JMIR Publications
The Leading eHealth Publisher

Targeted COVID-19 and Human Resource for Health News Information Extraction with a Multi-Component Deep Learning Framework

Mathieu Ravaut, Ruochen Zhao, Duy Phung, Vicky Mengqi Qin, Dusan Milovanovic, Anita Pienkowska, Iva Bojic, Josip Car, Shafiq Joty

Submitted to: JMIR AI
on: December 01, 2023

Table of Contents

Targeted COVID-19 and Human Resource for Health News Information Extraction with a Multi-Component Deep Learning Framework

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 SG
[2]Episteme Systems Geneva CH
[3]King's College London London GB
[4]Salesforce Research San Francisco US

Corresponding Author:
Mathieu Ravaut MSc
Nanyang Technological University
50 Nanyang Avenue
Singapore
SG

Abstract

Background: Global pandemics like COVID-19 put high strain on healthcare 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 WHO Epidemic Intelligence through 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 3,000 Internet sources across hundreds of jurisdictions.

Results: Rules-based classification with hand-designed rules narrows down the dataset to 8,508 articles with high relevancy confirmed in human-led evaluation. DeepCovid's automated information targeting component reaches a very strong binary classification performance of 98.98 ROC-AUC and 47.21 PR-AUC. Its information extraction component attains a good performance in automatic extractive summarization with 47.76 mean ROUGE score. 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. DeepCovid approach can contribute to an agile and timely global view, providing complementary information to scientific literature.

(JMIR Preprints 01/12/2023:55059)
DOI: https://doi.org/10.2196/preprints.55059

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Targeted COVID-19 and Human Resource for Health News Information Extraction with a Multi-Component Deep Learning Framework

Mathieu Ravaut[1*+], Ruochen Zhao[1*], Duy Phung[1], Vicky Mengqi Qin[1], Dusan Milovanovic[2], Anita Pienkowska[1], Iva Bojic[1], Josip Car[1#], Shafiq Joty[1,3#]

[1] Nanyang Technological University, Singapore
[2] Episteme Systems
[3] Salesforce Research

[*] Indicates joint first authorship contributions
[+] Indicates corresponding author: mathieuj001@e.ntu.edu.sg
[#] Indicates joint last authorship contributions

Abstract

Background:
Global pandemics like COVID-19 put high strain on healthcare 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.

Objectives:
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 WHO Epidemic Intelligence through 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 3,000 Internet sources across hundreds of jurisdictions.

Results:
Rules-based classification with hand-designed rules narrows down the dataset to 8,508 articles with high relevancy confirmed in human-led evaluation. DeepCovid's automated information targeting component reaches a very strong binary classification performance of 98.98 ROC-AUC and 47.21 PR-AUC. Its information extraction component attains a good performance in automatic extractive summarization with 47.76 mean ROUGE score. 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. DeepCovid approach can contribute to an agile and timely global view, providing complementary information to scientific literature.

Keywords

Covid-19 ; Natural language processing ; deep learning ; classification ; summarization

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] The 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 that are mostly passively conducted are often limited by data quality and timeliness, restricting the accurate and quick or even instantaneous identification of outbreaks and subsequent impacts, and adoption of effective intervention.[2,5] PHS has evolved over time as technological advances open the window for more accurate and timely information collection and analysis.[6]

Data-driven Artificial Intelligence (AI) 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 provides a great potential for the application of Natural Language Processing (NLP), a subset of AI, 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 which 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, Yahoo! and online forums like Reddit,[14] to name a few.

There is a growing number 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 datasets that could overlook under-represented population groups (generalizability) or contain misinformation (validity).[19-21] On the other hand, Open Source Intelligence, including published and broadcasted news reports, may play a central role in national security, including regarding health emergencies, which often are highly covered. Yet, such news sources have been less leveraged in the existing models and literature.[19,22] Varol et al. is 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]

While most of the 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 healthcare services, yet most directly affected by the pandemic. Compared to 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, 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 a NLP framework named DeepCovid which 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 which can be leveraged in other PHS applications and support health policy makers with strategic intelligence.

Methods

In this section, we describe each component of DeepCovid. The overall system which aims at classifying, arranging and reducing a big volume of data comprising news articles is pictured in Figure 1.

