*sustainability*

MDPI

*Article*

# Thematic Analysis as a New Culturomic Tool: The Social Media Coverage on COVID-19 Pandemic in Italy

**Massimo Aria [1,2,*], Corrado Cuccurullo [2,3], Luca D'Aniello [2,4], Michelangelo Misuraca [5] and Maria Spano [1,2]**

[1] Department of Economics and Statistics, University of Naples Federico II, 80126 Naples, Italy; maria.spano@unina.it
[2] K-Synth Spin-Off, University of Naples Federico II, 80126 Naples, Italy; corrado.cuccurullo@unicampania.it (C.C.); luca.daniello@unina.it (L.D.)
[3] Department of Economics, University of Campania Luigi Vanvitelli, 81043 Capua, Italy
[4] Department of Social Sciences, University of Naples Federico II, 80138 Naples, Italy
[5] Department of Business Administration and Law, University of Calabria, 87036 Arcavacata di Rende, Italy; michelangelo.misuraca@unical.it
* Correspondence: massimo.aria@unina.it; Tel.: +39-081-675187

**Abstract:** The COVID-19 pandemic influenced people's everyday lives because of the health emergency and the resulting socio-economic crisis. People use social media to share experiences and search for information about the disease more than before. This paper aims at analysing the discourse on COVID-19 developed in 2020 by Italian tweeters, creating a digital storytelling of the pandemic. Employing *thematic analysis*, an approach used in bibliometrics to highlight the conceptual structure of a research domain, different time slices have been described, bringing out the most discussed topics. The graphical mapping of these topics allowed obtaining an easily readable representation of the discourse, paving the way for novel uses of thematic analyses in social sciences.

**Keywords:** text analytics; topic detection; thematic mapping



**Citation:** Aria, M.; Cuccurullo, C.; D'Aniello, L.; Misuraca, M.; Spano, M. Thematic Analysis as a New Culturomic Tool: The Social Media Coverage on COVID-19 Pandemic in Italy. *Sustainability* **2022**, *14*, 3643. https://doi.org/10.3390/su14063643

Academic Editor: Haywantee Ramkissoon

Received: 8 February 2022
Accepted: 18 March 2022
Published: 20 March 2022



## 1. Introduction

Starting from December 2019, a new viral infection spread worldwide. The *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2), commonly known as COVID-19, appeared initially in Wuhan (China) and in a few weeks became a global pandemic [1]. Several theories about where the pathogen originated were proposed. There is a broad consensus in favour of a natural origin [2], as in other previous similar zoonoses [3], even if some characteristics are not fully explained by this hypothesis [4]. Italy was directly interested in the contagion at the end of January 2020, as one of the first countries in Europe together with France, but an early plight was experienced from the end of February 2020, when a small cluster of cases was detected in Lombardy, in Northern Italy. Due to the rapid spread of the disease and the occurrence of new clusters in different areas of the country, the Italian government decided to take drastic actions imposing a complete shutdown of all non-essential private and public activities (also including schools) in the regions effected by the first clusters on the 1st of March, and subsequently in the whole national territory on the 9th of March. In the following months, the lockdown was gradually eased, maintaining public safety and health measures such as social distancing, wearing masks and limiting numbers at public gatherings.

The exceptional nature of these measures, designed to curb further new cases of COVID-19 throughout the country, had a significant impact on people's feelings and behaviours. The fear for a newfound infection coming from Eastern Asia, and the initial lack of reliable sources about its riskiness, pushed people to search information online and to share their opinions on social media platforms. In recent years, the Internet has become an integral part of daily life, and the development of Web 2.0 tools contributed to

defining a new virtual environment to communicate and collaborate [5]. The widespread of social networks like Facebook, Reddit and Twitter allowed a rapid circulation of news, more than other traditional media, but also an increase of misinformation induced by the dissemination of fake contents [6]. The over-production of comments and opinions about COVID-19 origin and effects caused, alongside the spread of the infection, an actual worldwide infodemic [7,8].

In this paper, we aim at analysing the massive amount of comments shared by Italian users on Twitter during the first year of the pandemic, between February and December 2020, when narratives and moods related to the COVID-19 emerged and spread in social media. As stated above, during a time of social distancing and limited personal contacts, social media became a primary place to interact and helped people to remain connected, but at the same time facilitated the diffusion of conspiracy theories about COVID-19 and modified the social perception of a large section of the population. To investigate people's views and highlight the emerging social dynamics induced by the pandemic, we implemented a strategy based on the automatic detection of topics discussed by people, in the framework of a *culturomic* approach [9,10]. Typically, this approach is used to analyse human behaviour and discover cultural trends in large textual bodies derived from digital sources. According to [11], data-driven research can change the nature of social science through the application of new paradigms and methods. The use of social media content has become a powerful tool to evaluate and track macro-scale trends in human society, providing a deep understanding of its dynamics. Culturomic studies largely employ statistical techniques to identify linkages between linguistic patterns and socio-cultural phenomena by analysing huge quantities of digital data, mainly focusing on topic detection, semantic polarity and usage frequency in textual bodies. These tasks are often carried out considering time and geographical location, as well as the social and political environment since the context has a significant role in human behaviours and relations.

The problem of extracting the topics conveyed in a text written in natural language has been faced in different ways by scholars [12]. From an exploratory point of view, several techniques were applied. When there is not any preliminary information concerning the texts included in the textual body then their content in terms of words, or the aim is exploring the textual body without considering the conditional effect of text meta-data, several unsupervised approaches can be carried out to automatically mine the information embodied in texts [13]. Network-based methods, in particular, drew increasing attention since they allow exploring textual data by considering both single words as well as higher-order structures obtained by linking different words [14], preserving their semantic relatedness.

To enhance topic visualisation and interpretation and to track the topical trends over a given time-span, here we borrowed the methodological framework offered by the so-called *thematic analysis* [15], an approach broadly used in bibliometrics to explore the conceptual structure of a research domain. The use of such an approach allowed us to extract and automatically label the different topics related to COVID-19 and highlight the evolution of the discourse about the pandemic, offering the scientific community an interesting insight on this current issue as well as a strategy that can be easily implemented to investigate other social and cultural phenomena.

## 2. From Words to Figures and Back Again: A Statistical Approach to Topic Detection

Reading a substantial textual body to discover interesting patterns and trends is time-consuming. One of the primary problems in analysing natural language texts is that the content does not follow a predefined data model. Thus, it is challenging to handle texts directly from a quantitative standpoint, from the perspective of a knowledge discovery process [16]. Through a *parsing* routine, it is possible to scan a text and identify all the different words used to create its content. This procedure allows decomposing texts following a *bag of words* coding scheme, disregarding both grammatical relations and syntactic categories. The main result of parsing is the achievement of a list containing all the unique words used along the textual body, commonly known as "vocabulary".

A pre-processing stage is necessary to reduce the vocabulary's dimension and to filter non-informative words. According to the *vector space model* [17], each text can be then represented as a vector $\mathbf{d_i}$ in the space spanned by the $q$ words belonging to the vocabulary:

$$\mathbf{d_i} = \{w_{i1}, \ldots, w_{ij}, \ldots, w_{iq}\}, \quad (1)$$

where $w_{ij}$ represents the importance of the $j$-th word in the text-vector. This importance is usually measured by the *term frequency*—i.e., the number of word occurrences in the text—but other weighting methods can also be applied. It is possible, for example, to use binary weights and assign 1 to words appearing in the text and 0 to words not appearing in the text. More structured methods can be used instead to take into account the discriminative power of each word within the texts, as in the *term-frequency/inverse document-frequency* weighting method [18]. To statistically analyse the textual body, by now represented as a set of structured data, all the text-vectors can be juxtaposed and arranged in a matrix $\mathbf{D}$ with $p$ rows (the texts included in the textual body) and $q$ columns (the words listed in the vocabulary). Because of the bag of words scheme, in $\mathbf{D}$ there are not any hints about the words' context of use. This contextual information can be partially recovered by transmuting a binary version of matrix $\mathbf{D}$ in a $q$-dimensional integer matrix $\mathbf{A} = \mathbf{D}^{\mathsf{T}}\mathbf{D}$, whose generic element $a_{jj'}$ $(j \neq j')$ represents the number of texts in which two words $j$ and $j'$ co-occur (i.e., they both appear in the texts). The $a_{jj}$ elements on the principal diagonal of $\mathbf{A}$ count the total number of texts containing the single word $j$. This latter matrix can also be seen as an adjacency matrix and diagrammatically represented as a graph $\mathbb{G}$. Each vertex of $\mathbb{G}$ symbolises a word used in the textual body, while each edge between two vertices expresses the co-occurrence of two distinct words.

Both matrices $\mathbf{D}$ and $\mathbf{A}$ can be used to explore and visualise the most relevant topics conveyed in the textual body. A topic is what discourse is about [19]. Generally, it is possible to represent a topic by considering the words (or keywords) it is frequently associated with. In this sense, the quantitative formalisation of a topic expressing given concepts or themes in a text can be derived from the definition of words' cluster [20], since words' clusters should have a high internal cohesion as well as a high external separation [21].

Several approaches have been proposed in the literature for topic detection starting from these textual data representations. Factorial-based methods (e.g., *lexical correspondence analysis* [22,23], *latent semantic analysis* [24]) aim at representing the most relevant information by considering linear combinations of the original words and visualising these latent association structures onto factorial maps. The main drawback of the factorial-based approaches is that the topics are quite difficult to interpret. To overcome this problem, they are often performed together with a clustering step in a sequential manner as in *tandem analysis* [25]. The so-called topic modelling is a well-known alternative offering a probabilistic viewpoint of the topic detection problem. It includes a family of generative statistical models used to discover semantic patterns reflecting the underlying (latent) topics. Topic models consider a probabilistic decomposition of the matrix $\mathbf{D}$, where the probability that a word occurs in a text is modelled as a mixture of conditionally independent multinomial distributions. From its most popular formalisation (i.e., *latent Dirichlet allocation* [26], where topic distribution is drawn by a Dirichlet distribution) researchers have derived some interesting proposals directly applied on the matrix $\mathbf{A}$ (e.g., *biterm topic model* [27]). One of the main limitations is that topic modelling requires setting a priori the number of topics to be detected. Even if some methods for automatically setting the parameters have been proposed in the literature (e.g., [28]), there is a lack of shared solutions to cope with this issue. Network-based approaches have the advantage of automatically determining the appropriate number of topics. The relationships between words belonging to a text are represented as a graph [29,30] derived from the matrix $\mathbf{A}$. The detection of the most relevant topics is obtained by performing on the matrix $\mathbf{A}$ community detection procedures, to highlight if in the network there are groups of words sharing common characteristics and/or playing similar roles within the graph. Several proposals based on community detection algorithms (e.g., *KeyGraph* [14] , *ClusTop* [31], the strategy of [32]) were produced.

