Text Classification on COVID-19: a Transformer-based Approach

[2]Sofia Camilla Todeschini, [1,*]Marco Canaparo, [1,2]Elisabetta Ronchieri

[1]INFN CNAF, Bologna, Italy

[2]Department of Statistical Sciences, University of Bologna, Bologna, Italy

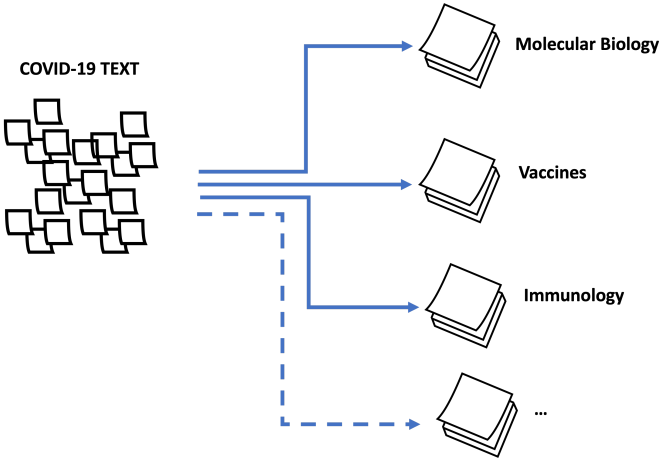
[*]*Speaker*

March 26, 2024

International Symposium on Grids and Clouds 2024
24 - 29 March 2024
Academia Sinica, Taipei, Taiwan

Table of Contents

Text classification aims to **assign labels** or **tags to textual units**, such as sentences, queries, and documents.



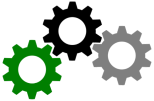
The **COVID-19 pandemic** has led to an **unprecedented amount of scientific publications**, growing at a pace never seen before.

Systematic Reviews (SR)