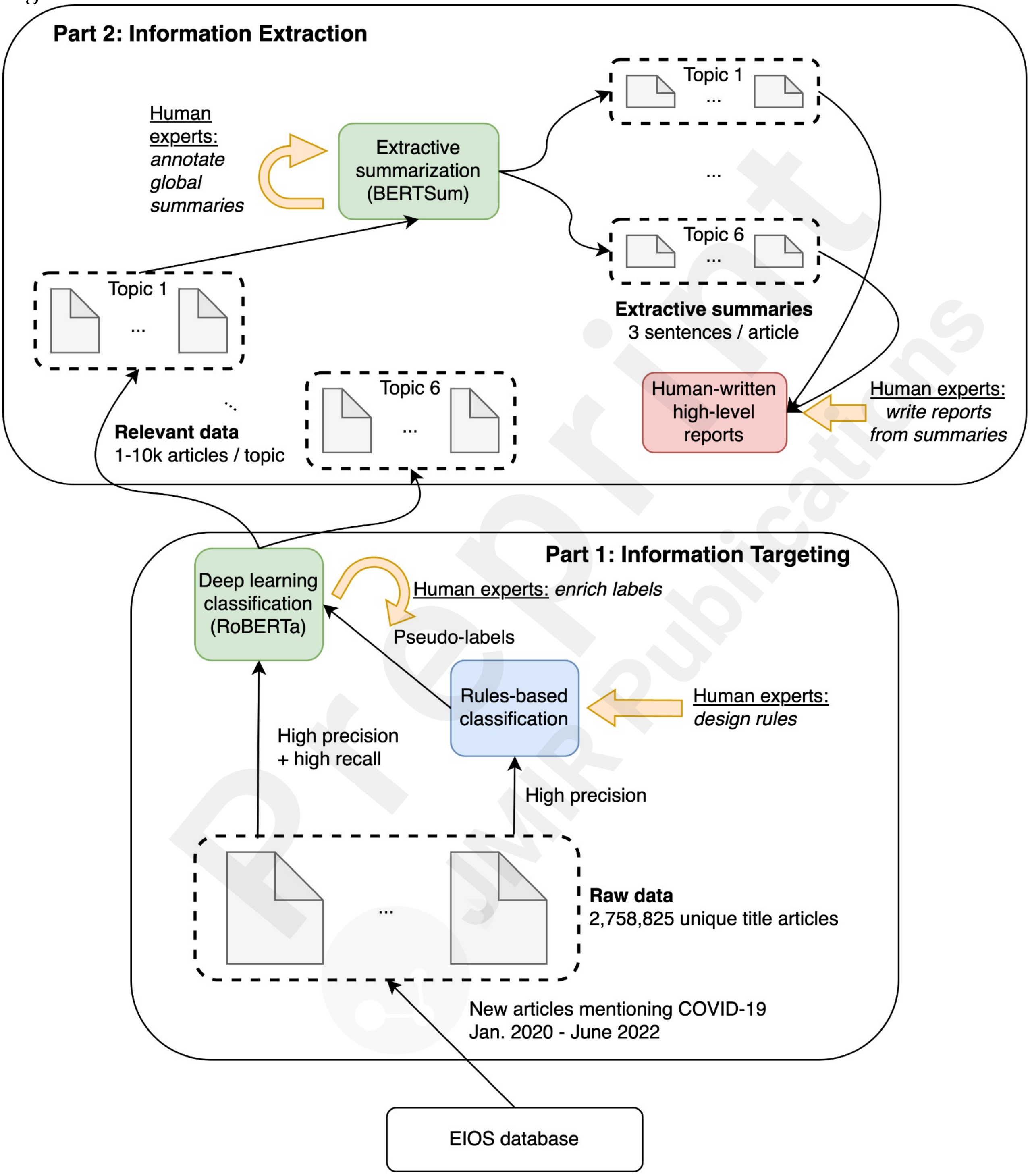

Part 2: Information Extraction
Human experts: annotate global summaries
Extractive summarization (BERTSum)
Topic 1
...
Topic 6
...
Extractive summaries
3 sentences / article
Human-written high-level reports
Human experts: write reports from summaries
Topic 1
...
Relevant data
1-10k articles / topic
Topic 6
...
Part 1: Information Targeting
Deep learning classification (RoBERTa)
Human experts: enrich labels
Pseudo-labels
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

Figure 1. DeepCovid model architecture overview (read from bottom to top). Colored blocks correspond to machine learning models. Golden arrows indicate actions necessitated from human experts. 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 three sentences. Corresponding summaries are then analyzed by human experts to produce reports.

Rules-Based Classification

After preliminary data cleaning which included deduplication, we built inclusion rules for each one of the six predetermined topics separately, validating choices through human assessment of precision. The end goal was to narrow down 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 healthcare professions and COVID-19 category.

Rules were designed on both the title of the article and the body. Rules rely on sets of manually identified keywords listed by domain experts, and logical operators OR and AND. Rules can be inclusive, meaning to keep the article if some 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 the Supplementary Material C.

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 countries tags from EIOS. Reducing to a single country tag reduces noise and enables to create pools of relevant articles per country, which allows to further synthesize key information. When country names are present in the title or first article sentence, we map the article to the most frequent such country. Otherwise, we use a deep learning embeddings approach. Specifically, we collect all ‘LOC’ (denoting location) and ‘PERSON’ (denoting a person’s name, e.g., Barack Obama) entities from the *spacy* library[31] in the article body, concatenate them, and encode them with a *RoBERTa* model.[32] *RoBERTa* CLS token embeddings then yields a representation with the desired behavior. We also encode each country name with the same *RoBERTa* model and return the country whose representation maximizes the cosine similarity with the article representation.

Deep Learning-Based Classification

Architecture

The rules-based classification described above 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 their 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 avoiding 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 which 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 multi-label binary classification fashion.

Following the success of large pre-training language models for natural language understanding,[32-37] we selected *RoBERTa-base*[32] as a backbone language model. As input to *RoBERTa*, we concatenate the title and article body, and truncate the resulting string to the maximum input length of 512 tokens. We added two fully connected layers, one with hidden size 768, followed by *ReLU* non-linearity,[38] and the last one with hidden size 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.

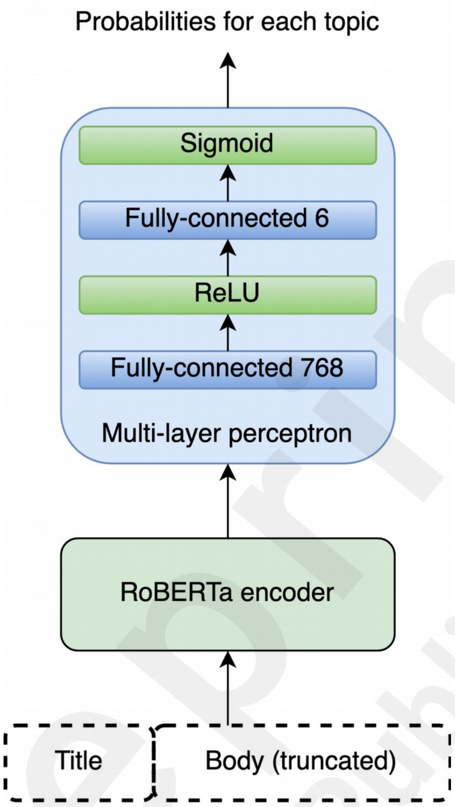


Figure 2. Deep learning classifier architecture.

Deep Classification Labels Construction

We trained the deep classifier with the multi-label 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 labelling, in the training set, we kept all positive samples but subsampled randomly 100,000 negative articles. No class re-balancing was performed on the validation set.

