

# Information Retrieval in an Infodemic: The Case of COVID-19 Publications

D Teodoro\*, S Ferdowsi\*, N Borissov, E Kashani, D Vicente  
Alvarez, J Copara, R Gouareb, N Naderi, P Amini

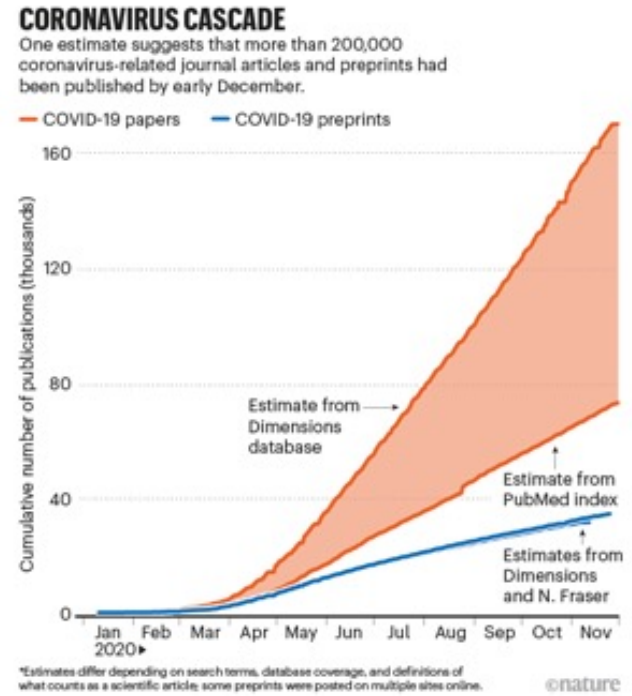
GESAN Journal Club  
19.01.2022

# Outline

- Introduction
- TREC COVID
- Information retrieval model
- Challenge results
- Discussion

# Introduction

- The **COVID-19 pandemic** resulted in an explosive surge of activities within scientific communities and across many disciplines
  - Stakeholders are **unable to keep up** with the fast-evolving body of knowledge disseminated
- **Infodemic**: overabundance of information online and offline with often negative impacts on the population
  - **Confusion** and desensitization among audiences



Else H. How a torrent of COVID science changed research publishing-in seven charts. Nature. 2020:553-.

# COVID Infodemic

 France 24

## Teachers in France stage massive walk-out over Covid confusion

Teachers in France stage massive walk-out over Covid confusion ... on keeping schools open to ease pressure on parents through the pandemic.

4 days ago



# COVID Infodemic

- More than 200 Covid-19 papers have been retracted
  - Elementary calculation errors, lack of transparency, conclusions not supported by the data, etc.

## **New big data study of 145 countries show COVID vaccines makes things worse (cases and deaths)**

I missed this study. So did the mainstream media for some reason. But this study is yet another independent analysis that is difficult to refute: we have been misled by the CDC, FDA, and NIH.



Steve Kirsch  
2 hr ago

♥ 174    💬 118    ➦

\*<https://retractionwatch.com/retracted-coronavirus-covid-19-papers/>

# WHO Framework for Managing the COVID-19 Infodemic



**Action area 1:** strengthening the scanning, review and verification of evidence and information



**Action area 2:** strengthening the interpretation and explanation of what is known, fact-checking statements, and addressing misinformation



**Action area 3:** strengthening the amplification of messages and actions from trusted actors to individuals and communities that need the information

# Objective

To investigate an **information retrieval model** supported by deep language models **to improve search and discovery** of COVID-19 scientific literature

# The TREC-COVID challenge

- **A query set** capturing relevant search questions of researchers during the pandemic
- **Run in 5 rounds** with a total of more than 50 teams
- Started with **30 topics** in round 1 and added **5 new** topics in each new round
  - 50 topics in round 5