IDENTIFY

SELECT AND ASSESS

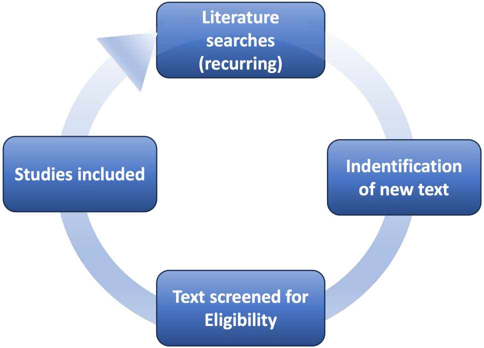
SYNTHESIZE AND FILTER



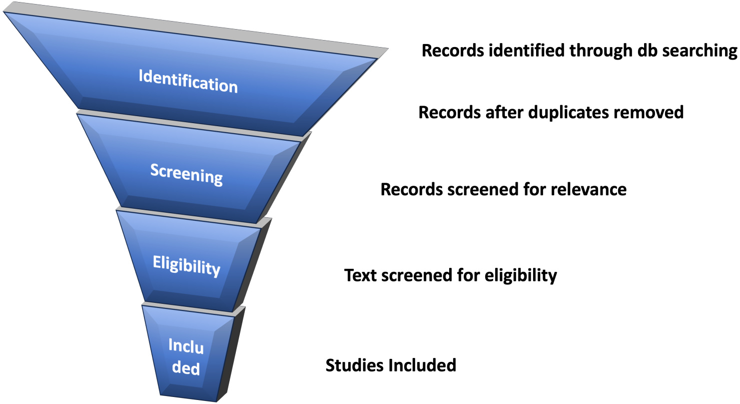
PRO & CONS of Systematic Review

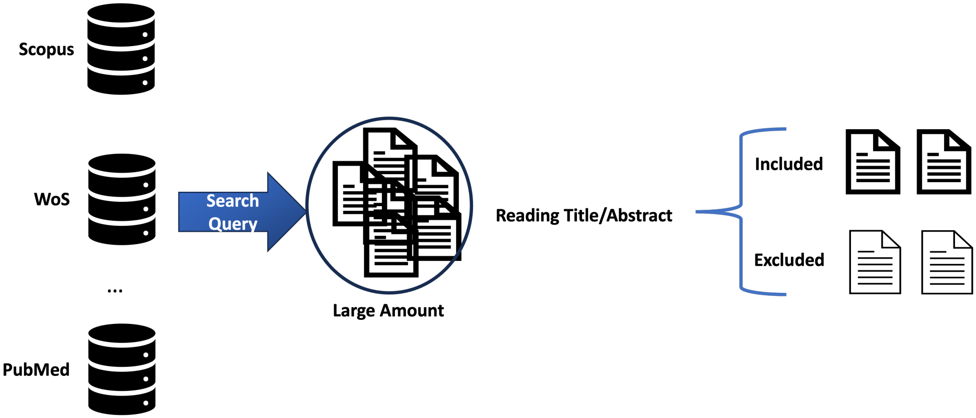
PRO	CONS
Trasparent	Time consuming and resource intensive
Replicable & Updatable	Unpredictable
Reduce bias	Dependent on primary studies
Provide best evidence for decision making	Findings not directly relevant
Findings easily assessable	Commissioning is costly
Can improve future research design	Difficult to locate relevant unpublished research

- Allows updating information as soon as new evidence becomes available.
- Living evidence can narrow the gap between knowledge and practice.
- Maintaining living evidence systems requires ongoing manual curation by skilled human resources.
- Title and/or abstract screening is still one of the most time-consuming tasks.

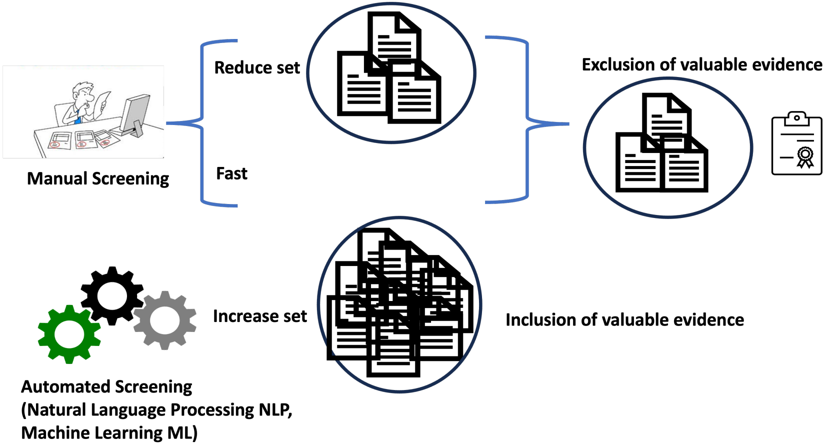


The **Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA)** is **a** reporting **guideline** that comprises a checklist of 27 items and a **four-phase** flow diagram recommended **for** reporting in **SR**.



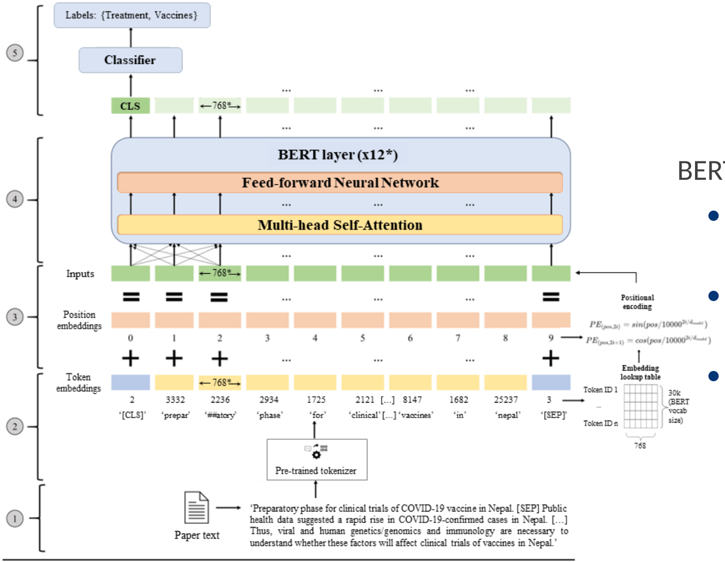

Scopus
WoS
...
PubMed
Search Query
Large Amount
Reading Title/Abstract
Included
Excluded