After training the first version of the model, we made inference on the entire 28-month dataset and sorted down articles by decreasing 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 6,512 positive articles in the training set, for 100,000 negative ones (positive ratio: 6.11%). The final validation set was made of 270,324 articles, including 723 positives and 269,601 negatives (positive ratio: 0.27%). We fine-tuned again the deep classifier 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.[39] We used a batch size of 4 and evaluated the model every 5,000 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 Curve (AUC) metric and performed early stopping, saving a new checkpoint whenever the validation AUC improves.

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

Extractive Summarization

Once relevant articles have been narrowed down 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,[40-42] a problem partly fueled by the fact that commonly used fine-tuning datasets themselves contain hallucinations.[40,43] Given the critical use case for DeepCovid, we decided to use an extractive summarization model.[44,45] In the following, we described how we built two 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 rules-based classification. Annotators were asked to highlight between one and three sentences forming a **global extractive summary** of the article. We obtained annotations for 4,062 unique articles, with at least 300 annotations per topic.

To ensure human agreement, we collected labels from three different humans for each article for one of the topics. Human labels were lists of selected sentences, and we used Fleiss Kappa[46] and Gwet AC1[47] as metrics to measure agreement. The two are complementary as Gwet AC1 does not account for chance, unlike Fleiss Kappa. Fleiss Kappa was 34.23, and for this metric random agreement stands at 0. 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,[48] and we concluded that labelers agreed enough in this task for us to collect a single human annotation per data point. The distribution of sentence positions selected by human annotators is shown in Supplementary Material D.

On top of these human global summarization labels, we also made use of pseudo-labels 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 (e.g., the article is kept if at least one sentence contains one of the keywords). We treated such sentences as pseudo-labels for extractive summarization, and built a set of 7,491 pseudo-labels.

Summarization Fine-tuning

We used *BERTExt* as a sentence selection model, a state-of-the-art extractive summarization model.[49] 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 HuggingFace *transformers* library.[50] To fine-tune jointly for both sets of labels described above, we doubled the prediction head. This means that the model assigned two probabilities to each sentence of the article: one to predict if the sentence should be in the global summary, and another one to predict if the sentence should be in the topic-focused pseudo-summary. 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 three sentences of this final ranking as 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* dataset first (CNN/DM).[30] CNN/DM is arguably the most widely used dataset in both extractive and abstractive summarization[30,51,52] and comprises more than 300,000 news articles with corresponding human-written highlights (bullet points) serving as abstractive summaries. Following prior work,[53] we built extractive summary labels by greedily matching each bullet point summary sentence to the source sentence maximizing ROUGE-1 with it.

We randomly sampled 1,800 (300 for each topic) articles to form a validation set, leaving the remaining 6,708 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 lack 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,[29] and performed early stopping. We trained with the Adam optimizer with a learning rate of 1e^{-5}.[39] 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 dataset.

System Flow

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

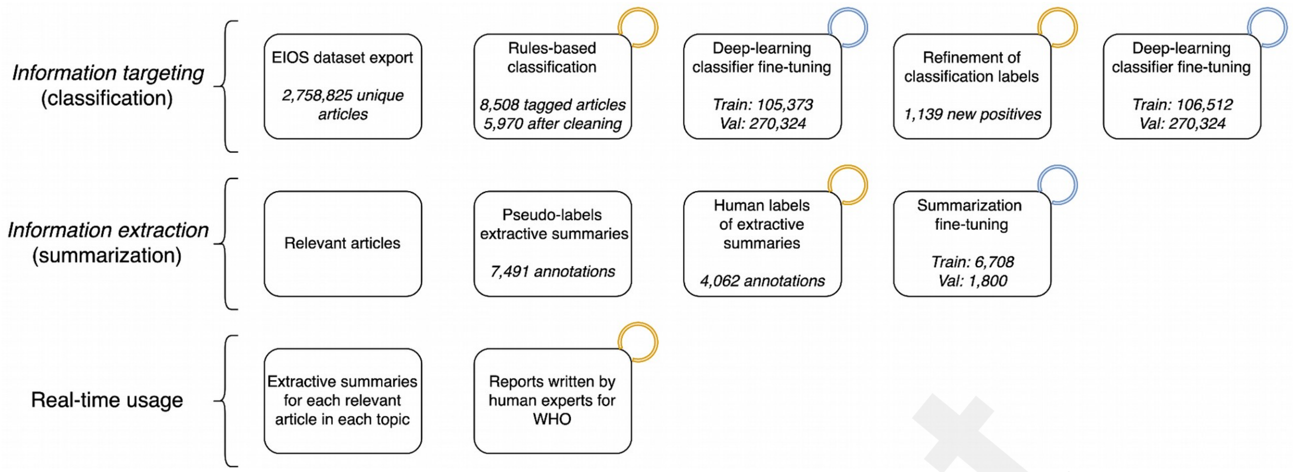

Information targeting
(classification)
EIOS dataset export
2,758,825 unique articles
Rules-based classification
8,508 tagged articles
5,970 after cleaning
Deep-learning classifier fine-tuning
Train: 105,373
Val: 270,324
Refinement of classification labels
1,139 new positives
Deep-learning classifier fine-tuning
Train: 106,512
Val: 270,324
Information extraction
(summarization)
Relevant articles
Pseudo-labels extractive summaries
7,491 annotations
Human labels of extractive summaries
4,062 annotations
Summarization fine-tuning
Train: 6,708
Val: 1,800
Real-time usage
Extractive summaries for each relevant article in each topic
Reports written by human experts for WHO

Figure 3. Flow chart for DeepCovid. We show the step by step process transforming a raw, 2.8M news articles dataset (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.

