b] The articles tagged by rules (among a random sample of 50) that were confirmed as relevant to the topic by human experts; overall mean: 43/50, 86%.

[c] Subset of articles that also belong to the topic listed in the column.

[d] Not applicable.

By construction, the rules-based classification identified a high *precision* subset of news articles (86% relevancy rate). However, it has no mechanism to ensure high *recall*, which is one of the motivations behind subsequently training the deep classifier. After training the first version of the deep classifier (tagged as the "initial model"), we made an inference on the entire data set and asked human evaluators to examine the articles that did not past the rules but among the top 500 highest predictions (the "Relevant" column). This corresponds to articles initially missed by the rules yet flagged as extremely relevant by the deep learning model. Such a relabeling process enabled us to enrich the rules-based labels with human annotations while avoiding a human inspection of 2.8 million news articles. Human annotation for this phase was done with the same volunteers as in the previous phase. We then trained the deep classifier again (tagged as the "final model") with the cleaner labels and evaluated it with the AUC. To understand what relative ranking the deep classifier assigned to articles marked positive by rules, we also report the Precision@k.N and Recall@k.N, where N is the number of articles marked positive by the rules-based process and k is an integer (eg, 1, 2, or 10). Table 2 reports the relevance and performance of the results.

The deep learning classifier achieved a consistent and very high AUC across topics (88.54 on average), attesting to both the strength of the signal singled out by the rules and the capacity of the deep classifier to accurately learn it. Indeed, if human-curated lexical rules were poorly designed, a high-capacity pretrained language model would struggle to capture their topic and linguistic style such as words, word patterns, and phrases. The high percentage of relevant negatives among high prediction scores also shows promising capacity in the model to ensure higher recall.

Table 2. Deep learning classifier performance on the classification validation set.

Topic	Relevance of the initial model		Performance of the final model					
	High-score negatives[a], n	Relevant, n (%)[b]	ROC-AUC[c]	PR-AUC[d]	Prec@[e] 2N	Prec@ 10N	Rec@[f] 2N	Rec@ 10N
Topic 1	309	255 (82.5)	99.33	22.77	20.00	5.87	40.00	58.67
Topic 2	299	234 (78.3)	94.87	24.55	19.64	5.79	39.29	57.86
Topic 3	272	270 (99.3)	99.92	42.75	36.56	9.78	73.12	97.85
Topic 4	151	150 (99.3)	99.98	84.62	48.75	10.00	97.50	100.00
Topic 5	220	161 (73.2)	99.86	39.65	29.85	8.06	59.70	80.60
Topic 6	230	227 (98.7)	99.93	68.72	42.13	9.81	84.26	98.15
All topics, mean	216.17	168.93 (88.5)	98.98	47.21	32.82	8.22	65.65	82.19

[a] Articles initially missed by the lexical rules (negatives) but were among the top 500 highest predicted score by the deep learning model.

[b] High-score negatives identified as relevant to the topic by human experts.

[c] ROC-AUC: area under the receiver operating characteristic curve.

[d] PR-AUC: area under the precision recall curve.

[e] Prec@: precision scores at different thresholds.

[f] Rec@: recall scores at different thresholds.

Information Extraction With Extractive Summarization

The subsequent module of DeepCovid tackled information extraction, which identified the key takeaways among articles previously marked as relevant. We evaluated the summarization performance with the standard ROUGE metric [50], averaging its 3 commonly used versions ROUGE-1/2/L. We report the mean ROUGE achieved by the extractive summarization component of DeepCovid on each topic in Table 3, alongside ablated versions with which the model had access to less training supervision. We experimented with sentences selected by lexical rules ("selected sentences"), sentences annotated by humans ("human sentences"), and fine-tuning on the CNN/DM news summarization data set.

