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

Topic 109 companion sheet

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

This file contains the following supplementary information for Topic 109 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
statistics	control->groups (32.5%)	control->groups->significant (8.2%)	control->groups->significant->difference (4.3%)

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.
 0.841346	0.007713	0.000190	0.000000	0.628978
111	30	29	10	113

Based on the aforementioned measures, Topic 109 has been classified as a CROSS-CUTTING topic.

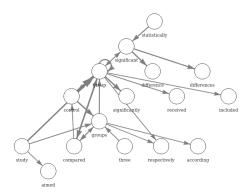


Figure 1: Network of top 25 bigrams (i.e., edges) by weight.



total different
respectively statistically
included compared observed
difference O Othree
groups O Othree
pour use of the control of the cont

observed
in post of higher
it is a significant

found the control of the control of

classified included observed differences among difference groups three compared results of significant study in total higher type received according

among
received according
three significant divided
differences control
included compared
study **Group**observed groups higher
significantly
difference
respectively
results

Out-degree Betweenness PageRank

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

novel>coronavirus
medianicaly entitation
social-distancing publica-health
severe>COVID-19 significantly>higher
study>aimed respiratory>syndrome
rist>lactors severes-acute
coronavirus>disease
COVID-19>patients
COVID-19-pandemic
care-unit control>group health>care
patients>COVID-19 significant>differences
acute>respiratory
significant>difference significantly>lower
mental>health
statistically>significant
syndrome>coronavirus

differences significant-groups differences-group group-significantly groups-control group-significantly group-significant group-significant group-significant group-significant group-significant groups-groups groups-groups group-significant-group control-groups control-group difference-group group-sompared study-group group-compared study-group group-difference

Control groups dudy-amend groups dudy-amend groups atury-amend groups dudy-amend groups dudy-amend groups dudy-amend groups-group dudy-amend groups-groups dudy-amend groups-groups dudy-amend groups-groups dudy-amend groups-groups

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

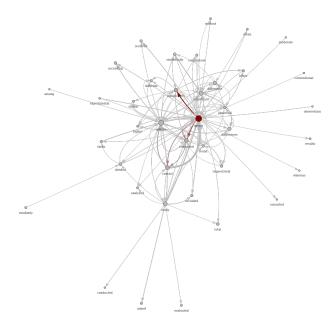


Figure 4: Filtered topic network (by weight). Layout based on Fruchterman-Reingold algorithm. Node size is proportional to topic-specific word probability provided by LDA. Edge width is proportional to topic-specific bigram weight provided by LDA2Net method. Node and edge color represent their betweenness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