Several studies on COVID-19 discourse on social media have been recently published. Many of these employed topic modelling to detect the main debated topics concerning the infection spread and the management of the subsequent health emergency in several countries (e.g., [33] for Spain, [34] for the United States, [35] for North America). Some other studies focused on network-based analyses to explore COVID-19 discourse (e.g., [36,37]). Among the different alternatives, network-based approaches seem a more suitable solution for topic analysis for two main reasons. First, the majority of community detection algorithms allow identifying topics without prior setting of their number, offering a completely automatic approach. Second, the topics discovered through the analysis can be represented as clusters/sub-networks of words and easily interpreted, also considering the strength of their relationships. Nevertheless, when a huge collection of texts is investigated, the usefulness of network visualisation tools could be limited because the number of words and topics to interpret rapidly increase.

An interesting strategy is offered by *thematic analysis*, an approach used in bibliometrics to explore the conceptual structure of a given research sphere [15]. This approach allows obtaining graphical representations that automatically summarise the main topics of a textual body. Moreover, the extracted topics could be characterised with respect to their structure and their role in the whole network.

*The Thematic Analysis Approach*

Starting from a matrix like **A**, it has been showed showed that the co-occurrence of two words belonging to a text can be expressed in terms of *association strength* [38]:

$$AS_{jj'} = \frac{a_{jj'}}{\hat{a}_{jj}\hat{a}_{j'j'}}, \tag{2}$$

where $a_{jj'}$ is the observed number of co-occurrences of words $j$ and $j'$, while $\hat{a}_{jj}$ and $\hat{a}_{j'j'}$ are the expected numbers of occurrences of $j$ and $j'$ under the assumption they are statistically independent. This measure is akin to cosine similarity used in text mining to evaluate the semantic relatedness of two words [39], with a primary difference concerning its probabilistic nature. The association strength is a normalised measure: a 0 value means that the two words never co-occur and a 1 value means that the words co-occur in all the texts. Performing on the adjacency matrix $\mathbf{A}^{\star}$—containing the association strength values between the different words—a *community detection* procedure [40], it is possible to identify $K$ groups of words representing different topics. Community detection can be seen as a method similar to clustering that allows grouping vertices in highly cohesive sub-graphs. In particular, the *Louvain algorithm* [41] showed high effectiveness with respect to other competing proposals [42].

The topics obtained through the community detection can be projected in a so-called *strategic diagram* [43], obtaining a thematic mapping of the surveyed domain, in accordance with two measures known as *Callon centrality* and *Callon density*:

$$CC_k = 10 \times \sum_{j \in k, h \in k'} AS_{jh} \qquad CD_k = 100 \times \sum_{j,j' \in k} \frac{AS_{jj'}}{n_k}, \tag{3}$$

where $AS_{jh}$ is the association strength between two words $j$ and $h$ belonging to two distinct topic $k$ and $k'$, $AS_{jj'}$ is the association strength between a couple of words $j$ and $j'$ belonging to a given topic $k$, and $n_k$ is the total number of words belonging to $k$. These measures express the role of a topic in organising the domain's conceptual structure. Callon centrality can be read as the relevance of the topic in the entire research domain, while Callon density can be read as a measure of the topic's development.

The strategic diagram (see Figure 1) allows highlighting four different kinds of topics, depending on the quadrant in which they are mapped:

- higher values of centrality and density define the *hot topics*, well developed and relevant for structuring the conceptual framework of the domain;

- higher values of centrality and lower values of density define the *basic topics*, significant for the domain and cross-cutting to its different areas;
- lower values of centrality and density define *peripheral topics*, not fully developed or marginally interesting for the domain;
- lower values of centrality and higher values of density define *niche topics*, strongly developed but still marginal for the domain under investigation.

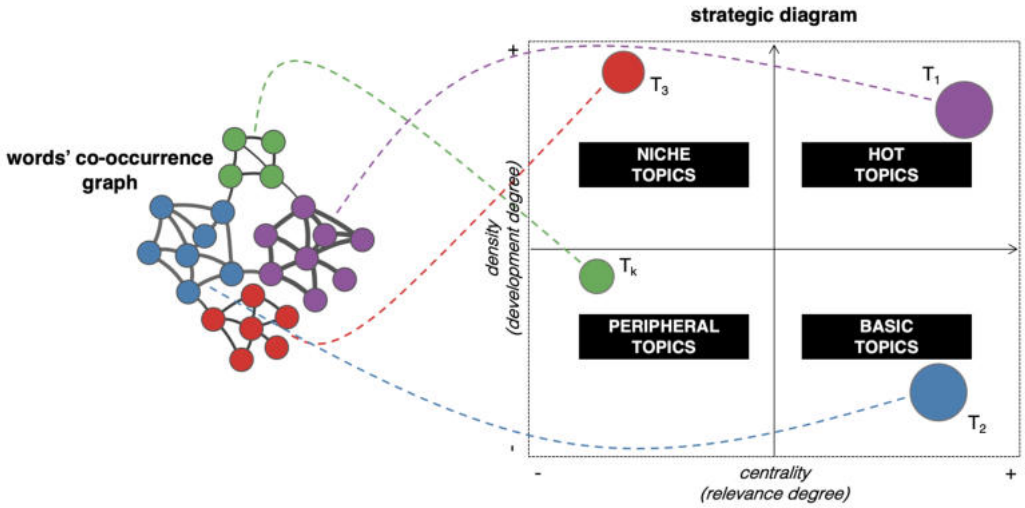


**Figure 1.** Construction of the thematic map from the words' co-occurrence graph.

Moreover, it is possible to express the complexity of each topic by scaling its representation on the diagram in accordance with the number of related words. To facilitate the reading of the map, each topic can be labelled with the associated most occurring keywords.

The above-described approach can be easily implemented in R or python, by expressing the following illustrative script in the proper programming language:

**Script:** Thematic Analysis

**Data:** a textual body
**Result:** a set of thematic maps

**begin**
  divide the collection into $t$ time slices
  **foreach** *time slice* **do**
    - pre-treat the textual body and transform each text in a vector
    - build a co-occurrence matrix $\mathbf{A}$ from the set of vectors
    - weight the elements of $\mathbf{A}$ to obtain an $AS$ matrix $\mathbf{A}^{\star}$
    - perform on $\mathbf{A}^{\star}$ the *Louvain* community detection
    - associate the $k$ detected communities to $k$ different topics
    **foreach** *topic k* **do**
      - calculate $CC_k$ and $CD_k$
      - plot the topic on the $t$ strategic diagram
    **end**
  **end**
**end**

Jointly analysing the conceptual structure of different temporal sub-periods, it is possible to shape the topical evolution of the domain, revealing the trajectories of the different topics across time (see Figure 2).

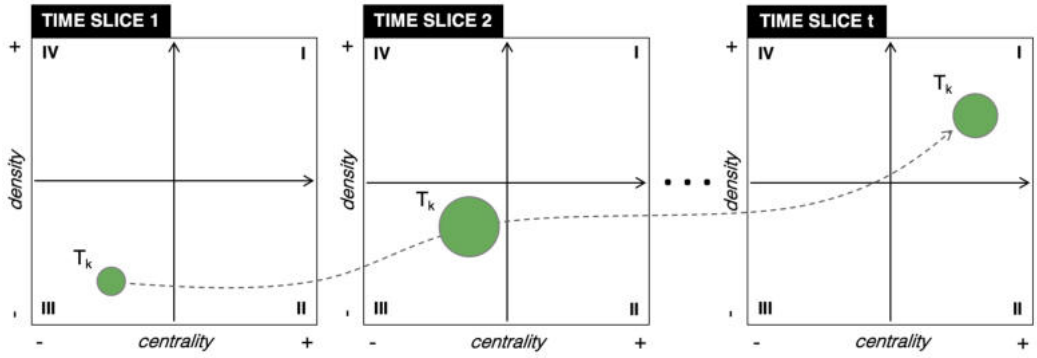


**Figure 2.** Topical evolution across different sub-period time slices.

In the following, the above-described strategy is implemented to study the coverage on Twitter of the COVID-19 pandemic in Italy during 2020, highlighting the set of topics discussed by the users in different sub-periods of the year, taking into account the different stages of the health, social and economic emergency.

## 3. Analysis Setup and Main Findings of the Study

To extract the topics concerning the COVID-19 pandemic and carry out the thematic analysis, we considered the large set of data offered by *40wita* [44], the most extensive repository holding tweets written in Italian about the COVID-19. This textual body is part of a more comprehensive project developed by the University of Turin, aiming at building a massive archive of tweets written in Italian [45]. The tweets were selected from the primary collection using a list of 43 different keywords that include both terms related to the disease itself (e.g., *covid19*, *coronavirus*) and other terms and hashtags popular in Italy during the emergency period (e.g., *#iorestoacasa* $\longrightarrow$ "I stay at home", *#andratuttobene* $\longrightarrow$ "everything will be fine"). We decided to focus, in particular, on the tweets published between February and December 2020, the period in which several decisions to contain the diffusion of the disease were taken at a central level (national government) as well as a local level (regional governments), including a strict lockdown of the whole Italian territory. In this way, we retrieved a set of 4,824,576 unique tweets (without retweets) accompanied by some meta-data, like the username, the publishing date, the retweet and the like count.

Data pre-processing of tweets was performed in two steps. Firstly, we stripped URLs, usernames, hashtags and emoticons, and then we normalised the text by removing special characters and any delimiter different from blank. Secondly, we lexicalised most frequent bigrams (i.e., couples of linked words like collocations were considered as unique entries of the vocabulary) and filtered out Italian stop-words (e.g., preposition, articles) and a list of context stop-words (e.g., words already used in the query). Figure 3 shows the pre-treating pipeline used to prepare the dataset before the analysis.