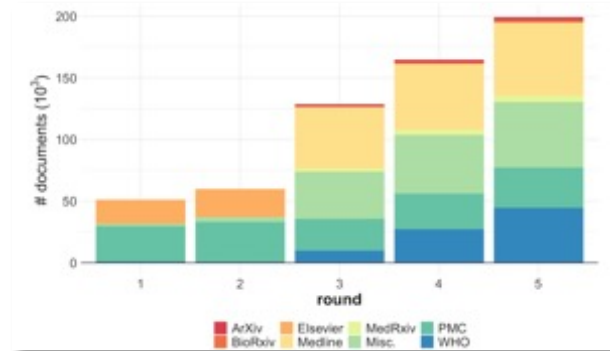


# The TREC-COVID challenge

Topic	Query	Question	Narrative
1	Coronavirus origin	What is the origin of COVID-19?	Seeking a range of information about the SARS-CoV-2 virus's origin, including its evolution, animal source, and first transmission into humans
25	Coronavirus biomarkers	Which biomarkers predict the severe clinical course of 2019-nCoV infection?	Looking for information on biomarkers that predict disease outcomes in people infected with coronavirus, specifically those that predict severe and fatal outcomes
50	mRNA vaccine coronavirus	What is known about an mRNA vaccine for the SARS-CoV-2 virus?	Looking for studies specifically focusing on mRNA vaccines for COVID-19, including how mRNA vaccines work, why they are promising, and a

# The CORD-19 dataset

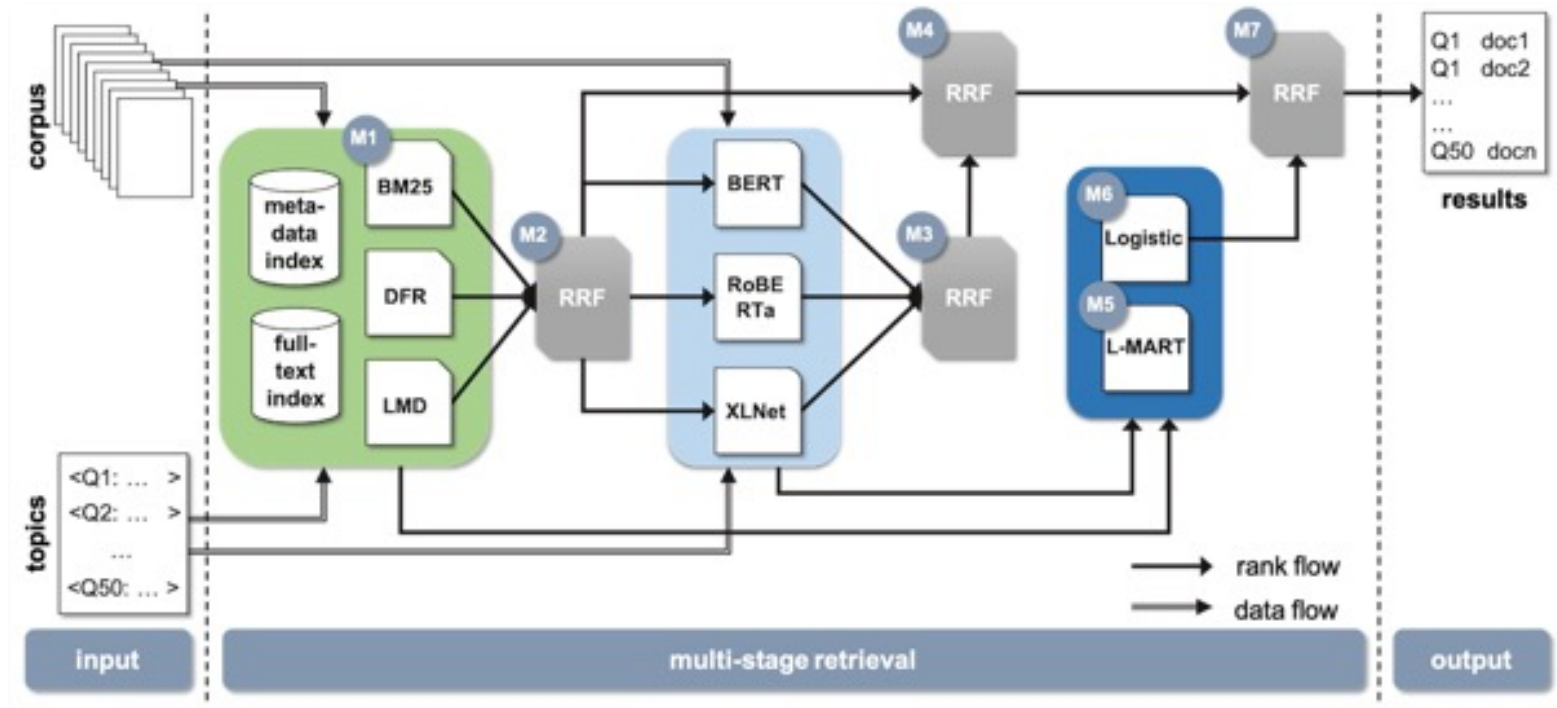
- **CORD-19 dataset:** Effort to gather publications, preprints, and reports related to the coronaviruses and acute respiratory syndromes from the AI2
- Large and dynamically growing **semi-structured dataset** from various sources
  - PubMed, PubMed Central (PMC), WHO, bioRxiv, medRxiv, arXiv, etc.
- **Content:** metadata, including title, abstract, and authors, among others, and the full text or link to full-text files when available



