ebadi_2021_understanding_the_temporal_evolution_of_COVID_19_research_through_machine_learning_and_natural_language_processing

Year

2021

Author(s)

Ebadi, Ashkan and Xi, Pengcheng and Tremblay, St{'e}phane and Spencer, Bruce and Pall, Raman and Wong, Alexander

Title

Understanding the temporal evolution of COVID-19 research through machine learning and natural language processing

Venue

Scientometrics

Topic labeling

Manual

Focus

Secondary

Type of contribution

Established approach

Underlying technique

Manual labeling

Topic labeling parameters

\

Label generation

The three experts manually labeled the generated topics after careful examination of the extensive set of keywords for each topic.

The seven labeled topics are: (1) Oncology, (2) Personal protective equipment (PPE), (3) Analytics, (4) Rehabilitation-panic, (5) High-risk groups, (6) Genomics, and (7) Intubation-oxygenation.

Motivation

\

Topic modeling

STM

Topic modeling parameters

Three STM models: (1) STM model built on the entire dataset with a monthly granularity, (2) STM model built only on PubMed dataset with a weekly granularity, and (3) STM model built on ArXiv dataset with a weekly granularity. We will refer to these models as STM-ALL, STM-PUBMED, and STM-ARXIV

Nr of topics: 2 to 20

Nr. of topics

7 (STM-ALL), 7 (STM-PUBMED) and 4 (STM-ARXIV)

Label

Manually assigned single or multi word labels

Label selection

3 labels per topic: “We used more than one expert as well as an odd number of experts to reduce the subjectivity effect in labeling topics as well as setting the optimal number of topics.“

Label quality evaluation

\

Assessors

Three domain experts

Domain

Paper: Health (COVID-19)

Dataset: Health (COVID-19)

Problem statement

In this study, we used multiple data sources, i.e., PubMed and ArXiv, and built several machine learning models to characterize the landscape of current COVID-19 research by identifying the latent topics and analyzing the temporal evolution of the extracted research themes, publications similarity, and sentiments, within the time-frame of January–May 2020.

Corpus

Origin: PubMed and ArXiv

Nr. of documents: 14,172

Details:

- January–May 2020

Document

Titles and abstracts

Pre-processing

- dropped articles with no/incomplete publication dates as well as those with neither titles nor abstracts
- converting the text to lowercase
- correcting special characters
- removing stop words using a customized English stop words list
- removing punctuations

```
@article{ebadi_2021_understanding_the_temporal_evolution_of_COVID_19_research_th
rough_machine_learning_and_natural_language_processing,
    abstract = {The outbreak of the novel coronavirus disease 2019 (COVID-19),
caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has
been continuously affecting human lives and communities around the world in many
ways, from cities under lockdown to new social experiences. Although in most
cases COVID-19 results in mild illness, it has drawn global attention due to the
extremely contagious nature of SARS-CoV-2. Governments and healthcare
professionals, along with people and society as a whole, have taken any measures
to break the chain of transition and flatten the epidemic curve. In this study,
we used multiple data sources, i.e., PubMed and ArXiv, and built several machine
learning models to characterize the landscape of current COVID-19 research by
identifying the latent topics and analyzing the temporal evolution of the
extracted research themes, publications similarity, and sentiments, within the
time-frame of January--May 2020. Our findings confirm the types of research
available in PubMed and ArXiv differ significantly, with the former exhibiting
greater diversity in terms of COVID-19 related issues and the latter focusing
more on intelligent systems/tools to predict/diagnose COVID-19. The special
attention of the research community to the high-risk groups and people with
complications was also confirmed.},
    author = {Ebadi, Ashkan and Xi, Pengcheng and Tremblay, St{\'e}phane and
```

```
Spencer, Bruce and Pall, Raman and Wong, Alexander},
    date-added = {2023-04-17 23:10:51 +0200},
    date-modified = {2023-04-17 23:10:51 +0200},
    day = {01},
    doi = {10.1007/s11192-020-03744-7},
    issn = {1588-2861},
    journal = {Scientometrics},
    month = {Jan},
    number = {1},
    pages = {725--739},
    title = {Understanding the temporal evolution of COVID-19 research through
machine learning and natural language processing},
    url = {https://link.springer.com/content/pdf/10.1007/
s11192-020-03744-7.pdf},
    volume = {126},
    year = {2021},
    bdsk-url-1 = {https://link.springer.com/content/pdf/10.1007/
s11192-020-03744-7.pdf},
    bdsk-url-2 = {https://doi.org/10.1007/s11192-020-03744-7}}
```

#Thesis/Papers/FS