Aiming to discover the most discussed topics on Twitter and track the evolution of the discourse about COVID-19 across time, we partitioned the overall time into four different sub-periods, according to some regulatory measures adopted by the Italian government during 2020. The first time slice ($T_1$) includes the period ranging from the 1st of February, the day after the discovery of two confirmed cases of COVID-19 in Italy, to the 9th of March, the enactment date of the Prime Minister's decree providing for the general lockdown, with school and non-essential business closures, social distancing and a ban on public gathering. The second time slice ($T_2$) includes the period ranging from the 10th of March, the beginning of the national quarantine (later called *Phase-1* by the government), to the 4th of May, the enactment date of the Prime Minister's decree providing for the gradual easing of the reopening of manufacturing and other activities. The third time slice ($T_3$) covers the period ranging from the 5th of May (*Phase-2*) to the 31st of August, including the end of the 97-day lockdown and the reopening of public spaces (*Phase-3*, started on the 15th of June) and the months in which Italians typically take summer holidays. The fourth time slice

($T_4$), finally, covers the period ranging from the 1st of September to the 31st of December, including the gradual resumption of activities after the summer holidays, the enactment date (on the 6th of November) of the Prime Minister's decree establishing different rules and restrictions at a regional level in accordance with the spread of the disease and the beginning of the vaccination campaign in Italy.

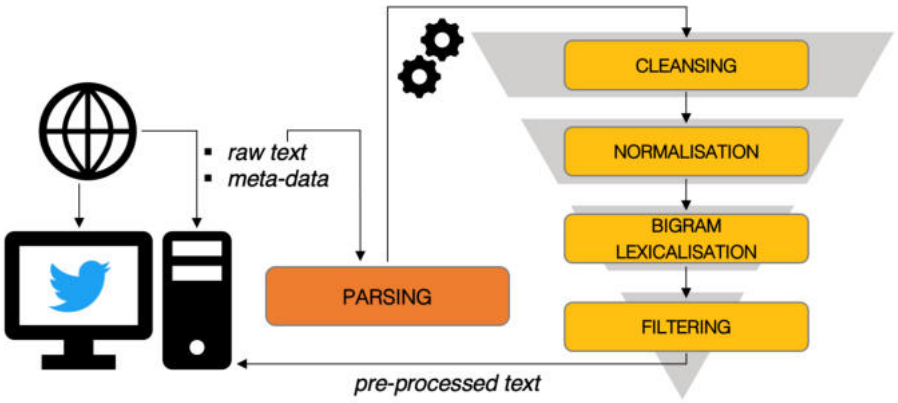


**Figure 3.** Diagrammatic scheme of the pre-treating pipeline.

At the end of the pre-treatment process, taking into account the publication date of the different posts, we obtained a set of 4,046,305 tweets on COVID-19 over the 4 considered sub-periods. In Table 1, some descriptive figures are reported.

**Table 1.** Descriptive statistics on the tweets posted in the four sub-periods $T_1$–$T_4$.

| Time Slice | N. of Tweets | Avg. Tweets per Day | Relative Std. Deviation |
|---|---|---|---|
| $T_1$: 1/2/2020–9/3/2020 | 517,508 | 13,619 | 0.927 |
| $T_2$: 10/3/2020–4/5/2020 | 1,744,975 | 30,614 | 0.314 |
| $T_3$: 5/5/2020–31/8/2020 | 845,484 | 7617 | 0.393 |
| $T_4$: 1/9/2020–31/12/2020 | 938,338 | 7629 | 0.381 |

We observed a more significant amount of tweets posted in the second time slice, covering the so-called Phase-1 of the lockdown imposed in the whole country by the Italian government, with the lowest daily posting variability. Conversely, we observed the lowest number of tweets in the first time slice but the highest daily posting variability. Figure 4 reports, in more detail, the day-wise distribution of tweets over the four sub-periods as well as the distribution of the COVID-19 daily new cases in Italy in the same period.

We observed a right-skewed distribution for the tweets posted in the analysed periods, with a peak of 51,490 tweets on the first day of Phase-1. Conversely, we observed a left-skewed distribution for the COVID-19 daily new cases, with a peak of 40,920 new cases in the middle of November, a week after the enactment of the Prime Minister's decree aiming at clamping down the spread of the contagion at a regional level, through a colour-coded risk system based on a set of 21 parameters [46].

In order to perform the thematic analysis on the four sub-periods, we kept the first 1500 most occurring bigrams and built a matrix **D** with 4,046,305 rows and 1501 columns, considering for each post the publication date and the presence/absence of the different bi-grams. From the latter matrix **D**, we derived 4 co-occurrence matrices $\mathbf{A}_h^\star$ (with $h = 1, 2, 3, 4$), with 1500 rows and columns, for the distinct analysed sub-periods. The analysis was carried out by using an appropriately adapted version of the routines included in the R package *bibliometrix* [47], which allows mapping the topics discussed in each sub-period and tracking the topical evolution across time [48].

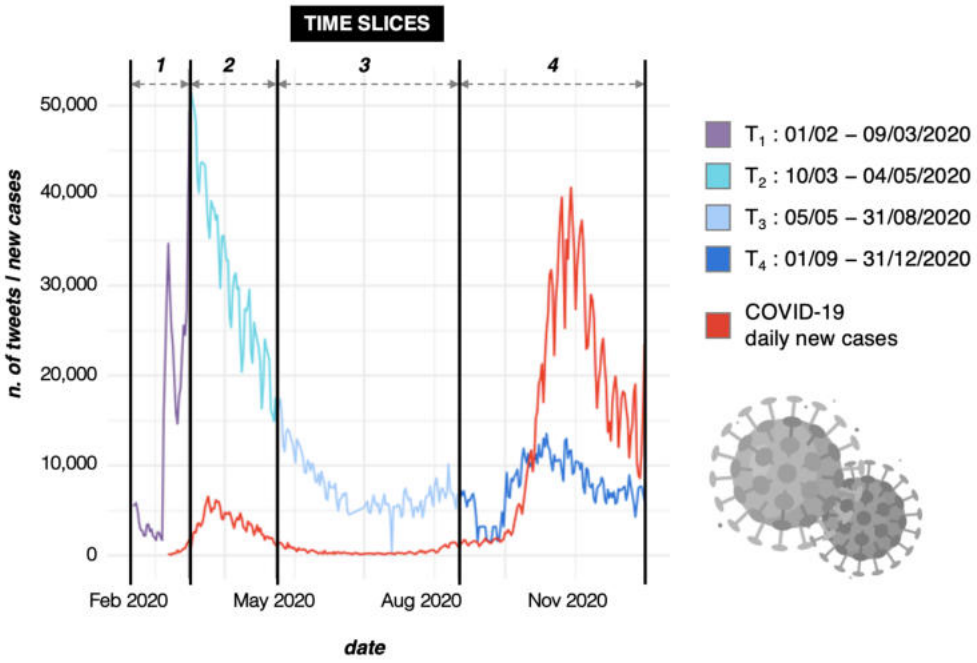


**Figure 4.** Daily tweets and COVID-19 new cases in Italy (February–December 2020).

*Mapping COVID-19 Topics and Tracking Their Evolution over 2020*

The first thematic map, shown in Figure 5, highlights the discourse on COVID-19 developed from the occurrence of the earlier case in Italy on the 1st of February (two Chinese tourists were hospitalised in Rome on the 30th of January) to the beginning of the strict lockdown, known as Phase-1, on the 9th of March.

We observed a prevalence of basic and niche topics (second and fourth quadrants, respectively). The basic topics covered the debate concerning the first news about the infection in Italy (e.g., *primo caso, pronto soccorso, caso sospetto* ⟶ "first case, emergency room, suspected case" pertains to the discovery of the potential patient zero in Italy; *terapia intensiva, nuovi casi, emilia romagna* ⟶ "critical care, new cases, Emilia Romagna" pertains to the spread of the infection in Northern Italy) and around the world (*primo morto, honk kong, prima vittima* ⟶ "first death, Honk Kong, first victim" pertains to the first death caused by COVID-19 in the world). Moreover, we observed some issues related to the actions taken by the Italian government in response to the infection (e.g., *zona rossa, scuole chiuse, tutta italia* ⟶ "red zone, closed schools, all over Italy" pertains to the shutdown and the closure of schools; *posti letto, sistema sanitario, operatori sanitari* ⟶ "beds, health-care system, health professionals" pertains to the teething troubles of the Italian health-care system in facing the contagion). Concerning the niche themes, we detected the first attempts of contrasting the COVID-19 infection (e.g., *australiano supera, vaccino australiano* ⟶ "Australian passed (the tests), Australian vaccine" pertains to the trials of a vaccine against the coronavirus conducted by the University of Queensland; *pazienti gravi, farmaco somministrato, anti artrite* ⟶ "seriously ill patients, administered drug, anti-arthritis" pertains to the initial use of anti-inflammatory therapy to treat patients in intensive care) as well as the polemics fuelled by politicians, pundits and other alleged experts (e.g., *topi vivi, luca zaia, cinesi mangiano* ⟶ "live mice, Luca Zaia, Chinese eat" pertains to the racist comments of the Veneto region governor about the dietary habits in China; *ilaria capua, meno letale, unica cura* ⟶ "Ilaria Capua, less-lethal, only cure" pertains to the opinion of a famed Italian virologist about the early cases; *vasco rossi, tampone costa, costa dollari* ⟶ "Vasco Rossi, swab expensive, costs dollars" pertains to the comments of a famous Italian singer about the influence of the U.S. health-care system high costs in effectively detecting the COVID-19 active cases).