# Teams

Team	Institution	Country
CSIROmed	Commonwealth Scientific and Industrial Research Organisation	Australia
anserini	University of Waterloo	Canada
covidex	University of Waterloo	Canada
xj4wang	University of Waterloo	Canada
HKPU	Hong Kong Polytechnic University	China
mpiid5	Max Planck Institute for Informatics	Germany
UCD_CS	University College Dublin	Ireland
uogTr	Glasgow Terrier Team	Scotland
Elhuyar_NLP_team	Elhuyar Foundation	Spain
<b>risklick</b>	<b>ours</b>	<b>Switzerland</b>
unique_ptr	Google Research	US
SFDC	Salesforce	US
udel_fang	University of Delaware	US
UIowaS	University of Iowa	US

# Multistage retrieval methodology



# First-stage retrieval

## Classic query-document **probabilistic weighting models**

- BM25

$$\begin{aligned}w(t, d, D) &= tf(t, d) \cdot idf(t, D) \\ &= \frac{f(t, d) \cdot (k_1 + 1)}{f(t, d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{avg_l}\right)} \cdot \log\left(\frac{|D|}{n_t}\right)\end{aligned}$$

- DFR

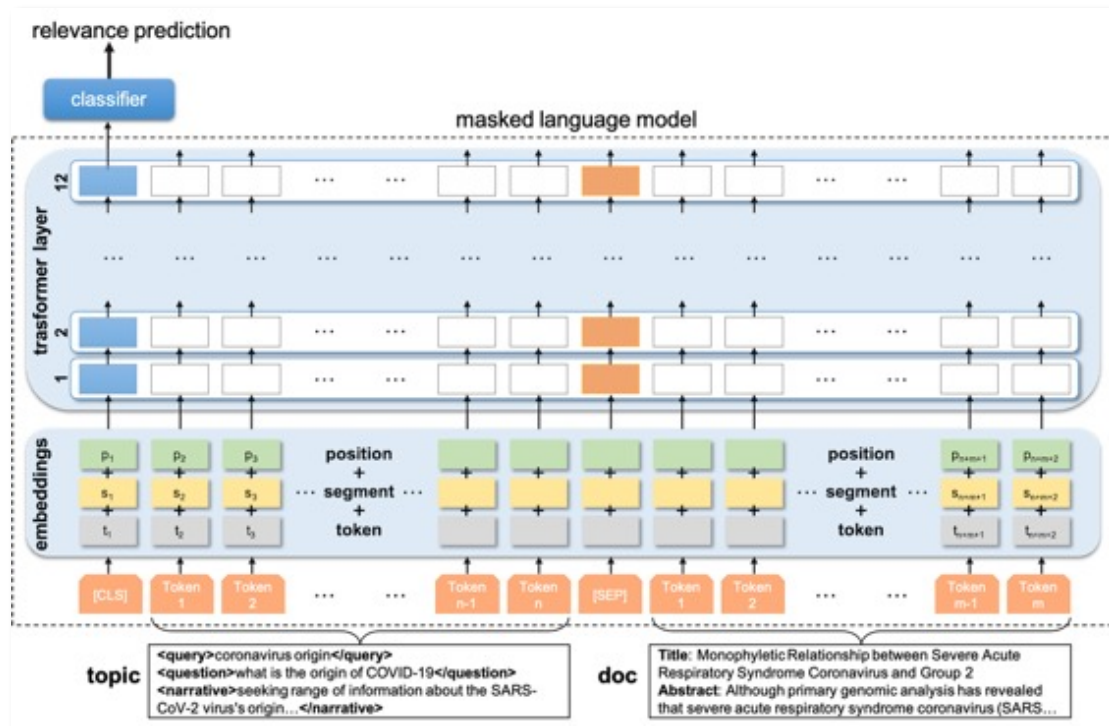
$$w(t, d, D) = k \cdot \log p_M(t \in d|D)$$