CNN/DM means that the model was fine-tuned on the news summarization benchmark CNN/DM first [46]. *Selected* refers to the model being fine-tuned with sentences flagged by the rules-based classification as labels (conveying a pseudosummary focused to each topic), and *human* refers to fine-tuning with human annotations that were designed to build a global summary. In practice, we used the model fine-tuned with all 3 options (denoted as the *final model*), even though it reached slightly less performance than *CNN/DM + human*, as we found its predicted summaries were more focused toward the topics of interest.

Table 3. Extractive summarization Recall-Oriented Understudy for Gisting Evaluation (ROUGE) results (mean of ROUGE-1, ROUGE-2, and ROUGE-L).

Model supervision[a]	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Mean
None (random model weights)	21.90	19.76	27.65	25.37	25.35	25.34	24.23
Selected sentences	45.41	37.27	49.96	50.77	47.20	37.79	44.73
Human sentences	47.61	38.22	48.67	49.58	48.71	37.96	45.13
Selected + human sentences	52.15	38.49	47.19	47.78	51.85	42.00	46.58
CNN/DM[b]	45.33	36.80	45.86	48.57	48.19	35.81	43.43
CNN/DM + selected sentences	44.71	28.69	44.65	40.08	44.82	35.55	39.75
CNN/DM + human sentences	53.97	43.91	49.75	49.08	55.61	43.08	49.23
CNN/DM + selected sentences + human sentences (final model)	53.76	40.74	49.08	47.55	52.45	42.99	47.76

[a] Signal with which the extractive summarization model was trained.

[b] CNN-DM: CNN-DailyMail.

Discussion

Main Findings

The challenge posed by the pandemic offered an opportunity to improve PHS through the use of innovative NLP techniques [7]. Our newly developed framework DeepCovid has demonstrated how to semi-automatically extract precise, targeted news information on health workers concerning the COVID-19 pandemic. Leveraging a global, million-article scale news article database, this framework is able to provide global and population-level information on how COVID-19 impacts health workers that traditional methods (eg, survey, media monitoring) may not be able to do in a short time. With a generic and reusable method that can deal with a high volume of news articles published worldwide, DeepCovid can be used for any health care–related events such as a future similar pandemic and potentially be extended for events beyond the scope of health care, such as financial crises. The DeepCovid framework can assist policymakers with providing fast responses to future similar public health concerns.

Putting DeepCovid in place only requires 4 human actions (see Figure 1): (1) design classification rules to narrow the data set to relevant articles, (2) relabel (some of) the resulting positive and negative articles, (3) label a small set of global extractive summaries to seed the summarization model, and (4) finally, aggregate extractive summaries into reports. All 4 steps require a moderate volume of work from human workers, on the order of a few hours to a few days from 2 humans, which is several orders of magnitude lower than what would be required to manually go through such a scale of data as that which we applied the system, proving the efficiency of DeepCovid. Furthermore, a simpler version of DeepCovid bypassing human actions (2) and (3) leads to a system with reasonable performance, as the key human interventions are the initial (1) and final (4) ones.

Existing work using machine learning to address system-level challenges arising from the COVID-19 pandemic does not cover the multiple impacts of the pandemic on the worldwide health workforce. We note one study that predicted the mental health of Chinese medical workers with logistic regression analysis [54]. In the realm of NLP applications, another study predicted sentiment from tweets by Indian citizens using BERT to assess public opinion during a lockdown [55]. Another paper leveraged long short-term memory networks to predict the number of deaths from WHO data in 3 countries [56]. The most relevant system to ours is CO-Search [57], which is a multicomponent deep learning pipeline enabling the user to find relevant documents with regards to a query; answer questions; and summarize them, leveraging scientific publications from the CORD-19 challenge [58]. However, the CO-Search input data are wildly different from the news data in our study.