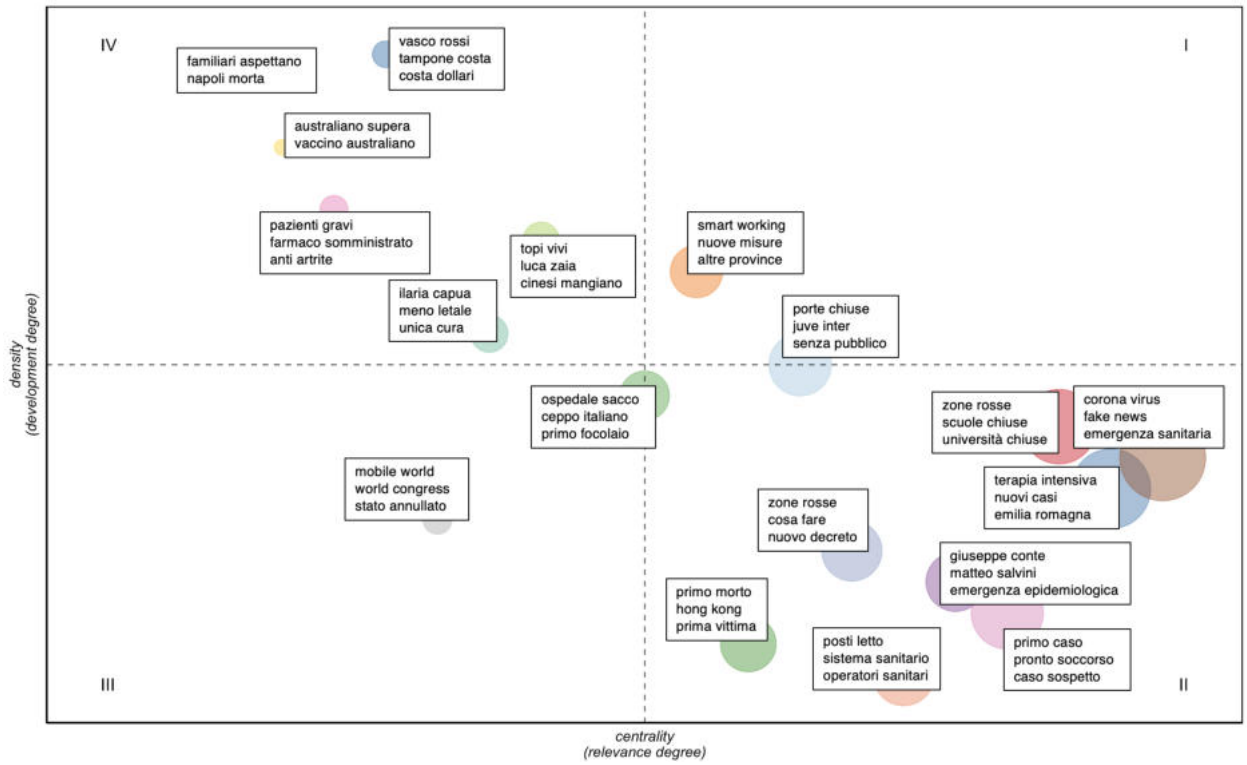


**Figure 5.** Thematic map of topics discussed in the period 1/2/2020–9/3/2020.

The second thematic map, shown in Figure 6, highlights the discourse across Phase-1, from the 10th of March to the 4th of May. Due to the spread of the infection (on the 11th of March, WHO Director-General announced that COVID-19 could be characterised as a pandemic), the uncertainty caused by the lack of reliable information and the stress caused by the extended quarantine, the emerging thematic structure concerning COVID-19 was more complex compared to the previous period.

First of all, the issues related to the health-care system in Italy became hot topics (e.g., *sanità lombarda, chiede aiuto, medici cubani* ⟶ "Lombard health-care, ask for help, Cuban doctors" pertains to the discussion about the initial deficiencies of the local governments in facing the health emergency and the support given by other countries; *medici senza frontiere, breve storia* ⟶ "doctors without borders, brief history" pertains to the support given by the Italian-French NGO *Medici senza frontiere* to the public health-care system). As in the previous period, basic topics were focused on the national government's actions (e.g., *giuseppe conte, emergenza epidemiologica, decreto legge* ⟶ "Giuseppe Conte, epidemiological emergency, decree-law" pertains to the regulatory action carried out to cope the emergency; *nuovi casi, terapia intensiva, protezione civile* ⟶ "new cases, critical care, civil protection" pertains to the support given by Italian Civil Protection to enhance medical facilities). It is interesting to note that the discussion about the use of anti-inflammatory therapy, which appeared as a niche topic in the first time slice, moved to this quadrant, becoming a basic topic in the discourse (*anti artrite, ospedali italiani, protocollo nazionale* ⟶ "anti-arthritis, Italian hospitals, national protocol"). In this period, together with the news about international tension (e.g., *repubblica ceca, mascherine inviate, sequestrato migliaia* ⟶ "Czech Republic, face mask shipped, seized thousands" pertains the seizure of masks sent by China for Italy's beleaguered hospitals; *fuori controllo, prigionieri politici, nasrin sotoudeh* ⟶ "out of control, political prisoner, Nasrin Sotoudeh" pertains to the call for an immediate release of political prisoners and detainees at risk of exposure in Iran), appeared as niche topics some issues related to the social impact of the lockdown in Italy (e.g., *bela madunina, commuove milano, milano suonando* ⟶ "Bela Madunina, touched Milan performing" pertains to the habit of Italians, during the lockdown, of playing and singing on the buildings' balconies

in an effort to boost morale; *salute pubblica, senso civico, assolutamente attenerci* ⟶ "public health, citizenship, adhere strictly" pertains to the appeals for caution of people during the quarantine period imposed by the government).

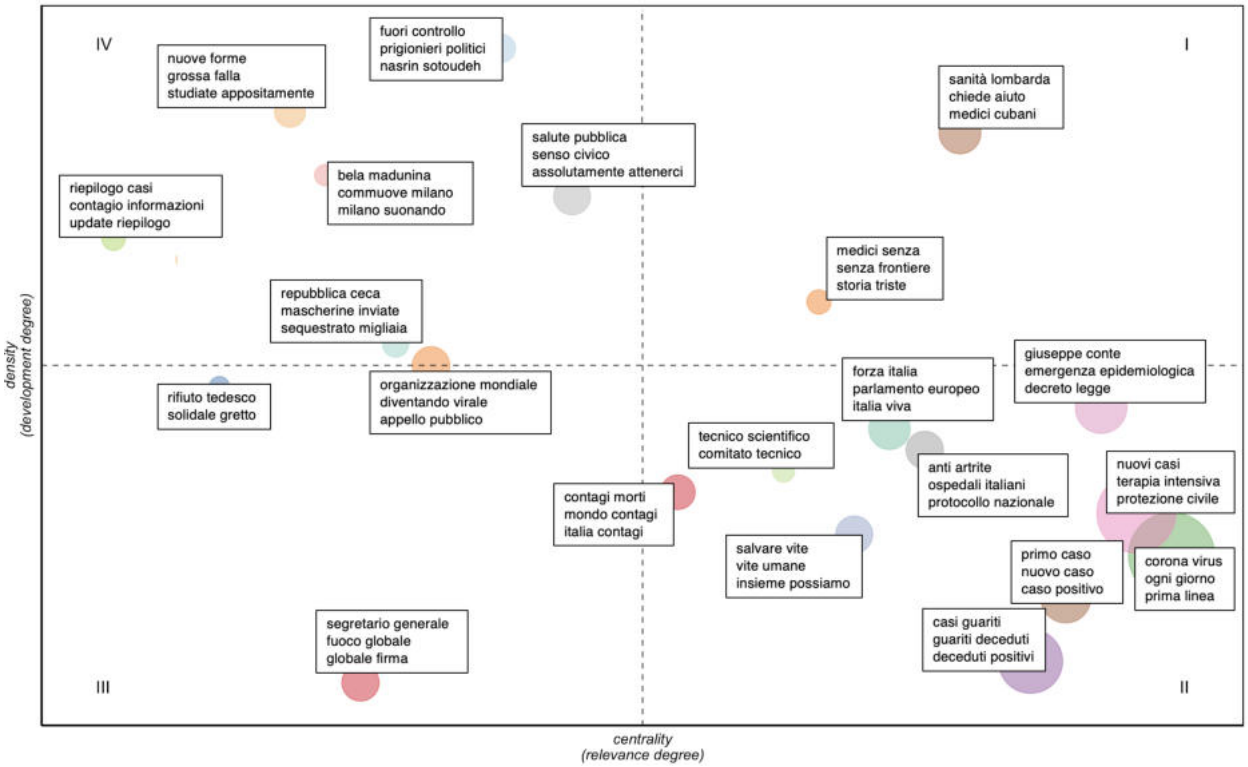


**Figure 6.** Thematic map of topics discussed in the period 10/3/2020–4/5/2020.

The third thematic map, shown in Figure 7, highlights the discourse across the Phase-2 and Phase-3 of the lockdown until the end of summer holidays in Italy, from the 5th of May to the 31st of August. Similarly to the second time slice, the thematic structure was articulated in several distinct topics, even if the number of tweets related to this period was about a half of the tweets related to the previous one (as reported above in Table 1).

Hot topics of this period were the prevention of cases among health professionals (*test sierologici, task force, personale sanitario* ⟶ "serology tests, . . . , medical staff" pertains the use of SARS-CoV-2 antibody tests to detect possible COVID-19 active cases among doctors and nurses). Basic topics still conveyed the news about the contagion (*nuovi casi, nuovi contagi, nuovi potitivi* ⟶ "new cases, new infections, new positives"; *nuovi decessi, altri nuovi, gran bretagna* ⟶ "new deaths, other new, Great Britain" pertains to the rise of the contagion in Great Britain) and the debate about the regulatory framework (e.g., *seconda ondata, senza mascherina, linee guida* ⟶ "second wave, without mask, guidelines" pertains to the risk of a new wave after the summer period and the requests of limiting the use of face masks outdoors; *contract tracing, fase analisi, tracing scenario* ⟶ ". . . , screening step, . . . " pertains to the use of a digital contact tracing app in Italy to prevent a new outbreak). Niche topics of the period were related to the contrasting opinions and initiatives related to the pandemic, its effects from an economic viewpoint, and how to deal with these effects (e.g., *truffa biologica, farcele domande* ⟶ "biological fraud, asking us questions" pertains to some conspiracy theories about COVID-19; *appena firmato, educazione opportunità, offrire educazione* ⟶ "just signed, education opportunity, offer education" pertains to the campaign of an Italian NGO to support poor children during the pandemic on the educational side; *aziende inquinanti, vogliono usare, continuare inquinare* ⟶ "polluting industries, plan to use, continue polluting" pertains to the campaign launched by an international community of citizens to ask for an after-COVID Green Renaissance to European leaders).