- LMD

$$w(t, d, D) = \frac{|d|}{|d| + \mu} \cdot p(t|d) + \frac{\mu}{|d| + \mu} \cdot p(t|D)$$

# Second-stage reranking

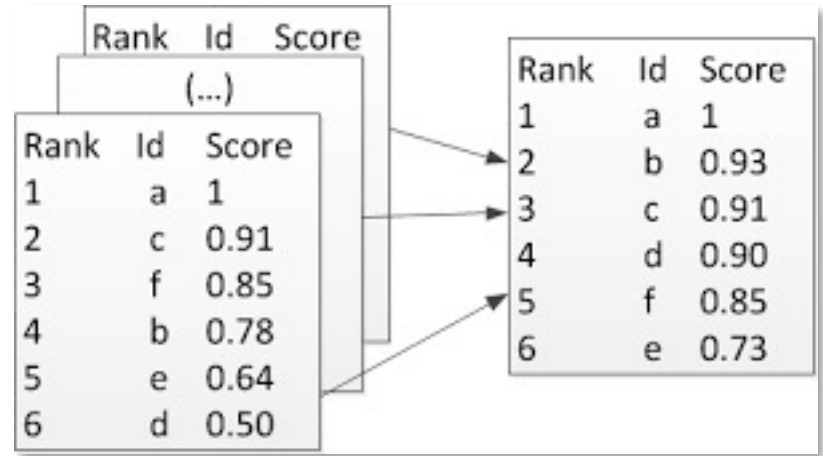
- Models
  - BERT
  - RoBERTa
  - XLNet
- Training set
  - 46k annotated query-doc pairs



# Combining model results

- Reciprocal rank fusion (RRF)
- Given a **set of documents**  $D$  to be sorted and a **set of ranking files**  $R = \{r_1 \dots r_n\}$ , each with a permutation on  $1 \dots |D|$ , the aggregated score is computed using the following equation:

$$s(q, d, R) = \sum_{i=1}^n \frac{1}{k+r_i(q,d)}$$



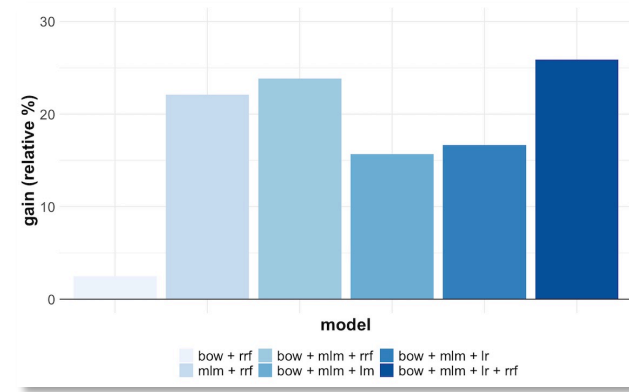
# Results

Run	Name	Description
1	bm25	Run based on the baseline BM25 model using the metadata index
2	bow + rrf	An RRF combination of BM25, DFR, and LMD models computed against the metadata and full-text indices
3	mlm + rrf	An RRF combination of BERT, RoBERTa, and XLNet models applied to run 2
4	bow + mlm + rrf	An RRF combination of runs 2 and 3
5	bow + mlm + lm	A LambdaMART-based model using features from the individual models used to create runs 2 and 3
6	bow + mlm + lr	A logistic regression model using features from the individual models used to create runs 2 and 3
7	bow + mlm + lr + rrf	An RRF combination of runs 2, 3, and 6