Screening Phase - Manual vs Automated



- To streamline the **screening phase**, we have employed pre-trained transformer models, such as BioBERT and PubMedBERT derived from Bidirectional Encoder Representations from Transformers (BERT).
- Transformers are based on **pre-training** and **self-attention**:
  - Pre-training:
    - cheap training time and data size for fine tuning.
    - mitigates overfitting problem.
  - Self-attention:
    - considers global dependencies between words, keeping track of the whole input sequence.
- BERT is the precursor pre-trained model.
- Further BERT-based models have been defined according to the specific corpus domain, such as biomedical and COVID-19.

BERTbase Architecture for text classification



BERTbase:

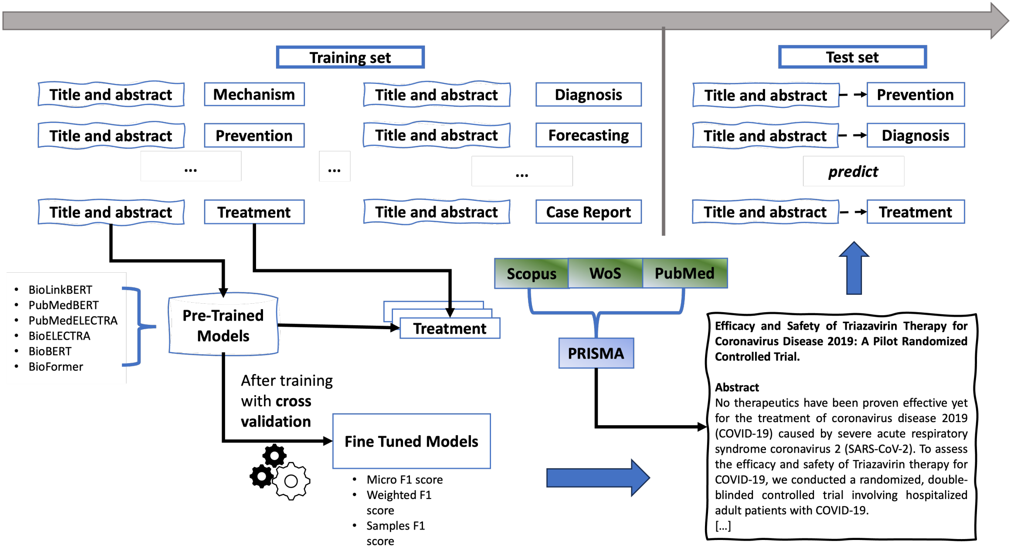
- 12 layers of transformers
- 768 hidden size
- 12 self-att.

Research Questions

RQ1 : Can transformer-based models successfully integrate into the living systematic review process on COVID-19 research?

RQ2 : Are the provided labels enough to classify documents?