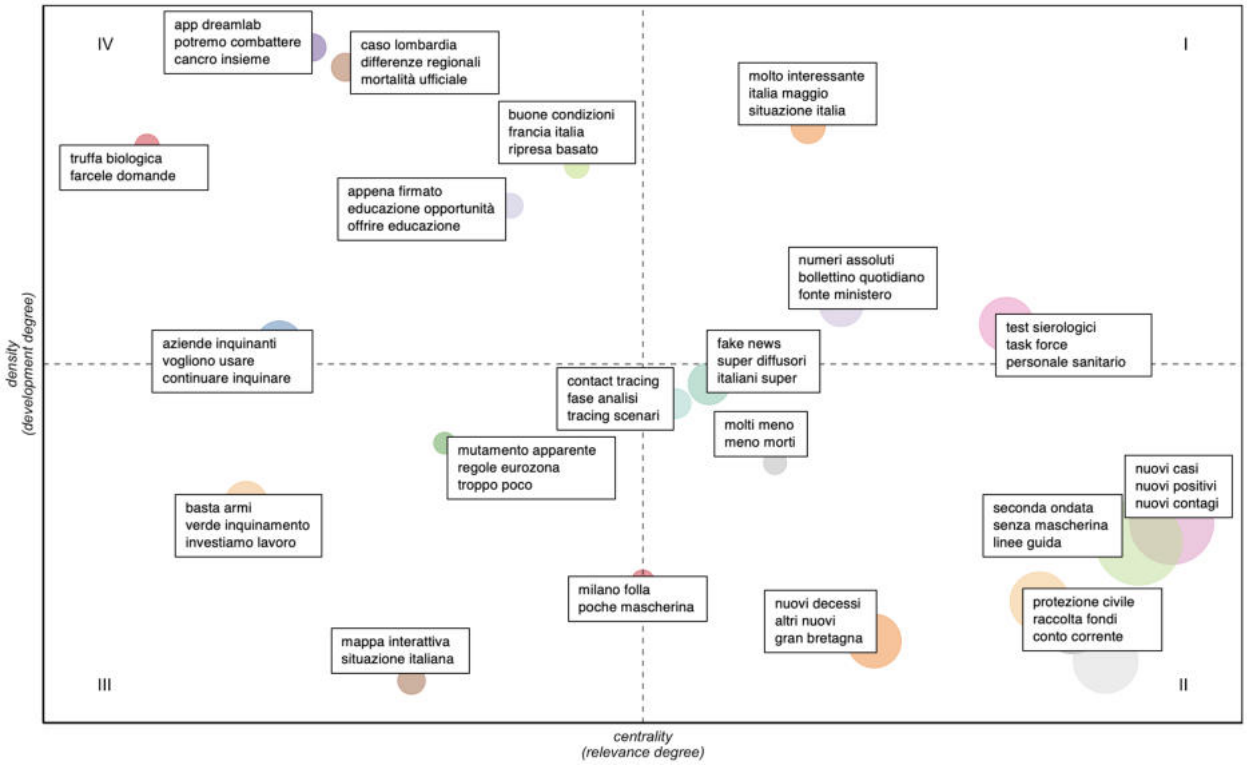


**Figure 7.** Thematic map of topics discussed in the period 5/5/2020–31/8/2020.

The last map, shown in Figure 8, highlights the discourse on COVID-19 that developed during the fourth quarter of 2020, from the 1st of September to the 31st of December.

The hot topics of the period conveyed issues related to the pharmaceutical companies developing a vaccine against COVID-19 (*grandi farmaceutiche, limitano accesso, sospendere brevetti* ⟶ "big pharmaceutical, limit access, suspend patents" pertains to the debate about the production of vaccine in the different countries, specifically less developed ones) and the renewed cases after the summer period (e.g., *infermiere stroncato, luciano quaglieri, filippo neri* ⟶ "nurse struck, Luciano Quaglieri, Filippo Neri" pertains to the death caused by COVID-19 of a nurse working at a Rome's hospital; *jerry scotti, carlo conti, controllo medico* ⟶ "Jerry Scotti, Carlo Conti, medical check" pertains news about Italian television presenters affected by the virus). The news about the contagion as well as issues related to the new actions of the government appeared again in the basic topics (*vaccino anti, zona rossa, seconda ondata* ⟶ "vaccine anti, red zone, second wave" pertains to the different measures implemented at a regional level following the extent of the contagion). Niche topics of the period were related to some conspiracy issues (e.g., *padre livio, progetto criminale, élites mondiali* ⟶ "Father Livio, crime design, world elite" pertains to the polemics fuelled by the director of a Catholic radio station concerning a view of COVID-19 as a divine punishment; *scopi politici, impostori usano, tamponi eseguiti* ⟶ "political ends, impostors use, performed swabs" pertains to the debate about the number of cases fuelled by COVID-19 sceptics). Interestingly, both in peripheral and niche topics appeared issues related to the minks' culling in more than 200 Danish mink farms in late 2020, offering a different standpoint. In the first case, the topic is more related to the story itself (*visoni positivi, diversi paesi, salute pubblica* ⟶ "positive minks, several countries, public health"), whereas in the second case, the topic is more related to animal rights and intensive farming (*essere crudeli, allevamenti visoni, chiusi definitivamente* ⟶ "be cruel, mink farms, permanently closed").

In the appendix, all the topics identified in the different time slices are reported in Tables A1–A4, specifying in which quadrant of the thematic map they appeared and translating in English the corresponding most important keywords.

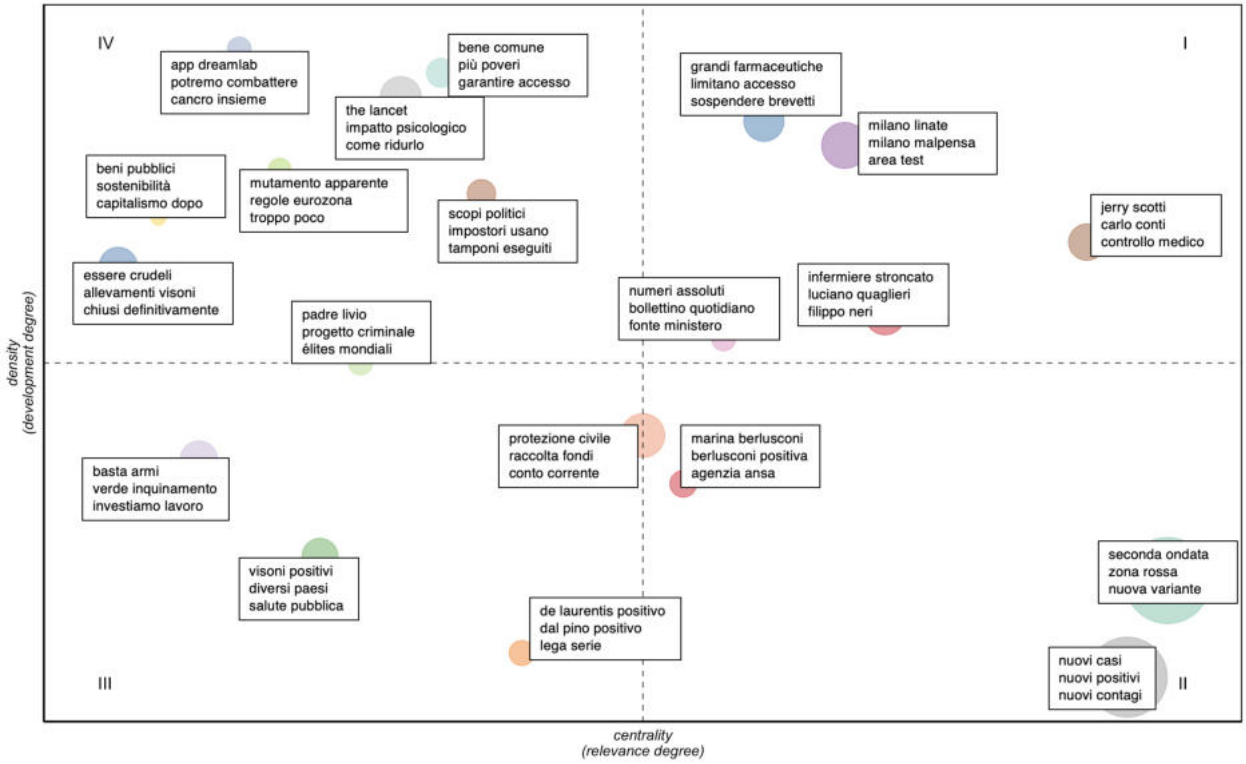


**Figure 8.** Thematic map of topics discussed in the period 1/9/2020–31/12/2020.

## 4. Discussion

Analysing the COVID-19 discourse on Twitter provided an opportunity to understand the information that the Italian public was exposed to. Interestingly, from a quantitative viewpoint, the number of tweets were very high during the first wave of COVID-19, with a peak at the beginning of the strict nationwide lockdown, whereas during the second wave (declining at the beginning of 2021) the number of tweets was relatively lower. A possible explanation relies on the fact that the analysed dataset was retrieved with a single set of keywords, not taking into account the events that occurred later, such as the development of COVID-19 vaccines and the early inoculation of health professionals in Italy. Nevertheless, the extended time span and the massive number of comments analysed in the study allowed highlighting the most discussed themes during the different phases of the first year of the pandemic.

From the thematic analysis on four distinct sub-periods, covering the evolution of the contagion and the consequent actions pursued by the Italian government to limit the diffusion of the disease, several key topics conveyed in the public discussions on COVID-19 appeared across time. Great attention was exceptionally devoted, during all the periods, to the behaviour of the government and its ability to respond in an effective way to the spread of the contagion, also in light of the ongoing political crisis caused by the changes in the governing coalition at the end of 2019 [49]. The discourse about COVID-19 was also primarily influenced by the daily reports released by the National Institute of Health and the Ministry of Health, inducing a raising fear and anxiety among people forced to change their everyday habits in a few months [50]. We noticed a certain stability in the basic topics, with a high relevance degree but a low development degree. In contrast, the niche topics changed from one period to another one, following the evolution of the pandemic and the occurrence of specific events, not enhancing their relevance in the thematic structure. We noticed a few hot topics concerning conspiracy theories and fake news about the contagion, but this lower detection may be related to the nature of the analysed social media (characterised by short texts) and its usual audience.

The study's main findings can be compared with the results obtained in previous studies concerning COVID-19 coverage on social media. In [51], a set of tweets written in Italian and posted in the last week of February 2020 has been analysed to reconstruct the meanings of the hashtag and the content related to the early steps of the outbreak and highlight the sentiment and opinions of the Italians. Users revealed to be worried, at first, by the news coming from the northern regions of Italy, where COVID-19 started its advance in the country and then frightened by the possibility of being physically locked up. In [52], the period of the stricter lockdown (coinciding with the second time slice analysed in the present study) was considered, highlighting the main buzzing topics on Twitter concerning the pandemic as well as the broader Italian socio-political framework. The topics emerging from the latter study (e.g., the polemics about the arguable claims of the Lombard Regional Councillor for Health) were not precisely identified by the thematic analysis here implemented because of the different set of keywords used to select the textual body and the focus on specific days rather than a longer time-span. Thanks to the availability of free online repositories of tweets written in different languages (for tweets on COVID-19 written in English see, for example, https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset, (accessed on 7 February 2022) [37]; for tweets written in French see, for example, https://github.com/calciu/COVID19-LockdownFr, (accessed on 7 February 2022); for tweets written in Spanish see, for example, https://data.mendeley.com/datasets/nv8k69y59d/2, (accessed on 7 February 2022) ), it is possible to apply the proposed strategy to analyse the COVID-19 discourse in countries other than Italy. Clearly, different languages may require a different pre-treatment to clean and normalise the texts, but the subsequent steps aiming at highlighting the topics and mapping them on a strategic diagram would perform in a similar way.