# Model performance

Run	Model	NDCG@20	P@20	Bpref	MAP	# rel
1	bm25	0.6320	0.6440	0.5021	0.2707	6533
2	bow + rrf	0.6475	0.6650	0.5174	0.2778	6695
3	mlm + rrf	0.7716	0.7880	0.5680	0.3468	6963
4	bow + mlm + rrf	0.7826	0.8050	0.5616	0.3719	<b>7006</b>
5	bow + mlm + lm	0.7297	0.7460	<b>0.5759</b>	0.3068	6834
6	bow + mlm + lr	0.7375	0.7450	0.5719	0.3439	6976
7	bow + mlm + lr + rrf	<b>0.7961</b>	<b>0.8260</b>	0.5659	<b>0.3789</b>	6939



# Challenge performance

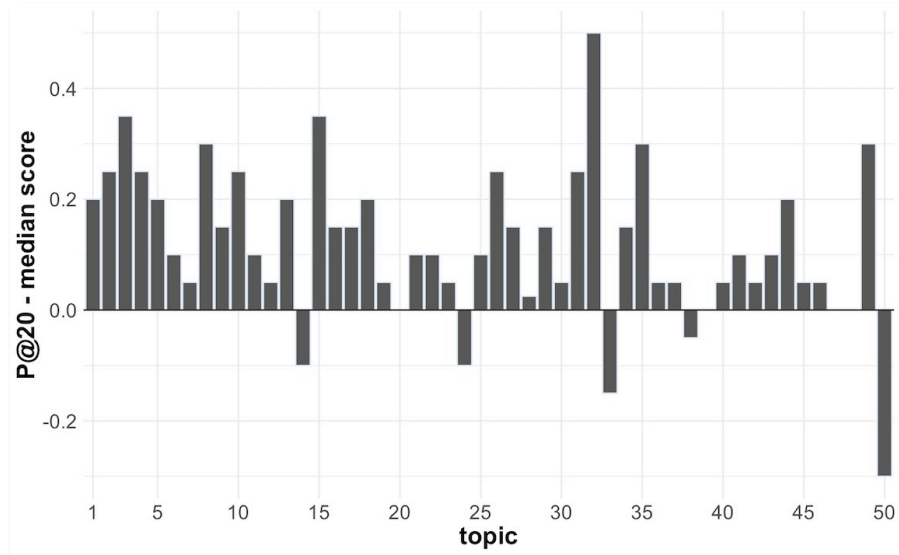
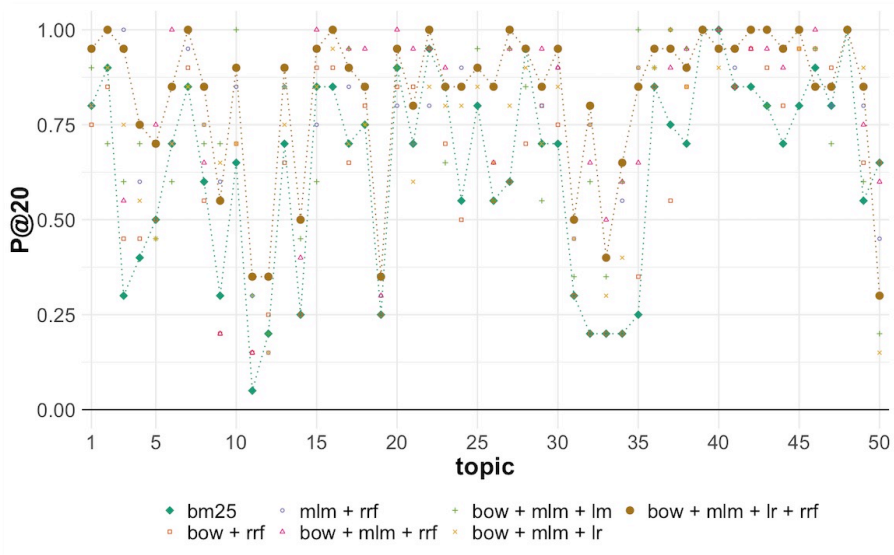


Team	NDCG@20	P@20	Bpref	MAP
unique_ptr	0.8496	0.8760	0.6378	0.4731
covidex	0.8311	0.8460	0.5330	0.3922
Elhuyar_NLP_team	0.8100	0.8340	0.6284	0.4169
<b>risklick (ours)</b>	0.7961	0.8260	0.5759	0.3789
udel_fang	0.7930	0.8270	0.5555	0.3682
CIR	0.7921	0.8320	0.5735	0.3983
uogTr	0.7921	0.8420	0.5709	0.3901
UCD_CS	0.7859	0.8440	0.4488	0.3348
sabir	0.7789	0.8210	0.6078	0.4061
mpiid5	0.7759	0.8110	0.5873	0.3903