With a data set the scale of EIOS, topic-specific precision and recall evaluation remain an open research issue. We showed that DeepCovid rules-based classification may reach high precision through human evaluation, but this is at the cost of 2 rounds of lexical filtering (EIOS and DeepCovid), and human precision evaluation itself is not perfect due to the subjectivity among raters. DeepCovid proposes a mechanism to boost recall of relevant articles through deep learning, yet “true” recall remains impossible to measure as it would involve an extremely costly human inspection of 2.8 million articles. Language models like the ones used in DeepCovid are not equipped with semantical understanding of what classification rules are designed to capture and merely rely on statistical co-occurrence patterns, which enables relevant articles to be expanded with other articles containing similar topics albeit phrased with different lexicality. Striving for perfect precision and recall may need other, complementary tools to deep pretrained language models, such as knowledge graphs.

With regards to optimization, DeepCovid's double objective makes training in a single phase complicated. Document classification and extractive summarization are different types of tasks, and reducing them to a single model addressing both might compromise performance, motivating our choice to keep separate modules, each proven to be a leading approach, even though this adds some complexity and requires 2 separate training processes. Another limitation of our work lies in the fact that human intervention remains compulsory at steps (1) and (4).

Recent progress in large language models (LLMs), sometimes referred to as foundation models, such as GPT-3 [59] or GPT-4 [60], opens a new perspective. Since these models can perform many complicated tasks in few-shot in-context learning [61], or even zero-shot, including summarization [62], we believe that they hold great promise for automating the final step (4) and could synthesize and combine insights from the set of extractive summaries, even more so by decomposing report writing into a template of specific instructions, which has been shown to dramatically boost performance of these models [63]. Acting as agents, LLMs can work hand-in-hand with human experts to create new annotations in cases where annotations are scarce [64], which in turn can be successfully used to fine-tune smaller language models. However, we highlight that LLMs are not a silver bullet since they are hidden behind a paywall and may hallucinate subtly, generating false content that only seasoned domain experts would spot at first glance [65]. We leave the evaluation of the performance of LLMs to better streamline DeepCovid to future work. Emergent capabilities of LLMs such as reasoning [63,66,67] may also be explored for information targeting: from a classification perspective in order to build classifiers (potentially bypassing the construction of lexical rules) and for evaluation of classification precision.

Limitations

The findings of this study should be interpreted in light of several limitations. Although DeepCovid can be a useful tool to extract information from open-source data and assist policymakers during the process of policymaking, it should not be the sole tool for decision-making. What is more important and essential to fight future similar emerging diseases is cross-jurisdictional and cross-functional coordination and collaboration [21].

First, our study was restricted to English-only news articles. This decision was based on the abundance of English sources compared with other languages. From the perspective of the data source, a model that is trained on English-only news articles is likely to miss information from non-English reported news, resulting in biased samples and underestimating the pandemic impact on underrepresented groups. Technically, a multilingual version of DeepCovid is very feasible. It would involve replacing each deep learning component with a multilingual model version (eg, mBERT instead of BERT for the information targeting encoder), which we leave to future work. With model improvement that is compatible with more languages and modalities, DeepCovid will better provide representative information for the global population.

Another limitation lies in the need for expert annotations to bootstrap fine-tuning for each component. This is time-consuming but critical for the final system performance. We envision that new capabilities of LLMs would in the future enable us to replace human annotators with LLM-generated annotations instead, particularly with powerful LLMs such as GPT-4. However, although annotation time would be reduced, using the GPT-4 application programming interface still bears a significant cost. Besides, annotations generated by LLMs would still need to be validated by human experts.

Although broad and valuable, the data set contains a relatively narrow type of news coverage, hence additional insights could be gained by expanding the data sources to social media channels and broadening the format to multimedia content such as videos. The data and findings are impacted by specific strategies for open-source collection that can manifest with, for example, underrepresentation of some countries. Additionally, the current work has not included the identification and exclusion of fake news or reporting biases. Further improvement focusing on bias removal techniques will be needed in order to remove bias from the training data inherited by DeepCovid.

Last, we highlight that DeepCovid synthesizes post hoc information, as news articles usually cover recent (yet, past) events. Findings from DeepCovid may be most useful if acted on early and may be of little use to predict future events.