Overview

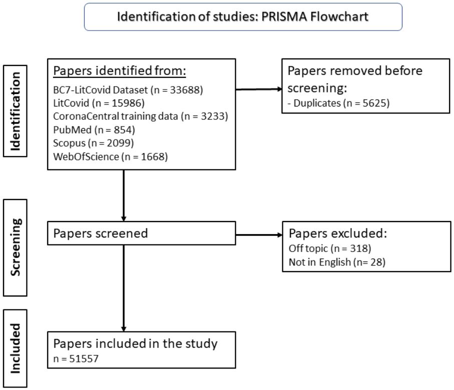

Training set
Title and abstract
Mechanism
Title and abstract
Diagnosis
Title and abstract
Prevention
Title and abstract
Forecasting
...
...
...
Title and abstract
Treatment
Title and abstract
Case Report
Test set
Title and abstract
Prevention
Title and abstract
Diagnosis
predict
Title and abstract
Treatment
BioLinkBERT
PubMedBERT
PubMedELECTRA
BioELECTRA
BioBERT
BioFormer
Pre-Trained Models
Treatment
Scopus
WoS
PubMed
PRISMA
After training with cross validation
Fine Tuned Models
Micro F1 score
Weighted F1 score
Samples F1 score
Efficacy and Safety of Triazavirin Therapy for Coronavirus Disease 2019: A Pilot Randomized Controlled Trial.
Abstract
No therapeutics have been proven effective yet for the treatment of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). To assess the efficacy and safety of Triazavirin therapy for COVID-19, we conducted a randomized, double-blinded controlled trial involving hospitalized adult patients with COVID-19.
[...]

e.g. Scopus search query keywords in title:

- artificial intelligence OR ... OR deep learning OR computer vision
- sars-cov-2 OR corona OR covid-19 OR coronavirus OR covid
- imaging OR image OR screen OR screening OR scan OR medical OR medical imaging OR medical image OR chest OR pulmonary OR lung OR simulation OR modeling OR tomographic OR therapy OR clinical OR gamma-ray OR radiation OR imaging systems OR ... OR quantitative imaging

From the titles and abstracts

- Identified almost 58k
- Included 51k
- Excluded only 5k



Search queries performed in July 2023.

Training and Test sets

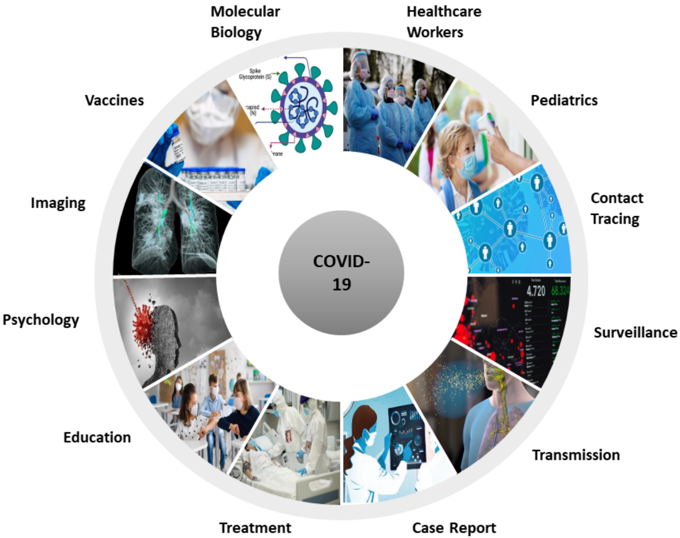
Corpus	Training set	Test set	Total	Labeled
BC7-LitCovid	29436	2293	31729	✓
LitCovid	14781	1205	15986	✓
CoronaCentral	1308	200	1508	✓
PubMed	0	718	718	
Scopus	0	708	708	
WebOfScience	0	908	908	
Total	45525	6032	51557	

Model name	Source	Instances	Words
PubMedBERT	PubMed	14M	3.2B
BioLinkBERT	PubMed	14M	3.2B
PubMedELECTRA	PubMed	14M	3.2B
BioELECTRA	PubMed	22M	4.2B
	PubMed Central (PMC)	3.2M	9.6B
BioBERT	PubMed	N/A	4.5B
	PubMed Central (PMC)		13.5B
BioFormer-LitCovid	PubMed	33M	N/A
	PubMed Central (PMC)	1M	
	LitCovid	164K	