# Neural re-ranking vs. ranking fusion

Model	NDCG@20	P@20	Bpref	MAP	# rel
BERT	0.6209	0.6430	0.5588	0.2897	6879
RoBERTa	0.6261	0.6440	0.5530	0.2946	6945
XLNet	0.6436	0.6570	0.5644	0.3064	6926
mlm + rrf	0.7716	0.7880	0.5680	0.3468	6963

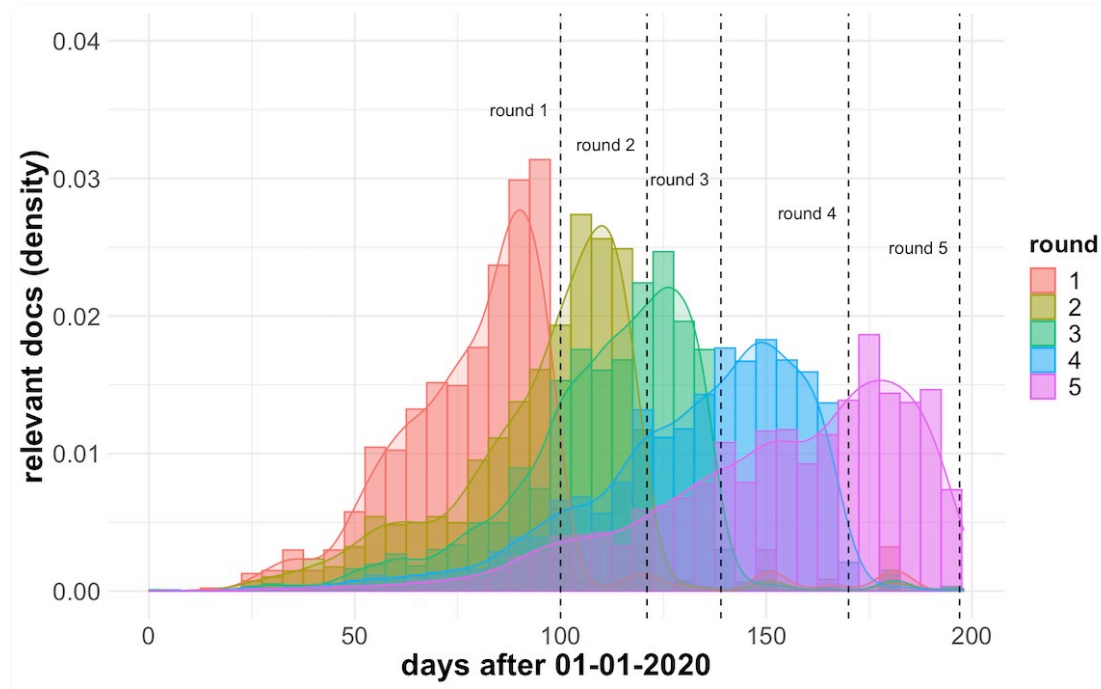
# Topic performance analyses



# Topic performance analyses

Topic ID	Information	Included	Excluded
11	guidelines for triaging patients infected with coronavirus	diagnosis ( <i>“early recognition of coronavirus”, “RT-PCR testing of SARS-CoV-2 for hospitalized patients clinically diagnosed”, ...</i> )	<i>“telephone triage of patients with respiratory complaints”</i>
12	best practices in hospitals and at home in maintaining quarantine	hospital preparedness ( <i>“improving preparedness for”, “preparedness among hospitals”, ...</i> )	<i>“home-based exercise note in Covid-19 quarantine situation”</i>

# Time-dependent relevance analyses



# Conclusion

- The use of the **multistage retrieval approach** significantly improved the search results of COVID-related literature
  - Gain in performance of 25.9% in terms of the NDCG@20 metric compared to a bag-of-words baseline
- The **ensemble of masked language models** brought the highest performance gain to the search pipeline
- The proposed information retrieval pipeline can provide a **potential solution to help stakeholders search and find** the relevant information in the unique situation caused by the COVID-19 pandemic
  - Very competitive results as judged by the official leaderboard of the challenge

Thank you