Conclusion

In this study, we introduced the DeepCovid system. Relying on 2 deep learning–powered components, DeepCovid automatically finds topic-focused relevant news articles among millions of candidates before writing succinct extractive summaries from them. We validated the performance of each component through both human evaluation and automatic metrics, confirming the high performance of the system: Information targeting can reach an AUC in the 98-99 range, and information extraction has an average ROUGE score of 47-48. Core elements of DeepCovid were successfully used to power the Workforce Intelligence from Open Sources project commissioned by the WHO. The findings are to be published in a separate paper. DeepCovid methodology also makes it suitable for use cases other than COVID-19, for instance global events with large news coverage from open sources.

Acknowledgments

This research was funded by the World Health Organization (WHO) courtesy of COVID-19 solidarity grants provided by the governments of Canada and France. This research is also supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-PhD/2021-01-001[T]).

We thank Onyema Ajuebor, Mathieu Boniol, Johannes Schnitzler, Giorgio Cometto, Thijs Cornelissen, and Alastair Donachie from the WHO for valuable insights on the data and feedback on the paper.

Conflicts of Interest

MR and RZ are PhD candidates at Nanyang Technological University (NTU). IB, DP, VMQ, and AP are full-time employees of NTU. DM was a full-time employee at the World Health Organization (WHO) Health Emergency Intelligence and Surveillance Systems (WSE) division, in the Intelligence and Surveillance Systems (ISY) department, in the Intelligence Innovation and Integration (III) unit during the time of the project. JC was a full-time employee at NTU during the time of the project. SJ was a full-time employee at NTU and part-time employee at Salesforce during the time of the project.

Multimedia Appendix 1

Rules-based classification.
[DOCX File , 19 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Extractive summarization labels.
[DOCX File , 40 KB-Multimedia Appendix 2]

Multimedia Appendix 3

EIOS classification.
[DOCX File , 15 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Data set statistics.
[DOCX File , 393 KB-Multimedia Appendix 4]