Pre-Trained Models Size Comparisons

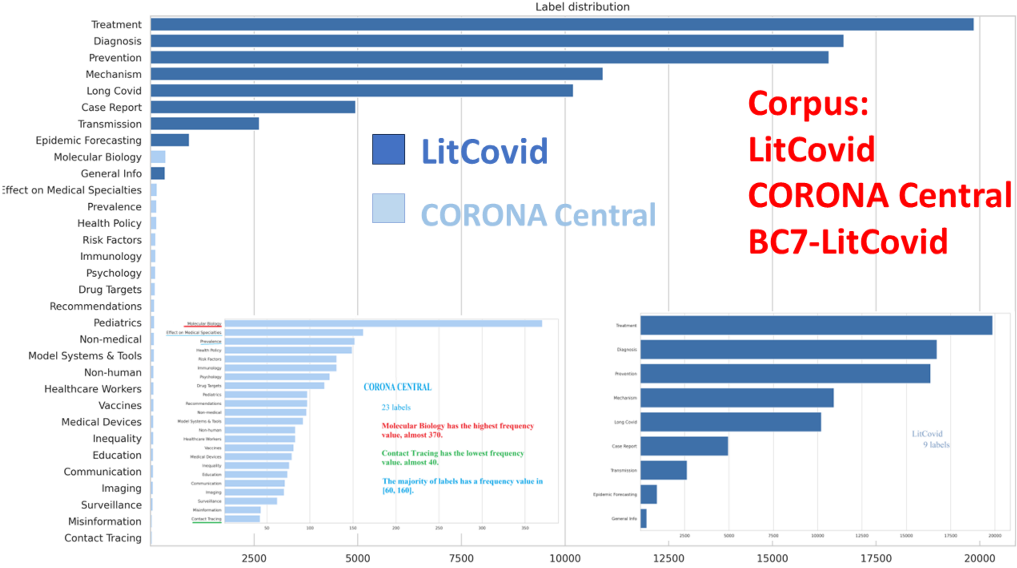
Size	Model	Hidden Size	Att. Heads	Hidden Layers	Parameters
Large	PubMedBERT-Large, BioLinkBERT-Large, PubMedELECTRA-Large, BioBERT-Large	1024	16	24	340M
Base	PubMedBERT-Base, BioLinkBERT-Base, PubMedELECTRA-Base, BioBERT-Base, BioELECTRA	768	12	12	110M
Tiny	BioFormer-LitCovid	512	8	8	14.5M

Examples of LitCovid and Corona Central Defined Labels



Instance Total Label Distribution in the Training Set

2 Methodology



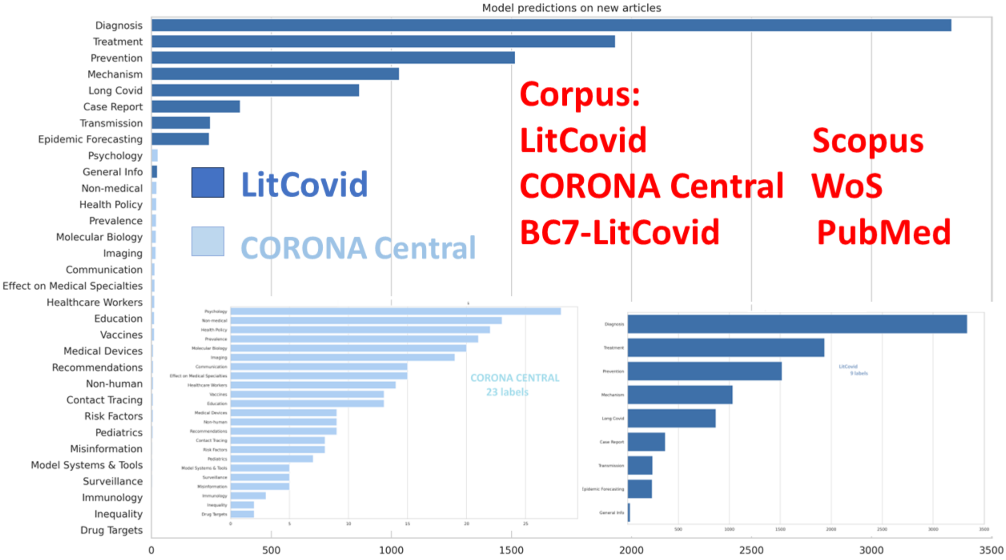
Performance F1 Score Metric Methods for Model Comparisons

- **Micro average** computes a global average F1 score by considering total true positives, false negatives, and false positives. Does not separate each class.
- **Macro-average** computes the arithmetic mean of all the per-class F1 scores.
- **Weighted average** computes the weighted average of all the per-class F1 score. The weight W depends on the number of instances of each class.
  $F1_{class1} \times W1 + F2_{class2} \times W2 + \ldots + FN_{classN} \times WN$
- **Samples average** computes F1 score metric for each instance, and returns their average (specifically meaningful for multilabel classification).