The use of thematic analysis offered different advantages with respect to other techniques. Above all, the co-occurrence data matrix behind the approach allows reconstructing the context of use of the different words encompassed in the textual body. Differently from factorial-based analyses and probabilistic-based analyses relying on the decomposition of a *documents* × *words* matrix, the detected topic structures are easily interpretable because each word retain its original meaning and at the same time the relations between words are based on their use in a specific context, viewing topics as higher-order structures embodied in the text collection. Moreover, the topic detection step is completely data-driven and automatic, not requiring the prior setting of any parameter—such as the number of topics—and not making hard the choice of the optimal partition through the interpretation of complex diagnostics for model selection. Another key characteristic representing a major advantage of the proposed approach is the possibility of mapping the different topics on a metric space, improving the readability of the results in comparison with classic network-based analyses. Computing the centrality and density of each topic and using these metrics to plot them in the thematic map allows discriminating the relevance and the development of the discussed issues, highlighting in a given time slice the most debated topics as well as those marginal topics that, in any case, contributed in building the discourse.

The results of thematic analyses can also be explained through the lens of *framing theory* [53]. Framing tries to depict how individuals, groups and societies perceive and communicate a given issue or event. Several authors (e.g., [54,55]) claimed that in recent years social media contents have significant influence over the ways reality is represented, and they can selectively direct information in the audience's minds. COVID-19 offered, from this point of view, an interesting research theme, with different studies applying framings of collections of Twitter posts [56,57] or of collections of newspapers articles [58,59]. In this context, a thematic analysis could be a valuable tool able to automatically highlight the topics linked to the different frames, helping researchers in exploring and visualising a massive quantity of textual data.

## 5. Conclusions and Final Remarks

The study presented in this paper highlighted how the COVID-19 discourse on social media developed in Italy during the first year of the pandemic through the analysis of eleven months of comments posted on Twitter in Italian from February to December 2020. As reported by several previous studies, the spread of the viral contagion was coupled with another peculiar contagion caused by the massive use of social media—known as infodemic—with an overproduction of unverified and inaccurate information, as well as of fake news [60,61]. In this sense, misinformation often became disinformation, so some authors preferred to use the term *disinfodemic* [62]. The thematic analysis allowed us to discover the main topics discussed in Italy, distinguishing different topical categories on the basis of their relevance and their development in the discourse. As in other approaches aiming at detecting the topics discussed in a given textual body through a co-occurrence network analysis, the strategy here presented is totally automatic and data-driven, leveraging a community detection procedure for identifying the optimal number of topics. The major advantage, in comparison with alternative solutions, is the visualisation of the topics and the possibility of highlighting their role in the analysed discourse in a synchronic and diachronic way, comparing in the latter case maps referred to different time slices.

Two primary limitations have to be considered for the present study. First of all, Twitter is not the most used social media in Italy, taking into account only microblogs and social networking sites on which textual contents are mainly posted. The audience of Twitter was about 10 million users in 2019, ranking behind Facebook, which had an audience three times greater in the same period [63]. Moreover, according to some unofficial sources [64], the gender and age composition of the Twitter audience is unbalanced, with male users representing two-thirds of the total (as against an amount fewer than a half for the Italian population) and an average age of 32 years old (about 45 years old in Italy). Even if it is not easy to precisely profile Twitter users from a demographic viewpoint, because of the lack of reliable data, the hints suggest that the discourse generated on this social media can not be considered representative of the entire population's opinions. Nevertheless, in [7] some remarkable evidence from a comparative study concerning COVID-19 discourse on different social media in addition to Twitter has been reported, showing very similar topics and dynamics. Another limitation could be found, as mentioned above, in the use of a single set of keywords for all the analysed sub-periods, disregarding the different key events that occurred during the pandemic. A variation in the keywords used to retrieve the tweets posted in the different time slices and a narrower time window for each sub-period may improve the informative power of the thematic analysis. Despite the shortcomings inherent to the applied research design, the use of a strategy based on thematic analysis offers a new analytical tool for studying different social and cultural phenomena from the textual data conveyed in social media. Future developments of this research frontier will consider the automatic labelling of emerging topics and the use of covariates (e.g., the geographical localisation of users) to have interesting insights on specific sub-groups of users involved in the phenomena under investigation. Moreover, the extension of the proposed approach to deal with textual bodies written in different languages will be considered. The employment of a cross-language thematic analysis could enrich the analysis of social and cultural phenomena, offering the possibility of tracking the topics belonging to a given issue of interest in the light of the country-specific differences.

**Author Contributions:** Conceptualisation: M.A., C.C., M.M. and M.S.; methodology: M.A. and C.C.; formal analysis: L.D., M.M. and M.S.; software: M.A., L.D. and M.S.; writing—original draft preparation: M.M. and M.S.; writing—review and editing: M.A., C.C., L.D., M.M. and M.S.; supervision: C.C.; project administration: funding acquisition: M.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study is part of the research project *Topic extraction and modelling in large textual databases*, funded by the University of Calabria (Fondo Ateneo ex60%, 2021).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** All data sources have been cited in the manuscript. Interested readers can contact the authors of this paper for more details.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## Appendix A

**Table A1.** Topics appearing in the 1/2–9/3/2020 time slice.

| Quadrant | Labels | |
|---|---|---|
| | **Italian** | **English** |
| I (*hot topics*) | smart working<br>nuove misure<br>altre province | smart working<br>new measures<br>other provinces |
| II (*basic topics*) | porte chiuse<br>juve inter<br>senza pubblico | closed-doors<br>Juve Inter<br>without audience |
| | zona rossa<br>scuole chiuse<br>università chiuse | red zone<br>closed schools<br>closed universities |
| | corona virus<br>fake news<br>emergenza sanitaria | corona virus<br>fake news<br>health-care emergency |
| | terapia intensiva<br>nuovi casi<br>emilia romagna | critical care<br>new cases<br>Emilia Romagna |
| | zone rosse<br>cosa fare<br>nuovo decreto | red zones<br>what to do<br>new decree-law |
| | giuseppe conte<br>matteo salvini<br>emergenza epidemiologica | Giuseppe Conte<br>Matteo Salvini<br>epidemiological emergency |
| | primo morto<br>hong kong<br>prima vittima | first death<br>Hong Kong<br>first victim |
| | primo caso<br>pronto soccorso<br>caso sospetto | first case<br>emergency room<br>suspected case |
| | posti letto<br>sistema sanitario<br>operatori sanitari | beds<br>health-care system<br>health professionals |
| III (*peripheral topics*) | ospedale sacco<br>ceppo italiano<br>primo focolaio | Sacco Hospital<br>italian strain<br>first outbreak |
| | mobile world<br>world congress<br>stato annullato | Mobile World<br>world congress<br>been canceled |

**Table A1.** *Cont.*

| Quadrant | Labels | |
|---|---|---|
| | **Italian** | **English** |
| IV (*niche topics*) | ilari capua<br>meno letale<br>unica cura | Ilaria Capua<br>less-lethal<br>only cure |
| | topi vivi<br>luca zaia<br>cinesi mangiano | live mice<br>Luca Zaia<br>Chinese eat |
| | pazienti gravi<br>farmaco somministrato<br>anti artrite | seriously ill patients<br>administered drug<br>anti-arthritis |
| | australiano supera<br>vaccino australiano | Australian passed (the tests)<br>Australian vaccine |
| | familiari aspettano<br>napoli morta | relatives are waiting<br>Naples dead |
| | vasco rossi<br>tampone costa<br>costa dollari | Vasco Rossi<br>swab expensive<br>costs dollars |

**Table A2.** Topics appearing in the 10/3–4/5/2020 time slice.

| Quadrant | Labels | |
|---|---|---|
| | **Italian** | **English** |
| I (*hot topics*) | sanità lombarda<br>chiede aiuto<br>medici cubani | Lombard health-care<br>ask for help<br>Cuban doctors |
| | medici senza<br>senza frontiere<br>storia triste | doctors without<br>without borders<br>brief history |
| II (*basic topics*) | giuseppe conte<br>emergenza epidemiologica<br>decreto legge | Giuseppe Conte<br>epidemiological emergency<br>decree-law |
| | forza italia<br>parlamento europeo<br>italia viva | Forza Italia<br>European Parliament<br>Italia Viva |
| | tecnico scientifico<br>comitato tecnico | technical scientific<br>technical committee |
| | anti artrite<br>ospedali italiani<br>protocollo nazionale | anti-arthritis<br>Italian hospitals<br>national protocol |
| | contagi morti<br>mondo contagi<br>italia contagi | infection deaths<br>world infections<br>Italy infections |
| | nuovi casi<br>terapia intensiva<br>protezione civile | new cases<br>critical care<br>civil protection |
| | salvare vite<br>vite umane<br>insieme possiamo | save lives<br>human lives<br>together we can |
| | primo caso<br>nuovo caso<br>caso positivo | first case<br>new case<br>positive case |

**Table A2.** *Cont.*

| Quadrant | Labels | |
|---|---|---|
| | **Italian** | **English** |
| II (*basic topics*) | corona virus<br>ogni giorno<br>prima linea | corona virus<br>every day<br>frontline |
| | casi guariti<br>guariti deceduti<br>deceduti positivi | recovered cases<br>recovered deaths<br>positive deaths |
| III (*peripheral topics*) | segretario generale<br>fuoco globale<br>globale firma | Secretary General<br>global fire<br>global signature |
| | rifiuto tedesco<br>solidale gretto | german refusal<br>supportive petty |
| | organizzazione mondiale<br>diventando virale<br>appello pubblico | world organisation<br>becoming viral<br>public appeal |
| IV (*niche topics*) | repubblica ceca<br>mascherine inviate<br>sequestrato migliaia | Czech Republic<br>face mask shipped<br>seized thousands |
| | riepilogo casi<br>contagio informazioni<br>update riepilogo | case summary<br>infection information<br>update summary |
| | bela madunina<br>commuove milano<br>milano suonando | Bela Madunina<br>touched Milan<br>Milan performing |
| | salute pubblica<br>senso civico<br>assolutamente attenerci | public health<br>citizenship<br>adhere strictly |
| | nuove forme<br>grossa falla<br>studiate appositamente | new forms<br>big gap<br>specially designed |
| | fuori controllo<br>prigionieri politici<br>nasrin sotoudeh | out of control<br>political prisoner<br>Nasrin Sotoudeh |