References

1. Pilipiec P, Samsten I, Bota A. Surveillance of communicable diseases using social media: A systematic review. PLoS One. 2023;18(2):e0282101. [FREE Full text] [doi: 10.1371/journal.pone.0282101] [Medline: 36827297]
2. Nsubuga P, White ME, Thacker SB, Anderson MA, Blount SB, Broome CV, et al. Public Health Surveillance: A Tool for Targeting and Monitoring Interventions. In: Jamison DT, Breman JG, Measham AR, editors. Disease Control Priorities in Developing Countries. New York, NY. Oxford University Press; 2011.
3. Thacker SB, Berkelman RL. Public health surveillance in the United States. Epidemiol Rev. 1988;10(1):164-190. [doi: 10.1093/oxfordjournals.epirev.a036021] [Medline: 3066626]
4. Narasimhan V, Brown H, Pablos-Mendez A, Adams O, Dussault G, Elzinga G, et al. Responding to the global human resources crisis. Lancet. May 01, 2004;363(9419):1469-1472. [doi: 10.1016/S0140-6736(04)16108-4] [Medline: 15121412]
5. Hope K, Durrheim DN, d'Espaignet ET, Dalton C. Syndromic Surveillance: is it a useful tool for local outbreak detection? J Epidemiol Community Health. May 2006;60(5):374-375. [FREE Full text] [doi: 10.1136/jech.2005.035337] [Medline: 16680907]
6. Tsao S-F, Chen H, Tisseverasinghe T, Yang Y, Li L, Butt ZA. What social media told us in the time of COVID-19: a scoping review. Lancet Digit Health. Mar 2021;3(3):e175-e194. [FREE Full text] [doi: 10.1016/S2589-7500(20)30315-0] [Medline: 33518503]
7. Budd J, Miller BS, Manning EM, Lampos V, Zhuang M, Edelstein M, et al. Digital technologies in the public-health response to COVID-19. Nat Med. Aug 2020;26(8):1183-1192. [doi: 10.1038/s41591-020-1011-4] [Medline: 32770165]
8. Al-Garadi MA, Yang YC, Sarker A. The role of natural language processing during the COVID-19 pandemic: health applications, opportunities, and challenges. Healthcare (Basel). Nov 12, 2022;10(11):2270. [FREE Full text] [doi: 10.3390/healthcare10112270] [Medline: 36421593]
9. Hall K, Chang V, Jayne C. A review on natural language processing models for COVID-19 research. Healthc Anal (N Y). Nov 2022;2:100078. [FREE Full text] [doi: 10.1016/j.health.2022.100078] [Medline: 37520621]
10. He P, Liu X, Gao J, Chen W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv. Preprint posted online on October 6, 2021. [doi: 10.48550/arXiv.2006.03654]
11. He P, Gao J, Chen W. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. arXiv. Preprint posted online on March 24, 2023. [doi: 10.48550/arXiv.2111.09543]
12. Weng J, Lim EP, Jiang J, He Q. TwitterRank: finding topic-sensitive influential twitterers. 2010. Presented at: WSDM '10: third ACM international conference on Web search and data mining; February 4-6, 2010; New York, NY. [doi: 10.1145/1718487.1718520]
13. Ortigosa A, Martín JM, Carro RM. Sentiment analysis in Facebook and its application to e-learning. Computers in Human Behavior. 2014;31:527-541. [doi: 10.1016/j.chb.2013.05.024]
14. Low DM, Rumker L, Talkar T, Torous J, Cecchi G, Ghosh SS. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-19: observational study. J Med Internet Res. Oct 12, 2020;22(10):e22635. [FREE Full text] [doi: 10.2196/22635] [Medline: 32936777]
15. Liu Y, Whitfield C, Zhang T, Hauser A, Reynolds T, Anwar M. Monitoring COVID-19 pandemic through the lens of social media using natural language processing and machine learning. Health Inf Sci Syst. Dec 2021;9(1):25. [FREE Full text] [doi: 10.1007/s13755-021-00158-4] [Medline: 34188896]
16. Patel R, Smeraldi F, Abdollahyan M, Irving J, Bessant C. Analysis of mental and physical disorders associated with COVID-19 in online health forums: a natural language processing study. BMJ Open. Nov 05, 2021;11(11):e056601. [FREE Full text] [doi: 10.1136/bmjopen-2021-056601] [Medline: 34740937]
17. Marshall C, Lanyi K, Green R, Wilkins GC, Pearson F, Craig D. Using natural language processing to explore mental health insights from UK tweets during the COVID-19 pandemic: infodemiology study. JMIR Infodemiology. 2022;2(1):e32449. [FREE Full text] [doi: 10.2196/32449] [Medline: 36406146]
18. Evans SL, Jones R, Alkan E, Sichman JS, Haque A, de Oliveira FBS, et al. The emotional impact of COVID-19 news reporting: a longitudinal study using natural language processing. Human Behavior and Emerging Technologies. 2023;19:1-14. [doi: 10.1155/2023/7283166]
19. Zhao Y, He X, Feng Z, Bost S, Prosperi M, Wu Y, et al. Biases in using social media data for public health surveillance: A scoping review. Int J Med Inform. Aug 2022;164:104804. [doi: 10.1016/j.ijmedinf.2022.104804] [Medline: 35644051]
20. Aiello AE, Renson A, Zivich PN. Social media- and internet-based disease surveillance for public health. Annu Rev Public Health. Apr 02, 2020;41:101-118. [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094402] [Medline: 31905322]
21. Brownstein JS, Rader B, Astley CM, Tian H. Advances in artificial intelligence for infectious-disease surveillance. N Engl J Med. Apr 27, 2023;388(17):1597-1607. [doi: 10.1056/NEJMra2119215] [Medline: 37099342]

