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

Topic 75 companion sheet

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

This file contains the following supplementary information for Topic 75 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	
numbers	one->three (22.8%)	one->three->study (9.2%)	one->three->study->five (1.5%)	

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.
0.794277	0.007406	0.000089	0.059564	0.626822
102	15	2	73	112

Based on the aforementioned measures, Topic 75 has been classified as a CROSS-CUTTING topic.

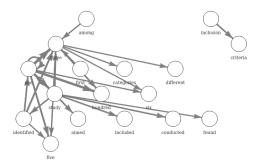


Figure 1: Network of top 25 bigrams (i.e., edges) by weight.

developed included found including among eight six Exposion five study as inclusion five study and inclusion five study and included five study and including among eight study and including five study and including five study and including among eight size.

categories percent
met conducted
inclusion
aimed hundred
still Criteria
mainseven
puthree
of identified
differentanother

developed

polypic inclusion

polypic included

polypic criteria among

ten cone studyeight

different three polypic

found hundred included

main five six

identified categories

percent including

sevence patronpus on significant of the sevence of

according
different
differ

developed categories

public criteria found six among one of the control of the categories

public criteria found one of the categories

Out-degree Betweenness PageRank

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

systematic-review
healthcare-workers
docvib-15-spaces
systematic-review
healthcare-workers
docvib-15-spaces
covid-15-spaces
systematic-review
healthcare-workers
docvib-15-spaces
covid-15-spaces
systematic-review
healthcare-workers
covid-15-spaces
healthcare-workers
covid-15-spaces
healthcare-workers
healthcare

identified>three
three>five>three>five>three
five>three three>three>three
six>three three>three
six>three three>three
six>three three>three
six>three three>three
six>three
one>study>three
one>study>three
noe>study>three>three>three>tidentified>one
one>corporation
six three>tidentified>one
one>corporation
six three>tidentified>tidentified>tidentified
oriteria>three
one>identified
criteria>one

study>included three>one inclusion>criteria identified-five study>aimed three>

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

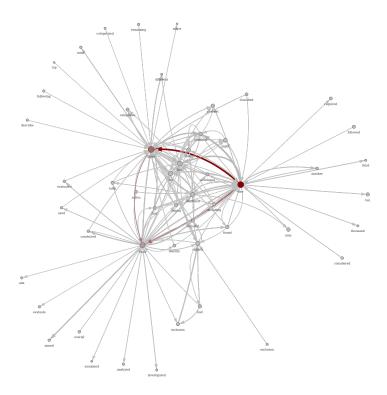


Figure 4: Filtered topic network (by weight). Layout based on Fruchterman-Reingold algorithm. Node size is proportional to topic-specific word probability provided by LDA. Edge width is proportional to topic-specific bigram weight provided by LDA2Net method. Node and edge color represent their betweenness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