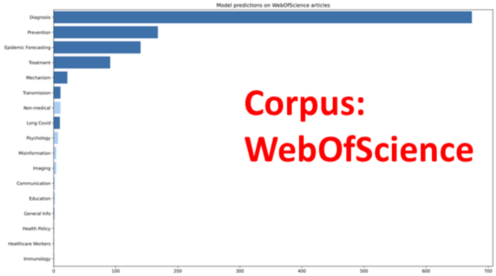
Table of Contents

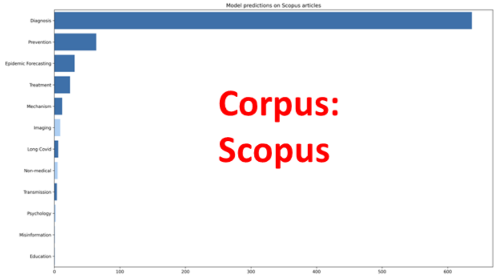
Model Labels Prediction in the Test Set

3 Results

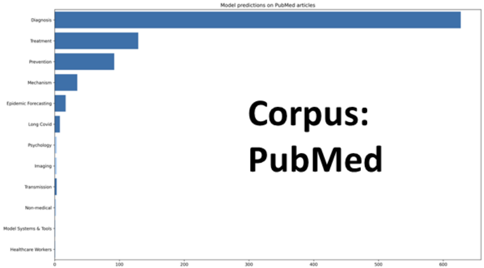


Model Labels Prediction in PubMed, Scopus and WoS

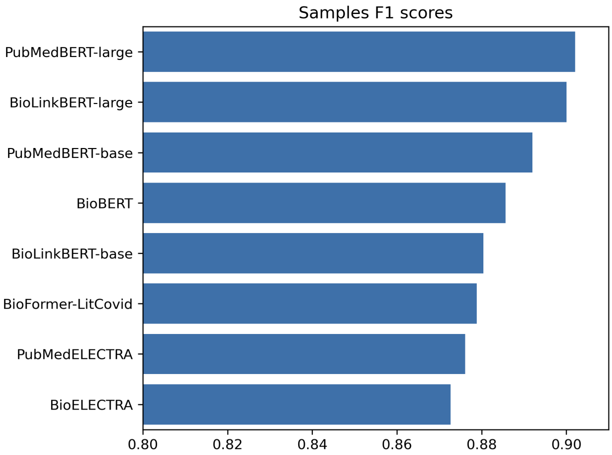




WebOfScience and Scopus have the same first 5 labels.



Performance F1 Metric with Samples Method



The **PubMedBERT-large** model got the best performance.

Method	F1
Micro	0.8967
Weighted	0.9010
Samples	0.9021

Table of Contents

Conclusion

- RQ1: Can transformer-based models successfully integrate into the living systematic review process on COVID-19 research?
  - Yes, our results show that transformer-based models can be integrated to classify **papers on COVID-19** achieving up to **0.9021** of F1-value with the **PubMedBERT-large** model.
  - Model is **domain specific**.
- Transformers have a positive impact on **speed**, **efficiency** and **customizability** of systematic reviews.
- RQ2: Are the provided labels enough to classify documents? Our research indicates that the labels extracted from LitCovid and Corona Central provide only broad information (such as diagnosis or treatment) regarding the content of publications.

Future Work

- Addition of labels to enable a more specific classification.
  - e.g. Treatment. Which kind? Which organ is involved?
  - e.g. Diagnosis. Which method? Which part of the body?
- How to organize labels:
  - Flat Text Classification: classes are defined independently of one another.
  - Hierarchical Text Classification: pre-defined classes are organized in a tree-like structure.
- Update extraction results.
- Explorer other models.

Any Questions?