22. Varol EM, Kocaman V, Haq HU, Talby D. Understanding COVID-19 news coverage using medical NLP. arXiv. Preprint posted online on March 19, 2022. [doi: 10.48550/arXiv.2203.10338]
23. Gupta A, Katarya R. Social media based surveillance systems for healthcare using machine learning: A systematic review. J Biomed Inform. Aug 2020;108:103500. [FREE Full text] [doi: 10.1016/j.jbi.2020.103500] [Medline: 32622833]
24. Pham QT, Le XTT, Phan TC, Nguyen QN, Ta NKT, Nguyen AN, et al. Impacts of COVID-19 on the life and work of healthcare workers during the nationwide partial lockdown in Vietnam. Front Psychol. 2021;12:563193. [FREE Full text] [doi: 10.3389/fpsyg.2021.563193] [Medline: 34489769]
25. Gupta N, Dhamija S, Patil J, Chaudhari B. Impact of COVID-19 pandemic on healthcare workers. Ind Psychiatry J. Oct 2021;30(Suppl 1):S282-S284. [FREE Full text] [doi: 10.4103/0972-6748.328830] [Medline: 34908710]
26. explosion / spaCy. GitHub. URL: https://github.com/explosion/spaCy [accessed 2017-11-07]
27. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv. Jul 26, 2019. URL: https://arxiv.org/abs/1907.11692 [accessed 2019-07-26]
28. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. Presented at: Conference of the North American Chapter of the Association for Computational Linguistics; June 2-7, 2019; Minneapolis, MN.
29. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research. 2021;21:1-67.
30. Zhang Z, Han X, Liu Z, Jiang X, Sun M, Liu Q. ERNIE: Enhanced Language Representation with Informative Entities. 2019. Presented at: 57th Annual Meeting of the Association for Computational Linguistics; July 28-August 2, 2019; Florence, Italy. [doi: 10.18653/v1/p19-1139]
31. Clark K, Luong TL, Le QV, Manning C. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. 2020. Presented at: International Conference on Learning Representations; April 26-May 1, 2020; Addis Ababa, Ethiopia.
32. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. 2020. Presented at: 58th Annual Meeting of the Association for Computational Linguistics; July 5-10, 2020; Virtual meeting. [doi: 10.18653/v1/2020.acl-main.703]
33. Nair V, Hinton G. Rectified linear units improve restricted boltzmann machines. 2010. Presented at: 27th International Conference on International Conference on Machine Learning; June 21-24, 2010:807-814; Haifa, Israel. [doi: 10.5555/3104322.3104425]
34. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. 2015. Presented at: 3rd International Conference for Learning Representations; May 7-9, 2015; San Diego, CA.
35. Maynez J, Narayan S, Bohnet B, McDonald R. On Faithfulness and Factuality in Abstractive Summarization. 2020. Presented at: 58th Annual Meeting of the Association for Computational Linguistics; July 5-10, 2020; Virtual meeting. [doi: 10.18653/v1/2020.acl-main.173]
36. Kryscinski W, McCann B, Xiong C, Socher R. Evaluating the Factual Consistency of Abstractive Text Summarization. 2020. Presented at: 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); November 16-20, 2020; Virtual meeting. [doi: 10.18653/v1/2020.emnlp-main.750]
37. Goyal T, Durrett G. Evaluating Factuality in Generation with Dependency-level Entailment. 2020. Presented at: Findings of the Association for Computational Linguistics: EMNLP 2020; November 16-20, 2020; Virtual meeting. [doi: 10.18653/v1/2020.findings-emnlp.322]
38. Narayan S, Cohen SB, Lapata M. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. 2018. Presented at: Conference on Empirical Methods in Natural Language Processing; October 31-November 4, 2018; Brussels, Belgium. [doi: 10.18653/v1/d18-1206]
39. Erkan G, Radev DR. LexRank: graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research. 2004:457-479. [doi: 10.1613/jair.1523]
40. Zhong M, Liu P, Chen Y, Wang D, Qiu X, Huang X. Extractive Summarization as Text Matching. 2020. Presented at: 58th Annual Meeting of the Association for Computational Linguistics; July 5-10, 2020; Virtual meeting. [doi: 10.18653/v1/2020.acl-main.552]
41. McHugh M. Interrater reliability: the kappa statistic. Biochem Med. 2012;22(3):276-282. [doi: 10.11613/bm.2012.031]
42. Gwet K. Computing inter - rater reliability and its variance in the presence of high agreement. Brit J Math & Statis. Dec 24, 2010;61(1):29-48. [doi: 10.1348/000711006x126600]
43. Karn SK, Chen F, Chen Y-Y, Waltinger U, Schütze H. Few-Shot Learning of an Interleaved Text Summarization Model by Pretraining with Synthetic Data. 2021. Presented at: Second Workshop on Domain Adaptation for NLP; April 20, 2021; Kyiv, Ukraine.
44. Liu Y, Lapata M. Text Summarization with Pretrained Encoders. 2019. Presented at: Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); November 3-7, 2019; Hong Kong, China. [doi: 10.18653/v1/d19-1387]
45. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-Art Natural Language Processing. 2020. Presented at: Conference on Empirical Methods in Natural Language Processing: System Demonstrations; November 16-20, 2020; Virtual meeting. [doi: 10.18653/v1/2020.emnlp-demos.6]