**Table A3.** Topics appearing in the 5/5–31/8/2020 time slice.

| Quadrant | Labels | |
|---|---|---|
| | **Italian** | **English** |
| I (*hot topics*) | molto interessante<br>italia maggio<br>situazione italia | very interesting<br>Italy May<br>Italy situation |
| | numeri assoluti<br>bollettino quotidiano<br>fonte ministero | absolute numbers<br>daily bulletin<br>source ministry |
| | test sierologici<br>task force<br>personale sanitario | serology tests<br>task force<br>medical staff |

**Table A3.** *Cont.*

| Quadrant | Labels | |
|---|---|---|
| | **Italian** | **English** |
| II (*basic topics*) | fake news<br>super diffusori<br>italiani super | fake news<br>super spreader<br>Italians super |
| | contact tracing<br>fase analisi<br>tracing scenario | contact tracing<br>screening step<br>tracing scenario |
| | molti meno<br>meno morti | many less<br>fewer deaths |
| | nuovi casi<br>nuovi positivi<br>nuovi contagi | new cases<br>new infections<br>new positives |
| | seconda ondata<br>senza mascherina<br>linee guida | second wave<br>without mask<br>guidelines |
| | nuovi decessi<br>altri nuovi<br>gran bretagna | new deaths<br>other new<br>Great Britain |
| | protezione civile<br>raccolta fondi<br>conto corrente | civil protection<br>fundraising<br>bank account |
| III (*peripheral topics*) | milano folla<br>poche mascherina | Milan crowd<br>few masks |
| | mappa interattiva<br>situazione italiana | interactive map<br>Italy situation |
| | basta armi<br>verde inquinamento<br>investiamo lavoro | stop guns<br>green pollution<br>investing employment |
| | mutamento apparente<br>regole eurozona<br>troppo poco | apparent change<br>Eurozone rules<br>too little |
| IV (*niche topics*) | aziende inquintanti<br>vogliono usare<br>continuare inquinare | polluting industries<br>plan to use<br>continue polluting |
| | appena firmato<br>educazione opportunità<br>offrire educazione | just signed<br>education opportunity<br>offer education |
| | truffa biologica<br>farcele domande | biological fraud<br>asking us questions |
| | buone condizioni<br>francia italia<br>ripresa basato | good condition<br>France Italy<br>recovery based |
| | caso lombardia<br>differenze regionali<br>mortalità ufficiale | Lombardy case<br>regional differences<br>official mortality |
| | app dreamlab<br>potremo combattere<br>cancro insieme | dreamlab app<br>we could fight<br>cancer together |

**Table A4.** Topics appearing in the 1/9–31/12/2020 time slice.

| Quadrant | Labels | |
|---|---|---|
| | **Italian** | **English** |
| I (*hot topics*) | grandi farmaceutiche<br>limitano accesso<br>sospendere brevetti | big pharmaceutical<br>limit access<br>suspend patents |
| | milano linate<br>milano malpensa<br>area test | Milan Linate<br>Milan Malpensa<br>area test |
| | jerry scotti<br>carlo conti<br>controllo medico | Jerry Scotti<br>Carlo Conti<br>medical check |
| | infermiere stroncato<br>luciano quaglieri<br>filippo neri | nurse struck<br>Luciano Quagliari<br>Filippo Neri |
| | numeri assoluti<br>bollettino quotidiano<br>fonte ministero | absolute numbers<br>daily bulletin<br>source ministry |
| II (*basic topics*) | marina barlusconi<br>berlusconi positiva<br>agenzia ansa | Marina Berlusconi<br>Berlusconi positive<br>ANSA agency |
| | seconda ondata<br>zona rossa<br>nuova variante | second wave<br>red zone<br>new variant |
| | nuovi casi<br>nuovi positivi<br>nuovi contagi | new cases<br>new positives<br>new infections |
| III (*peripheral topics*) | protezione civile<br>raccolta fondi<br>conto corrente | civil protection<br>fundraising<br>bank account |
| | de laurentis positivo<br>del pino positivo<br>lega serie | De Laurentis positive<br>Del Pino positive<br>League serie |
| | visioni positivi<br>diversi paesi<br>salute pubblica | positive minks<br>several countries<br>public health |
| | basta armi<br>verde inquinamento<br>investiamo lavoro | stop guns<br>green pollution<br>investing employment |
| IV (*niche topics*) | padre livio<br>progetto criminale<br>élites mondiali | Father Livio<br>crime design<br>world elite |
| | essere crudeli<br>allevamenti visioni<br>chiusi definitivamente | be cruel<br>mink farms<br>permanently closed |
| | scopi politici<br>impostori usano<br>tamponi eseguiti | political ends<br>impostors use<br>performed swabs |
| | mutamento apparente<br>regole eurozona<br>troppo poco | apparent change<br>Eurozone rules<br>too little |

**Table A4.** *Cont.*

| Quadrant | Labels | |
|---|---|---|
| | **Italian** | **English** |
| IV (*niche topics*) | beni pubblici<br>sostenibilità<br>capitalismo dopo | public goods<br>sustainability<br>capitalism after |
| | the lancet<br>impatto psicologico<br>come ridurlo | The Lancet<br>psychological impact<br>how to reduce it |
| | bene comune<br>più poveri<br>garantire accesso | common good<br>more poor<br>ensure access |
| | app dreamlab<br>potremo combattere<br>cancro insieme | dreamlab app<br>we could fight<br>cancer together |

## References


1. World Health Organization. WHO Director-General's Opening Remarks at the Media Briefing on COVID-19. Available online: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (accessed on 15 January 2022).
2. Platto, S.; Wang, Y.; Zhou, J.; Carafoli, E. History of the COVID-19 pandemic: Origin, explosion, worldwide spreading. *Biochem. Biophys. Res. Commun.* **2021**, *538*, 14–23. [CrossRef] [PubMed]
3. Taylor, L.H.; Latham, S.M.; Woolhouse, M.E. Risk factors for human disease emergence. *Philos. Trans. R. Soc. Lond. B: Biol. Sci* **2001**, *356*, 983–989. [CrossRef] [PubMed]
4. Rasmussen, A.L. On the origins of SARS-CoV-2. *Nat. Med.* **2021**, *27*, 9. [CrossRef] [PubMed]
5. Westerman, D.; Spence, P.R.; Van Der Heide, B. Social Media as information source: Recency of updates and credibility of information. *J. Comput.-Mediat. Commun.* **2014**, *19*, 171–183. [CrossRef]
6. Pulido, C.M.; Ruiz-Eugenio, L.; Redondo-Sama, G.; Villarejo-Carballido, B. A New Application of Social Impact in Social Media for Overcoming Fake News in Health. *Int. J. Environ. Res. Public Health* **2020**, *17*, 2430. [CrossRef] [PubMed]
7. Cinelli, M.; Quattrociocchi, W.; Galeazzi, A.; Valensise, C.M.; Brugnoli, E.; Schmidt, A.L.; Zola, P.; Zollo, F.; Scala, A. The COVID-19 social media infodemic. *Sci. Rep.* **2020**, *10*, 16598. [CrossRef] [PubMed]
8. Liu, H.; Liu, W.; Yoganathan, V.; Osburg, V.-S. COVID-19 information overload and generation Z's social media discontinuance intention during the pandemic lockdown. *Technol. Forecast. Soc. Chang.* **2021**, *166*, 120600. [CrossRef] [PubMed]
9. Aiden, E.; Michel, J.-B. *Uncharted: Big Data as a Lens on Human Culture*; Riverhead Books: New York, NY, USA, 2013.
10. Michel, J.-B.; Shen, Y.K.; Aiden, A.P.; Veres, A.; Gray, M.K.; Pickett, J.P.; Hoiberg, D.; Clancy, D.; Norvig, P.; Orwant, J.; et al. Quantitative analysis of culture using millions of digitized books. *Science* **2011**, *331*, 176–182. [CrossRef] [PubMed]
11. Boyd, D.; Crawford, K. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Inf. Commun. Soc.* **2012**, *15*, 662–679. [CrossRef]
12. Ibrahim, R.; Elbagoury, A.; Kamel, M.S.; Karray, F. Tools and approaches for topic detection from Twitter streams: Survey. *Knowl. Inf. Syst.* **2018**, *54*, 511–539. [CrossRef]
13. Misuraca, M.; Spano, M. Unsupervised analytic strategies to explore large document collections. In *Text Analytics. Advances and Challenges*; Iezzi, D.F., Mayaffre, D., Misuraca, M., Eds.; Springer: Heidelberg, Germany, 2020; pp. 17–28.
14. Sayyadi, H.; Raschid, L. A graph analytical approach for topic detection. *ACM Trans. Internet Technol.* **2013**, *13*, 1–23. [CrossRef]
15. Cobo, M.J.; López-Herrera, A.G.; Herrera-Viedma, E.; Herrera, F. An approach for detecting, quantifying, and visualising the evolution of a research field: A practical application to the fuzzy sets theory field. *J. Infometr.* **2011**, *5*, 146–166. [CrossRef]
16. Loh, S.; Palazzo M. de Oliveira, J.; Leite Gastal, F. Knowledge discovery in textual documentation: Qualitative and quantitative analyses. *J. Doc.* **2001**, *57*, 577–590. [CrossRef]
17. Salton, G.; Wong, A.; Yang, C.S. A vector space model for automatic indexing. *Commun. ACM* **1975**, *18*, 613–620. [CrossRef]
18. Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523. [CrossRef]
19. Asher, N. Discourse topic. *Theor. Ling.* **2004**, *30*, 163–201. [CrossRef]
20. Wartena, C.; Brussee, R. Topic Detection by Clustering Keywords. In Proceedings of the 19th International Workshop on Database and Expert Systems Applications, Turin, Italy, 1–5 September 2008; pp. 54–58.
21. Balbi, S.; Misuraca, M.; Spano, M. A Two-Step Strategy for Improving Categorisation of Short Texts. In Proceedings of the 14th International Conference on Statistical Analysis of Textual Data, Rome, Italy, 12–15 June 2018; pp. 60–67.
22. Benzécri, J. *Histoire et Préhistoire de L'analyse des Données*; Dunod: Paris, France, 1982.
23. Lebart, L.; Salem, A.; Berry, L. *Exploring Textual Data*; Kluwer: Dordrecht, The Netherlands, 1988.