Results

Dataset

We used data from the EIOS database ranging from January 1st, 2020 to June 30th, 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 was filtered for English language and with keywords relevant to health workforce (See Supplementary Material A). Each article in the resulting dataset was tagged by EIOS in-house lexical classification patterns with at least one matching keyword (there could be more). Verification using the *langid* package confirms that more than 99.8% of articles are indeed in English.[26] The initial dataset contained 3,235,657 news articles, from 3,472 different unique sources, and tagged with 243 different locations. After removing duplicate articles based on the title, our final working dataset contained 2,758,825 unique news articles. Further statistics on the working dataset can be found in Supplementary Material B.

Information Targeting through Article-level Classification

The information targeting component of DeepCovid serves the purpose of reducing noise in the dataset to narrow it down to only the relevant articles, for each of the six topics of interest from WHO. Namely, these topics of interest are 1/ policy, management and investments on 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 workers infections and deaths.

We first created a rules-based classification, which 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 the Supplementary Material C. 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 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 two different topics.

Rules-based classification number of positives (N) per topic, relevancy rate (precision), and overlap between topics are shown in Table 1. Overall, rules-based classification identifies a very small fraction of articles (8,508 in total, 0.053% on average across the six topics) with a high fraction of them (86%) being marked relevant by humans, proving its high precision. However, we highlight that this high precision is achieved after two rounds of articles selection through lexical rules (the ones in EIOS, and our subsequent proposed ones), 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 a human to another.[27,28] Besides, as seen in the confusion matrix, the overlap between topics is small: for instance, out of 1,125 articles identified for Topic 1, 9 or 0.8% of them also belong to Topic 2.

Table 1. Rules-based classifier on the January 2020 - April 2022 timeframe. **N** refers to the number of articles marked relevant by the rules-based system, **%** is the fraction with regards to the entire 2.8M dataset, and **Relevant (%)** indicates the fraction of articles tagged by rules (among a random sample of 50) which were confirmed as relevant to the topic by human experts. Numbers in the “Topic” columns correspond to the subset of articles (among N) which also belong to another topic.

	Positive rate	**Overlap between topics**

Topic	N	%	Relevant (%)	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
Topic 1	1,125	0.041	90		9	0	5	9	1
Topic 2	1,706	0.062	90	9		3	6	21	5
Topic 3	2,077	0.075	88	0	3		14	22	81
Topic 4	1,102	0.040	88	5	6	14		56	4
Topic 5	1,444	0.052	72	9	21	22	56		41
Topic 6	1,331	0.048	88	1	5	81	4	41	
Mean	**1,464**	**0.053**	**86**						

By construction, rules-based classification identifies 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 "**initial model**"), we make inference on the entire dataset, and ask human evaluators to examine articles not passing rules, yet among the top 500 highest predictions (**Relevant (%)** column). This corresponds to articles initially missed by the rules yet flagged as extremely relevant by the deep learning model. Such a re-labeling process enables us to enrich rules-based labels with human annotations, while avoiding a human inspection of 2.8M news articles. Human annotation for this phase is done with the same volunteers as in the previous phase. Then, we train the deep classifier again (tagged as "**final model**") with the cleaner labels and evaluate it with the Area Under the ROC Curve (AUC). To understand what relative ranking the deep classifier is assigning 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 rules-based, and k is an integer (e.g., 1, 2 or 10). Table 2 reports the relevance and performance of the results.

Table 2. Deep learning classifier performance on the classification validation set. **High-score negatives** correspond to articles initially missed by the lexical rules (negatives), yet among the top 500 highest predicted score by the deep learning model. **Relevant (%)** correspond to the fraction of such high-score negatives identified as relevant to the topic by human experts. **AUC** is the Area Under the ROC-Curve, **Prec@k** and **Rec@k** are precision and recall scores at diverse thresholds.

	Relevance - initial model		**Performance - final model**					
Topic	**High-score negatives**	**Relevant (%)**	**ROC-AUC**	**PR-AUC**	**Prec@ 2N**	**Prec@ 10N**	**Rec@ 2N**	**Rec@ 10N**
Topic 1	309	82.52	99.33	22.77	20.00	5.87	40.00	58.67
Topic 2	299	78.26	94.87	24.55	19.64	5.79	39.29	57.86
Topic 3	272	99.26	99.92	42.75	36.56	9.78	73.12	97.85
Topic 4	151	99.34	99.98	84.62	48.75	10.00	97.50	100.00

Topic 5	220	73.18	99.86	39.65	29.85	8.06	59.70	80.60
Topic 6	230	98.70	99.93	68.72	42.13	9.81	84.26	98.15
Mean	**216.17**	**88.54**	**98.98**	**47.21**	**32.82**	**8.22**	**65.65**	**82.19**

The deep learning classifier achieves a consistent and very high AUC across topics (88.54 on average), attesting both the strength of the signal singled out by rules and the capacity of the deep classifier to accurately learn it. Indeed, if human-curated lexical rules were poorly designed, a high-capacity pre-trained 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.

Information Extraction with Extractive Summarization

The subsequent module of DeepCovid tackles information extraction, which identifies the key takeaways among articles previously marked as relevant. We evaluate summarization performance with the standard ROUGE metric,[29] averaging its three 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 has access to less training supervision.

Table 3. Extractive summarization ROUGE results (mean of ROUGE-1/2/L). Model Supervision refers to the signal with which the extractive summarization model was trained. We experiment with sentences selected by lexical rules ("Selected sentences"), sentences annotated by humans ("Human sentences"), and also fine-tuning on the CNN/DM news summarization dataset.

Model supervision	**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	52.15	38.49	47.19	47.78	51.85	42.00	46.58
CNN/DM	45.33	36.80	45.86	48.57	48.19	35.81	43.43
CNN/DM + Selected	44.71	28.69	44.65	40.08	44.82	35.55	39.75
CNN/DM + Human	**53.97**	**43.91**	49.75	49.08	**55.61**	**43.08**	**49.23**
CNN/DM + Selected + Human (final model)	53.76	40.74	49.08	47.55	52.45	42.99	47.76

CNN/DM means that the model was fine-tuned on the news summarization benchmark

CNN/DailyMail first.[30] *Selected* refers to the model being fine-tuned with sentences flagged by rules-based classification as labels (conveying a pseudo-summary focused to each topic), and *Human* refers to fine-tuning with human annotations which were designed to build a global summary. In practice, we use the model fine-tuned with all three options (denoted as the *final model*), even though it reaches slightly less performance than *CNN/DM + Human*, as we found its predicted summaries were more focused towards the topics of interest.