46. Hermann KM, Kocisky T, Grefenstette E, Espeholt L, Kay W, Suleyman M, et al. Teaching Machines to Read and Comprehend. 2015. Presented at: Advances in Neural Information Processing Systems; December 7-12, 2015; Montreal, Canada.
47. Nallapati R, Zhou B, dos Santos C, Gulçehre C, Xiang B. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. 2016. Presented at: 20th SIGNLL Conference on Computational Natural Language Learning; August 11-16, 2016; Berlin, Germany. [doi: 10.18653/v1/k16-1028]
48. See A, Liu PJ, Manning CD. Get To The Point: Summarization with Pointer-Generator Networks. 2017. Presented at: 55th Annual Meeting of the Association for Computational Linguistics; July 30-August 4, 2017; Vancouver, Canada. [doi: 10.18653/v1/p17-1099]
49. Jia R, Cao Y, Tang H, Fang F, Cao C, Wang S. Neural Extractive Summarization with Hierarchical Attentive Heterogeneous Graph Network. 2020. Presented at: 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); November 16-20, 2020; Virtual meeting. [doi: 10.18653/v1/2020.emnlp-main.295]
50. Lin CY. ROUGE: A Package for Automatic Evaluation of Summaries. 2004. URL: https://aclanthology.org/W04-1013.pdf [accessed 2024-09-26]
51. Lui M, Baldwin T. An off-the-shelf language identification tool. 2012. Presented at: ACL 2012 System Demonstrations; July 10, 2012; Jeju Island, Korea.
52. Leonardelli E, Menini S, Aprosio AP, Guerini M, Tonelli S. Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement. 2021. Presented at: 2021 Conference on Empirical Methods in Natural Language Processing; November 7-11, 2021; Punta Cana, Dominican Republic. [doi: 10.18653/v1/2021.emnlp-main.822]
53. Pandey R, Purohit H, Castillo C, Shalin VL. Modeling and mitigating human annotation errors to design efficient stream processing systems with human-in-the-loop machine learning. International Journal of Human-Computer Studies. 2022;160:102772. [doi: 10.1016/j.ijhcs.2022.102772]
54. Wang X, Li H, Sun C, Zhang X, Wang T, Dong C, et al. Prediction of mental health in medical workers during COVID-19 based on machine learning. Front Public Health. 2021;9:697850. [FREE Full text] [doi: 10.3389/fpubh.2021.697850] [Medline: 34557468]
55. Chintalapudi N, Battineni G, Amenta F. Sentimental analysis of COVID-19 tweets using deep learning models. Infect Dis Rep. Apr 01, 2021;13(2):329-339. [FREE Full text] [doi: 10.3390/idr13020032] [Medline: 33916139]
56. Esteva A, Kale A, Paulus R, Hashimoto K, Yin W, Radev D, et al. COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. NPJ Digit Med. Apr 12, 2021;4(1):68. [FREE Full text] [doi: 10.1038/s41746-021-00437-0] [Medline: 33846532]
57. Wang LL, Lo K, Chandrasekhar Y, Reas R, Yang J, Burdick D, et al. CORD-19: The COVID-19 Open Research Dataset. 2020. Presented at: 1st Workshop on NLP for COVID-19 at ACL 2020; July 5-10, 2020; Virtual meeting.
58. Aldhyani TH, Alrasheed M, Alqarni AA, Alzahrani MY, Alahmadi AH. Deep learning and Holt-trend algorithms for predicting COVID-19 pandemic. Computers, Materials & Continua. 2021;67(2):2141-2160. [FREE Full text] [doi: 10.32604/cmc.2021.014498]
59. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language Models are Few-Shot Learners. 2020. Presented at: Neural Information Processing Systems 33 (NeurIPS 2020); December 6-12, 2020; Virtual meeting.
60. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv. Mar 22, 2023. URL: https://arxiv.org/abs/2303.12712 [accessed 2024-09-26]
61. Liu J, Shen D, Zhang Y, Dolan B, Carin L, Chen W. What Makes Good In-Context Examples for GPT-3? 2022. Presented at: Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures; March 1, 2022; Dublin, Ireland.
62. Goyal T, Li JJ, Durrett G. News Summarization and Evaluation in the Era of GPT-3. arXiv. URL: https://arxiv.org/abs/2209.12356 [accessed 2022-09-26]
63. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. 2022. Presented at: 36th International Conference on Neural Information Processing Systems; November 28-December 9, 2022; New Orleans, LA.
64. Xu C, Sun Q, Zheng K, Geng X, Zhao P, Feng J, et al. Wizardlm: Empowering large language models to follow complex instructions. 2024. Presented at: International Conference on Learning Representations (ICLR) 2024; May 7-11, 2024; Vienna, Austria.
65. Park JC, Arase Y, Hu B, Lu W, Wijaya D, Purwarianti A, et al. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. 2023. Presented at: 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics; November 1-4, 2023; Bali, Indonesia. [doi: 10.18653/v1/2023.ijcnlp-main.45]
66. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. 2022. Presented at: 36th International Conference on Neural Information Processing Systems; November 28-December 9, 2022; New Orleans, LA.
67. Wang X, Wei J, Schuurmans, D, Le Q, Chi E, Narang S, et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models. 2023. Presented at: The Eleventh International Conference on Learning Representations (ICLR); May 1-5, 2023; Kigali, Rwanda.

Abbreviations

AUC: area under the receiver operating characteristic curve
BERT: Bidirectional Encoder Representations from Transformers
CNN-DM: CNN-DailyMail
EIOS: Epidemic Intelligence from Open Sources
LLM: large language model
NLP: natural language processing
PHS: public health surveillance
ReLU: rectified linear unit
RoBERTa: Robustly Optimized BERT Pretraining Approach
ROC: receiver operating characteristic
ROUGE: Recall-Oriented Understudy for Gisting Evaluation
WHO: World Health Organization

Edited by H Liu; submitted 01.12.23; peer-reviewed by J Zheng, W Gan; comments to author 04.01.24; revised version received 17.03.24; accepted 10.07.24; published 30.10.24

Please cite as:
Ravaut M, Zhao R, Phung D, Qin VM, Milovanovic D, Pienkowska A, Bojic I, Car J, Joty S
Targeting COVID-19 and Human Resources for Health News Information Extraction: Algorithm Development and Validation
JMIR AI 2024;3:e55059
URL: https://ai.jmir.org/2024/1/e55059
doi: 10.2196/55059
PMID: