

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

Topic 117 companion sheet

G. Minello

C.R.M.A. Santagiustina

M. Warglien

This file contains the following supplementary information for Topic 117 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
report	reported->among (7.6%)	reported->among->less (3.2%)	reported->less->likely->report (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.
value	0.867302	0.007670	0.000104	0.000000	0.612798
rank	116	27	7	19	101

Based on the aforementioned measures, Topic 117 has been classified as a CROSS-CUTTING topic.

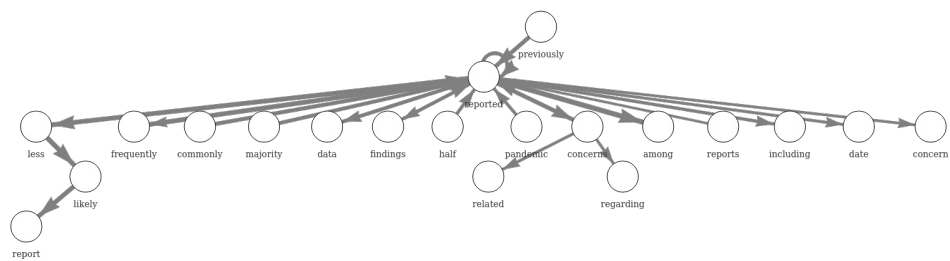


Figure 1: Network of top 25 bigrams (i.e., edges) by weight.

A word cloud of terms related to the COVID-19 pandemic. The most prominent word is 'reported', which is the largest and centered. Other significant words include 'concerns', 'pandemic', 'including', 'nearly', 'report', 'related', 'majority', 'prior', 'likely', 'among', 'since', 'raised', 'needed', 'regarding', 'less', 'findings', 'identified', 'april', 'although', 'half', 'frequently', and 'concern'. The words are arranged in a circular pattern around the central 'reported' word, with varying sizes and orientations.

A word cloud of terms related to the COVID-19 pandemic. The word "reported" is the largest and most central. Other prominent words include "concerns", "findings", "pandemic", "concern", "previously", "regarding", "half", "frequently", "date", "raised", "majority", "reporting", "report", "less", "among", "data", "related", "commonly", "since", "overall", "including", "suggest", and "reports". The words are arranged in a circular pattern around the central word "reported".

A word cloud of terms related to the COVID-19 pandemic. The words are arranged in a circular pattern, with 'reported' being the largest and most central word. Other prominent words include 'reporting', 'concerns', 'pandemic', 'report', 'among', 'regarding', 'including', 'frequently', 'overall', 'date', 'prior', 'previously', 'majority', 'likely', 'findings', 'march', 'less', 'related', 'since', 'half', 'suggest', 'april', 'concern', 'data', 'pandemic', 'reporting', 'concerns', 'pandemic', 'report', 'among', 'regarding', 'including', 'frequently', 'overall', 'date', 'prior', 'previously', 'majority', 'likely', 'findings', 'march', 'less', 'related', 'since', 'half', 'suggest', 'april', 'concern', 'data'.

identified reporting
among
data pandemic
half concerns
likely less majority
reported
since
including
raised
findings
related
concern
information
report although
reports
nearly
frequently
previously
commonly

A word cloud of terms related to the COVID-19 pandemic. The words are arranged in a circular pattern, with 'reported' being the largest and most central word. Other prominent words include 'concerns', 'report', 'findings', 'pandemic', 'information', 'identified', 'similar', 'regarding', 'one', 'related', 'date', 'among', 'data', 'likely', 'report', 'reporting', 'prior', 'including', 'suggesting', 'april', and 'experienced'. The words are in various shades of gray and sizes, creating a dynamic visual effect.

A word cloud of terms related to the COVID-19 pandemic. The words are arranged in a circular shape. The most prominent words are 'reported', 'information', 'pandemic', 'regarding', 'likely', 'among', 'date', 'reports', 'since', 'concerns', 'related', 'report', 'data', 'overall', 'including', 'reporting', 'frequently', 'concern', 'suggest', 'prior', 'majority', 'half', 'april', 'findings'. The words are in various sizes and orientations, with 'reported' being the largest and most central.

Out-degree

Betweenness

PageRank

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

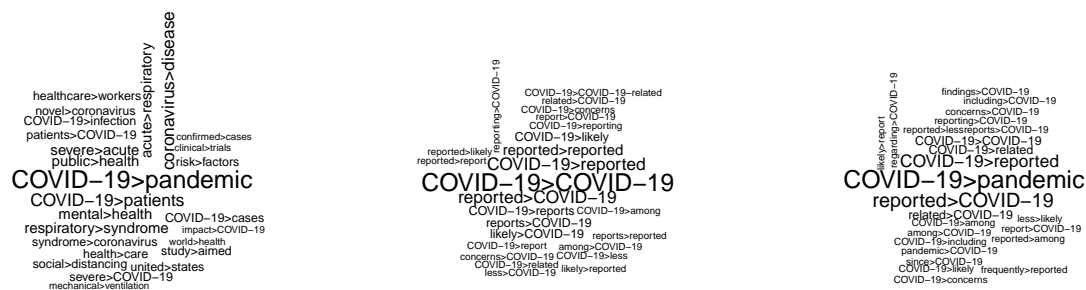


Figure 3: Top 25 bigrams (i.e., edges) by measure.

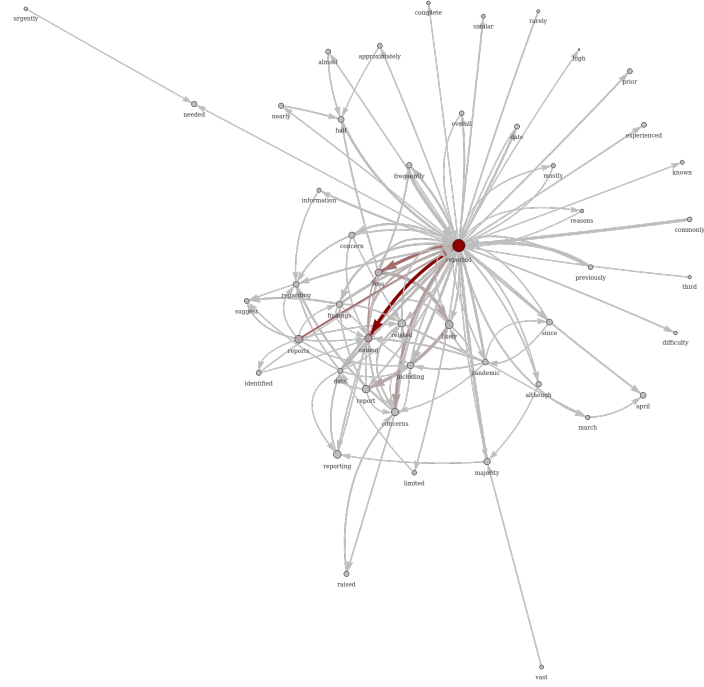


Figure 4: Filtered topic network (by weight). Layout based on Fruchterman-Reingold algorithm. Node size is proportional to topic-specific word probability provided by LDA. Edge width is proportional to topic-specific bigram weight provided by LDA2Net method. Node and edge color represent their betweenness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