Discussion

The challenge posed by the pandemic offers 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 on a global, million-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 may not be able to do in a short time (e.g., survey, media monitoring). With the generic and reusable method to deal with a high volume of news articles published worldwide, DeepCovid can be employed for any healthcare-related events such as a future similar pandemic, and potentially be extended for events of other nature beyond the scope of healthcare, such as financial crises. The DeepCovid framework can assist policymakers to provide fast responses to future similar public health concerns.

Setting DeepCovid in place only requires four human actions (see Figure 1): (1) to design classification rules to narrow down to relevant articles, (2) to re-label (some of) the resulting positive and negative articles, (3) to label a small set of global extractive summaries to seed the summarization model and (4) finally, to aggregate extractive summaries into reports. All four steps require a moderate volume of work from human workers, on the order of a few hours to a few days from two humans, which is several orders of magnitude lower than what would be required to manually go through such a scale of data as the one we applied the system on, 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 works using machine learning to address system-level challenges arising from the COVID-19 pandemic do not jointly cover multiple impacts of the pandemic on the worldwide health workforce. We note one study which predicts mental health of Chinese medical workers with logistic regression[54]. In the realm of NLP applications, another study predicts sentiment from tweets from Indian citizens using BERT to assess public opinion during lockdown[55]. A paper leverages Long Short Term Memory networks (LSTMs) to predict the number of deaths from WHO data in three countries[56]. The most relevant system to ours is CO-Search[57], which builds a multi-component 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]. Yet, CO-Search input data is wildly different from the news data in our study.

With a dataset of the scale of EIOS, topic-specific precision and recall evaluation remains 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 two 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 2.8M articles. Language models like the ones used in DeepCovid are not equipped with a semantical understanding of what classification rules are designed to capture, and merely rely on statistical co-occurrence patterns, which enables to expand relevant articles with other articles containing similar topics, albeit phrased with different lexicality. Striving for perfect precision and recall may need other, complementary tools to deep pre-trained language models, such as Knowledge

Graphs.

With regards to optimization, DeepCovid's double objective makes it complicated to be trained into a single phase. 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 two separate training processes. Besides, another limitation of our work lies in the fact that human intervention remains compulsory at steps (1) and (4) mentioned above.

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 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 which 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 also for evaluation of classification precision.

Limitations

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

Firstly, our study is restricted to English-only news articles. This decision was based on the abundance of English sources compared to other languages. From the perspective of data source, the model that was 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 much feasible. It would involve replacing each deep learning component with a multilingual model version (e.g., 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 of 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 by LLM-generated annotations instead, particularly with powerful LLMs such as GPT-4. However, although annotation time would be reduced, using the GPT-4 API still bears a significant cost. Besides, annotations generated by LLMs would still need to be validated by human experts.

Although broad and valuable, the dataset contains a relatively narrow type of news coverage, hence, additional insights could be gained through expanding sources to social media channels and broadening the format to multimedia content such as videos. The data and the findings are impacted by specific strategies for open-source collection which 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.

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

Conclusion

In this study, we introduced the DeepCovid system. Relying on two deep learning-powered components, DeepCovid automatically finds topic-focused relevant news articles among millions of candidates, before writing succinct extractive summaries out of 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 byWHO. The findings are to be published in a separate paper. DeepCovid methodology also makes it suitable to other use cases 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 World Health Organization for valuable insights on the data and feedback on the paper.

During this research, co-author Dusan Milovanovic was employed at the 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.

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) 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.

Abbreviations

AI - Artificial Intelligence
AUC - Area Under the Receiver Operating Characteristic Curve
EIOS - Epidemic Intelligence from Open Sources
LLMs - Large Language Models
NLP- Natural Language Processing
PHS - Public Health Surveillance
WHO - World Health Organization

References

1 Pilipiec, P., Samsten, I., & Bota, A. (2023). Surveillance of communicable diseases using social media: A systematic review. *PLoS One*, *18*(2), e0282101.

2 Nsubuga, P., White, M. E., Thacker, S. B., Anderson, M. A., Blount, S. B., Broome, C. V., ... & Trostle, M. (2011). Public health surveillance: a tool for targeting and monitoring interventions.

3 Thacker, S. B., & Berkelman, R. L. (1988). Public health surveillance in the United States. *Epidemiologic reviews*, *10*(1), 164-190.

4 Narasimhan, V., Brown, H., Pablos-Mendez, A., Adams, O., Dussault, G., Elzinga, G., ... & Chen, L. (2004). Responding to the global human resources crisis. *The Lancet*, *363*(9419), 1469-1472.

5 Hope, K., Durrheim, D. N., d'Espaignet, E. T., & Dalton, C. (2006). Syndromic surveillance: is it a useful tool for local outbreak detection?. *Journal of Epidemiology & Community Health*, *60*(5), 374-374.

6 Tsao, S. F., Chen, H., Tisseverasinghe, T., Yang, Y., Li, L., & Butt, Z. A. (2021). What social media told us in the time of COVID-19: a scoping review. *The Lancet Digital Health*, *3*(3), e175-e194.

7 Budd, J., Miller, B. S., Manning, E. M., Lampos, V., Zhuang, M., Edelstein, M., ... & McKendry, R. A. (2020). Digital technologies in the public-health response to COVID-19. *Nature medicine*, *26*(8), 1183-1192.

8 Al-Garadi, M. A., Yang, Y. C., & Sarker, A. (2022, November). The Role of Natural Language Processing during the COVID-19 Pandemic: Health Applications, Opportunities, and Challenges. In *Healthcare* (Vol. 10, No. 11, p. 2270). MDPI.

9 Hall, K., Chang, V., & Jayne, C. (2022). A review on Natural Language Processing Models for COVID-19 research. *Healthcare Analytics*, 100078.

10 He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

11 He, P., Gao, J., & Chen, W. (2021). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

12 Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010, February). Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining* (pp. 261-270).

13 Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in human behavior*, *31*, 527-541.

14 Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., & Ghosh, S. S. (2020). Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of medical Internet research*, *22*(10), e22635.

15 Liu, Y., Whitfield, C., Zhang, T., Hauser, A., Reynolds, T., & Anwar, M. (2021). Monitoring COVID-19 pandemic through the lens of social media using natural language processing and machine learning. *Health Information Science and Systems*, *9*(1), 25.

16 Patel, R., Smeraldi, F., Abdollahyan, M., Irving, J., & Bessant, C. (2021). Analysis of mental and physical disorders associated with COVID-19 in online health forums: a natural language processing study. *BMJ open*, *11*(11), e056601.

17 Marshall, C., Lanyi, K., Green, R., Wilkins, G. C., Pearson, F., & Craig, D. (2022). Using natural language processing to explore mental health insights from UK tweets during the COVID-19 pandemic: infodemiology study. *Jmir Infodemiology*, *2*(1), e32449.

18 Evans, S. L., Jones, R., Alkan, E., Sichman, J. S., Haque, A., de Oliveira, F. B. S., & Mougouei, D. (2023). The emotional impact of COVID-19 news reporting: A longitudinal

study using natural language processing. *Human Behavior and Emerging Technologies*, *2023*.

19 Zhao, Y., He, X., Feng, Z., Bost, S., Prosperi, M., Wu, Y., ... & Bian, J. (2022). Biases in using social media data for public health surveillance: A scoping review. *International Journal of Medical Informatics*, *164*, 104804.

20 Aiello, A. E., Renson, A., & Zivich, P. (2020). Social media-and internet-based disease surveillance for public health. *Annual review of public health*, *41*, 101.

21 Brownstein, J. S., Rader, B., Astley, C. M., & Tian, H. (2023). Advances in Artificial Intelligence for infectious-disease surveillance. *New England Journal of Medicine*, *388*(17), 1597-1607.

22 Varol, A. E., Kocaman, V., Haq, H. U., & Talby, D. (2022). Understanding COVID-19 news coverage using medical NLP. *arXiv preprint arXiv:2203.10338*.

23 Gupta, A., & Katarya, R. (2020). Social media based surveillance systems for healthcare using machine learning: a systematic review. *Journal of biomedical informatics*, *108*, 103500.

24 Pham, Q. T., Le, X. T. T., Phan, T. C., Nguyen, Q. N., Ta, N. K. T., Nguyen, A. N., ... & Ho, R. C. (2021). Impacts of COVID-19 on the life and work of healthcare workers during the nationwide partial lockdown in Vietnam. *Frontiers in Psychology*, *12*, 563193.

25 Gupta, N., Dhamija, S., Patil, J., & Chaudhari, B. (2021). Impact of COVID-19 pandemic on healthcare workers. *Industrial psychiatry journal*, *30*(Suppl 1), S282.

26 Lui, M., & Baldwin, T. (2012, July). langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations* (pp. 25-30).

27 Leonardelli, E., Menini, S., Aprosio, A. P., Guerini, M., & Tonelli, S. (2021). Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement. *arXiv preprint arXiv:2109.13563*.

28 Pandey, R., Purohit, H., Castillo, C., & Shalin, V. L. (2022). 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*, *160*, 102772.

29 Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74-81).

30 Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. *Advances in neural information processing systems*, *28*.

31 Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, *7*(1), 411-420.

32 Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

33 Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

34 Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *TheJournal of Machine Learning Research*, *21*(1), 5485-5551.

35 Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.

36 Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

37 Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

38 Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-*

10) (pp. 807-814).

39 Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

40 Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

41 Kryściński, W., McCann, B., Xiong, C., & Socher, R. (2019). Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

42 Goyal, T., & Durrett, G. (2020). Evaluating factuality in generation with dependency-level entailment. *arXiv preprint arXiv:2010.05478*.

43 Narayan, S., Cohen, S. B., & Lapata, M. (2018). Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

44 Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, *22*, 457-479.

45 Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., & Huang, X. (2020). Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*.

46 McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276-282.

47 Gwet, K. L. (2008). Computing inter‐rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29-48.

48 Karn, S. K., Chen, F., Chen, Y. Y., Waltinger, U., & Schütze, H. (2021). Few-shot learning of an interleaved text summarization model by pretraining with synthetic data. *arXiv preprint arXiv:2103.05131*.

49 Liu, Y., & Lapata, M. (2019). Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

50 Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

51 Nallapati, R., Zhou, B., Gulcehre, C., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

52 See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

53 Jia, R., Cao, Y., Tang, H., Fang, F., Cao, C., & Wang, S. (2020, November). Neural extractive summarization with hierarchical attentive heterogeneous graph network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3622-3631).

54 Wang, X., Li, H., Sun, C., Zhang, X., Wang, T., Dong, C., & Guo, D. (2021). Prediction of mental health in medical workers during COVID-19 based on machine learning. Frontiers in public health, 9, 697850.

55 Chintalapudi, N., Battineni, G., & Amenta, F. (2021). Sentimental analysis of COVID-19 tweets using deep learning models. Infectious disease reports, 13(2), 329-339.

56 Esteva, A., Kale, A., Paulus, R., Hashimoto, K., Yin, W., Radev, D., & Socher, R. (2020). Co-search: Covid-19 information retrieval with semantic search, question answering, and abstractive summarization. arXiv preprint arXiv:2006.09595.

57 Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., ... & Kohlmeier, S. (2020). Cord-19: The covid-19 open research dataset. ArXiv.

58 Aldhyani, T. H., Alrasheed, M., Alqarni, A. A., Alzahrani, M. Y., & Alahmadi, A. H. (2020). Deep learning and holt-trend algorithms for predicting covid-19 pandemic. MedRxiv, 2020-06.

