LDA2Net: Digging under the surface of COVID-19 topics in literature

Topic 61 companion sheet

G. Minello C.R.M.A. Santagiustina M. Warglien

This file contains the following supplementary information for Topic 61 of the manuscript "LDA2Net: Digging under the surface of COVID-19 topics in scientific literature":

- Human label and automatic n-gram label proposals (Table 1)
- Summary measures (Table 2)
- Network of top 25 bigrams (Figure 1)
- Wordclouds of top 25 words by node relevance measure (Figure 2)
- Wordclouds of top 25 bigrams by edge relevance measure (Figure 3)
- Filtered (0.99 percentile) topic network (Figure 4)

Table 1: Human and automatic label proposals. Automatic label candidate for largest word community of the topic. In parenthesis: absolute frequency of the walk out of a sample of size 1000.

Human label	2-gram label	3-gram label	4-gram label
data for the study	analysis->used (9.9%)	data->analysis->used (4.3%)	data->analysis->used->study (1.6%)

Here follows the set of topic-specific measures that have been used to classify the topic and to analyse its structural properties (see manuscript for details):

Table 2: Summary measures

	JSD	Mean propensity	Variance propensity	Modularity	Barrat Clustering Coeff.
value	0.842469	0.007974	0.000122	0.000159	0.641858
rank	112	44	12	46	115

Based on the aforementioned measures, Topic 61 has been classified as a CROSS-CUTTING topic.

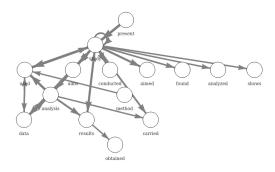


Figure 1: Network of top 25 bigrams (i.e., edges) by weight.



determine
analyzed
indicated conducted
present data investigate
present data investigate
study shows
analysis eg
indings results potained
found indicate
descriptive g applied
quantitative

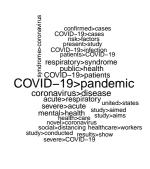
aimshow investigate carried carried services ults method applied of USEQ indicated found study shows aimsconducted aimsconducted aimsconducted ainvestigated



related various moreover present conducted present results obtained aims used obtained total ganalysis obtained garried galiferent aimed according investigated applied

Out-degree Betweenness PageRank

Figure 2: Top 25 unigrams (i.e., nodes) by measure.



results-analyzed results-analysis analysis-data results-study data-study-study-study study-study-study-study-study-study-study-study-study-study-analysis-study-analysis-study-aried-study-analysis-study-aried-study-analysis-analysis-study-aried-study-data-study-data-study-data-study-data-study-obtained-study

study-analyzed
study-loundstudy-analysis
analysis-results
analysis-results
analysis-results
analysis-results
analysis-results
analysis
analysis-analysis
study-used
study-used
study-aims study-aims
study-aims study-aims
study-aims
study-aims
study-aims
study-aims
study-aims
study-aims
study-aims
study-aims
study-aims
study-aims
study-aims
study-aims
study-aims
study-aims
study-aims
study-aims
study-aims
study-aims
results-study

Figure 3: Top 25 bigrams (i.e., edges) by measure.

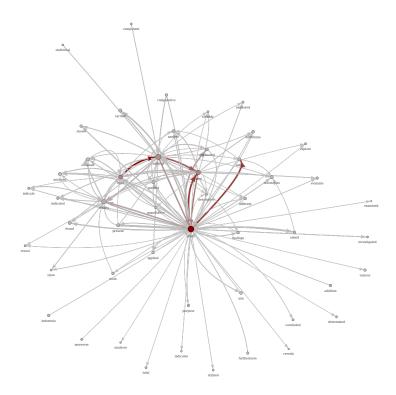


Figure 4: Filtered topic network (by weight). Layout based on Fruchterman-Reingold algorithm. Node size is proportional to topic-specific word probability provided by LDA. Edge width is proportional to topic-specific bigram weight provided by LDA2Net method. Node and edge color represent their betweenness centrality. Isolated nodes have been removed after filtration.