59 Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D.

(2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

60 Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

61 Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., & Chen, W. (2021). What Makes Good In-Context Examples for GPT-$3 $?. *arXiv preprint arXiv:2101.06804*.

62 Goyal, T., Li, J. J., & Durrett, G. (2022). News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

63 Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824-24837.

64 Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., ... & Jiang, D. (2023). Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

65 Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., ... & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

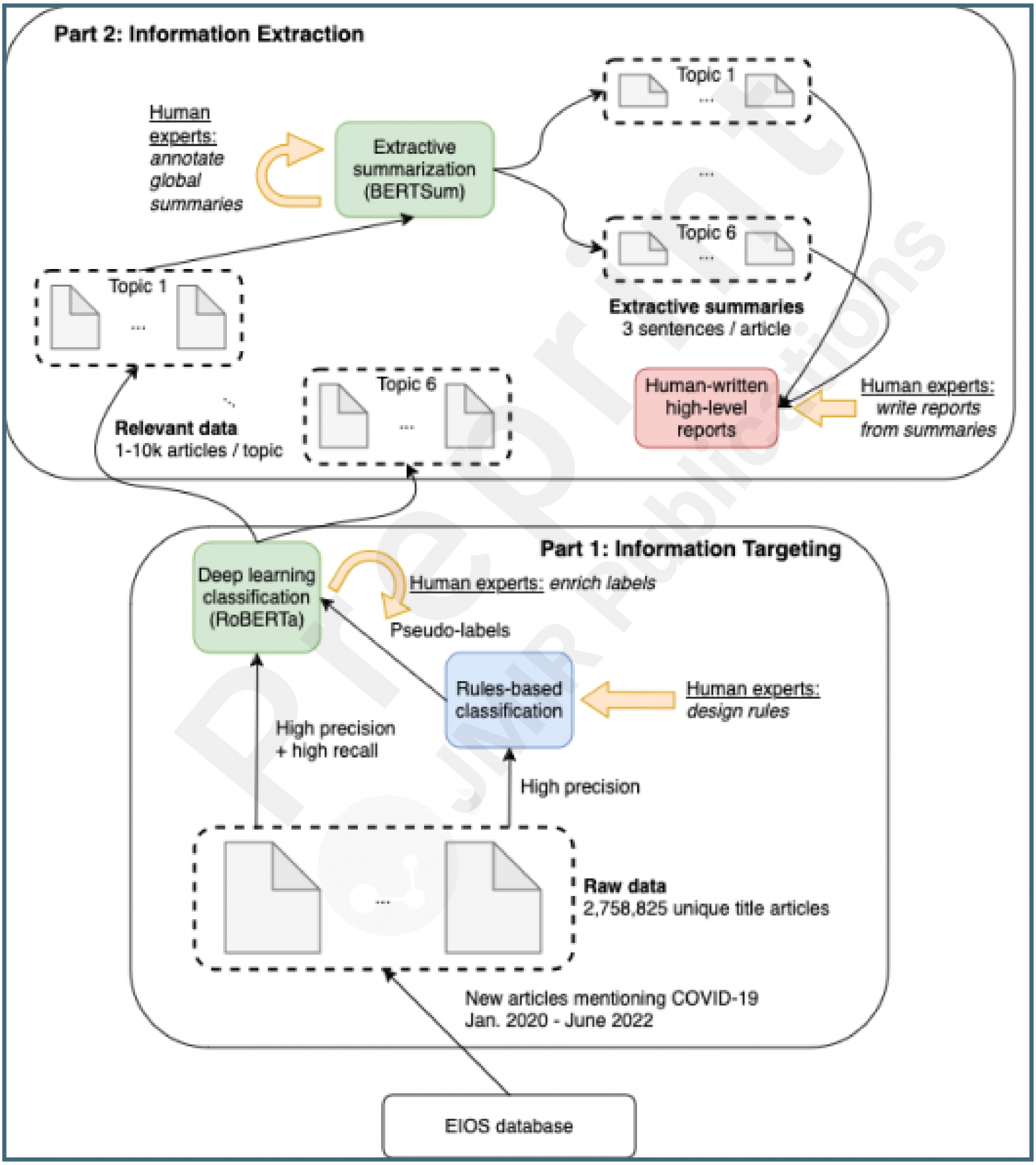
66 Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, *35*, 22199-22213.

67 Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

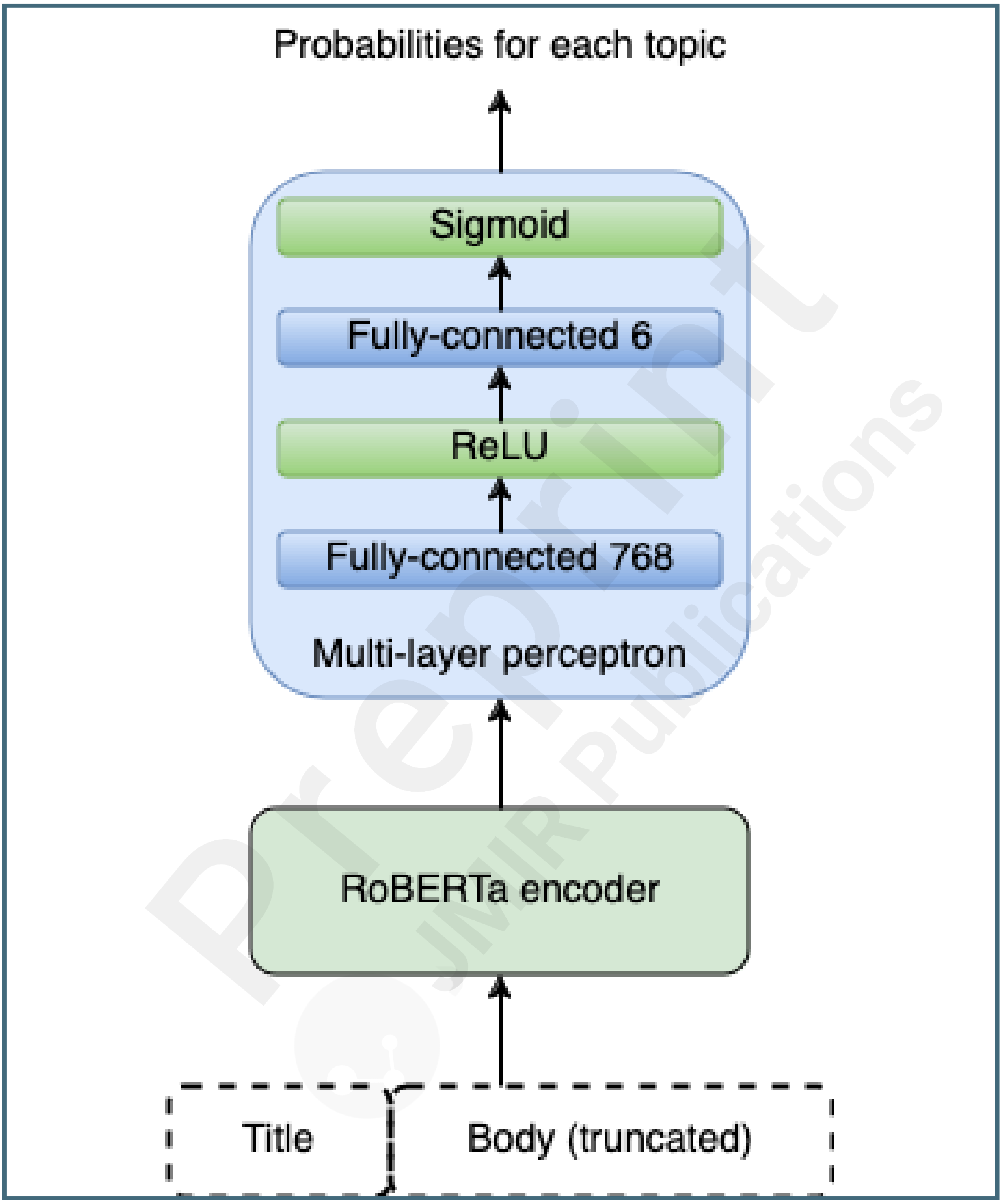
Supplementary Files

Figures

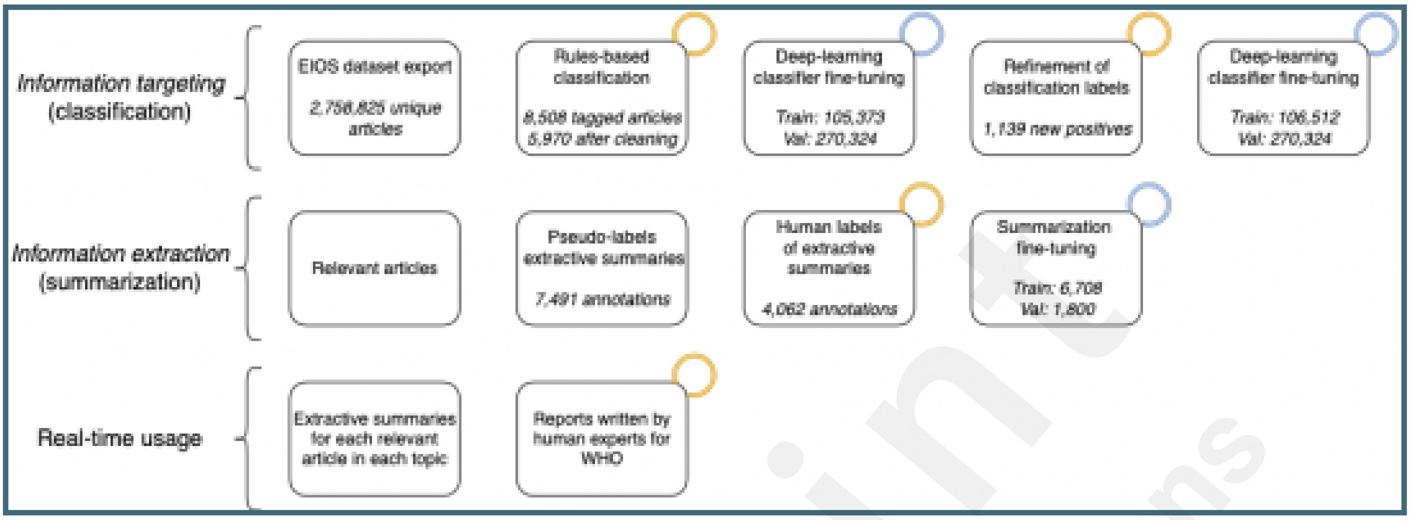
DeepCovid model architecture overview (read from bottom to top). Colored blocks correspond to machine learning models. Golden arrows indicate interventions necessitated from human experts. The first part of the system aims to find relevant articles, and it trains and applies a deep learning classifier 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 three sentences. Corresponding summaries are then analyzed by human experts to produce reports.


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
1-10k 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
Pseudo-labels
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

Deep learning classifier architecture.



Flow chart for DeepCovid. We show the step by step process transforming a raw, 2.8M news articles dataset (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.


Information targeting
(classification)
EIOS dataset export
2,756,825 unique articles
Rules-based classification
8,508 tagged articles
5,970 after cleaning
Deep-learning classifier fine-tuning
Train: 105,373
Val: 270,324
Refinement of classification labels
1,139 new positives
Deep-learning classifier fine-tuning
Train: 106,512
Val: 270,324
Information extraction
(summarization)
Relevant articles
Pseudo-labels extractive summaries
7,491 annotations
Human labels of extractive summaries
4,062 annotations
Summarization fine-tuning
Train: 6,708
Val: 1,800
Real-time usage
Extractive summaries for each relevant article in each topic
Reports written by human experts for WHO