24. Deerwester, S.; Dumais, S.; Furnas, G.; Landauer, T.; Harshman, R. Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407. [CrossRef]
25. Arabie, P.; Hubert, L. Advances in Cluster Analysis Relevant to Marketing Research. In *From Data to Knowledge. Studies in Classification, Data Analysis, and Knowledge Organization*; Gaul, W., Pfeifer, D., Eds.; Springer: Heidelberg, Germany, 1996; pp. 3–19.
26. Blei, D.; Ng, A.; Jordan, M. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
27. Yan, X.; Guo, J.; Lan, Y.; Cheng, X. A Biterm Topic Model for Short Texts. In Proceedings of the 22nd International Conference on World Wide Web, Rio De Janeiro, Brazil, 13–17 May 2013; pp. 1445–1456.
28. Griffiths, T.; Steyvers, M. Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5228–5235. [CrossRef]
29. Carley, K. Network Text Analysis: The network position of concepts. In *Text Analysis For The Social Sciences*; Roberts, C.W., Ed.; Routledge: New York, NY, USA, 1997; pp. 79–102.
30. Popping, R. *Computer-Assisted Text Analysis*; Sage: London, UK, 2000.
31. Lim, K.; Karunasekera, S.; Harwood, A. ClusTop: A clustering-based topic modelling algorithm for twitter using word networks. In Proceedings of the 2017 IEEE International Conference on Big Data, Boston, MA, USA, 11–14 December 2017; pp. 2009–2018.
32. Misuraca, M.; Scepi, G.; Spano, M. A network-based concept extraction for managing customer requests in a social media care context. *Int. J. Inf. Manag.* **2020**, *51*, 101956. [CrossRef]
33. Agüero-Torales, M.M.; Vilares, D.; López-Herrera, A.G. Discovering topics in Twitter about the COVID-19 outbreak in Spain. *Proces. Leng. Nat.* **2021**, *66*, 177–190.
34. Comito, C. How COVID-19 information spread in US The Role of Twitter as Early Indicator of Epidemics. *IEEE Trans. Serv. Comput.* **2021**, 1–12. [CrossRef]
35. Jang, H.; Rempel, E.; Roth, D.; Carenini, G.; Janjua, N.Z. Tracking COVID-19 Discourse on Twitter in North America: Infodemiology Study Using Topic Modeling and Aspect-Based Sentiment Analysis. *J. Med. Internet Res.* **2021**, *23*, e25431. [CrossRef] [PubMed]
36. Gutiérrez, I.; Guevara, J.A.; Gómez, D.; Castro, J.; Espínola, R. Community Detection Problem Based on Polarization Measures: An Application to Twitter: The COVID-19 Case in Spain. *Mathematics* **2021**, *9*, 443. [CrossRef]
37. Lamsal, R. Design and analysis of a large-scale COVID-19 tweets dataset. *Appl. Intell.* **2021**, *51*, 2790–2804. [CrossRef] [PubMed]
38. van Eck, N.; Waltman, L. How to normalise co-occurrence data? An analysis of some well-known similarity measures. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 1635–1651.
39. Egghe, L. On the relation between the association strength and other similarity measures. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 1502–1504. [CrossRef]
40. Fortunato, S. Community detection in graphs. *Phys. Rep.* **2010**, *486*, 75–174. [CrossRef]
41. Blondel, V.D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech Theory Exp.* **2008**, *2008*, P10008. [CrossRef]
42. Yang, Z.; Algesheimer, R.; Tessone, C.J. A comparative analysis of community detection algorithms on artificial networks. *Sci. Rep.* **2016**, *6*, 30750. [CrossRef] [PubMed]
43. Callon, M.; Courtial, J.P.; Laville, F. Co-word analysis as a tool for describing the network of interactions between basic and technological research—The case of polymer chemistry. *Scientometrics* **1991**, *22*, 155–205. [CrossRef]
44. Basile, V.; Caselli, T. 40twita 1.0: A Collection of Italian Tweets during the COVID-19 Pandemic. Available online: http://twita.di.unito.it/dataset/40wita (accessed on 10 December 2021).
45. Basile, V.; Lai, M.; Sanguinetti, M. Long-term Social Media Data Collection at the University of Turin. In Proceedings of the Fifth Italian Conference on Computational Linguistics, Turin, Italy, 10–12 December 2018. Available online: http://ceur-ws.org/Vol-2253/paper48.pdf (accessed on 10 December 2021).
46. Pelagatti, M.; Maranzano, P. Assessing the effectiveness of the Italian risk-zones policy during the second wave of COVID-19. *Health Policy* **2021**, *125*, 1188–1199. [CrossRef] [PubMed]
47. Aria, M.; Cuccurullo, C. Bibliometrix: An R-tool for comprehensive science mapping analysis. *J. Informetr.* **2017**, *11*, 959–975. [CrossRef]
48. Aria, M.; Misuraca, M.; Spano, M. Mapping the evolution of social research and data science on 30 years of Social Indicators Research. *Soc. Indic. Res.* **2020**, *149*, 803–831. [CrossRef]
49. Bull, M. The Italian government response to Covid-19 and the making of a prime minister. *Contemp. Ital. Politics* **2021**, *13*, 149–165. [CrossRef]
50. Zammitti, A.; Imbrogliera, C.; Russo, A.; Zarbo, R.; Magnano, P. The Psychological Impact of Coronavirus Pandemic Restrictions in Italy. The Mediating Role of the Fear of COVID-19 in the Relationship between Positive and Negative Affect with Positive and Negative Outcomes. *Eur. J. Investig. Health Psychol. Educ.* **2021**, 11, 697–710. [CrossRef]
51. Boccia Artieri, G.; Greco, F.; La Rocca, G. The construction of the meanings of #coronavirus on Twitter: An analysis of the initial reactions of the Italian people. *Int. Rev. Sociol.* **2021**, *31*, 287–309.
52. De Santis, E.; Martino, A.; Rizzi, A. An Infoveillance System for Detecting and Tracking Relevant Topics From Italian Tweets During the COVID-19 Event. *IEEE Access* **2020**, *8*, 132527–132538. [CrossRef]
53. Entman, R.M. Framing: Towards clarification of a fractured paradigm. *J. Commun.* **1993**, *43*, 51–58. [CrossRef]
54. López-Rabadán, P. Framing Studies Evolution in the Social Media Era. Digital Advancement and Reorientation of the Research Agenda. *Soc. Sci.* **2022**, *11*, 9. [CrossRef]

55. Valenzuela, S.; Piña, M.; Ramírez, J. Behavioral Effects of Framing on Social Media Users: How Conflict, Economic, Human Interest, and Morality Frames Drive News Sharing. *J. Commun.* **2017**, *67*, 803–826. [CrossRef]
56. Tahamtan, I.; Potnis, D.; Mohammadi, E.; Miller, L.E.; Singh, V. Framing of and Attention to COVID-19 on Twitter: Thematic Analysis of Hashtags. *J. Med. Internet Res.* **2021**, *23*, e30800. [CrossRef] [PubMed]
57. Wicke, P.; Bolognesi, M.M. Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter. *PLoS ONE* **2020**, *15*, e0240010. [CrossRef]
58. Ophir, Y.; Walter, D.; Arnon, D.; Lokmanoglu, A.; Tizzoni, M.; Carota, J.; D'Antiga, L.; Nicastro, E. The Framing of COVID-19 in Italian Media and Its Relationship with Community Mobility: A Mixed-Method Approach. *J. Health Commun.* **2021**, *26*, 161–173. [CrossRef]
59. Wang, D.; Mao, Z. From risks to catastrophes: How Chinese Newspapers framed the Coronavirus Disease 2019 (COVID-19) in its early stage. *Health Risk Soc.* **2021**, *23*, 93–110. [CrossRef]
60. Caldarelli, G.; De Nicola, R.; Petrocchi, M.; Pratelli, M.; Saracco, F. Flow of online misinformation during the peak of the COVID-19 pandemic in Italy. *EPJ Data Sci.* **2021**, *10*, 34. [CrossRef] [PubMed]
61. Guarino, S.; Pierri, F.; Di Giovanni, M.; Celestini, A. Information disorders during the COVID-19 infodemic: The case of Italian Facebook. *Online Soc. Netw. Media* **2021**, *22*, 100124. [CrossRef] [PubMed]
62. Posetti, J.; Bontcheva, K. Disinfodemic. Deciphering COVID-19 Disinformation. Policy Brief 1, UNESCO. Available online: https://en.unesco.org/sites/default/files/disinfodemic_deciphering_covid19_disinformation.pdf (accessed on 24 January 2022).
63. Autorità per le Garanzie nelle Comunicazioni. Osservatorio Sulle Comunicazioni n. 1/2021. Available online: https://www.agcom.it/documents/10179/22666659/Documento+generico+22-04-2021/30bb16e2-adb6-4de0-b1f5-4a4df2d8ec24?08February2022version=1.1 (accessed on 24 January 2022).
64. Datareportal. Digital 2020: Italy. Available online: https://datareportal.com/reports/digital-2020-italy (accessed on 24 January 2022).

**ATTRIBUZIONE DELLE PARTI**
**DICHIARAZIONE SOSTITUTIVA DI CERTIFICAZIONE**
(Art 46 D.P.R. 28.12.2000 n. 445 recante il "T.U. delle disposizioni legislative e regolamentari in materia di documentazione amministrativa")

Il sottoscritto, **Luca D'Aniello**, nato a Napoli il 16/06/1994, residente a Napoli (NA), in Via G. A. Campano 164;

consapevole della responsabilità cui possono andare incontro in caso di dichiarazione mendace o di esibizione di atto falso o contenente dati non più rispondenti a verità, nonché delle sanzioni penali richiamate ai sensi del comma 1. dell'art.45 del D.P.R. 445 del 28.12.2000 e per le ipotesi di falsità in atti e dichiarazioni mendaci, ai sensi degli artt. 46 e 47 del D.P.R. n. 445 del 28.12.2000

**D I C H I A RA**

che in relazione all'articolo "*Thematic Analysis as a New Culturomic Tool: The Social Media Coverage on COVID-19 Pandemic in Italy*", pubblicato su *Sustainability,* DOI: *https://doi.org/10.3390/su14063643*,
pur essendo frutto di una stretta e paritetica collaborazione fra gli autori che ne condividono l'impostazione, le ipotesi di ricerca ed i risultati conseguiti, ai fini dell'individuazione dell'apporto individuale nei lavori in collaborazione (art. 4 c. 2 e art. 5 c. 2 DM 76 del 7 giugno 2012), l'apporto individuale di **Luca D'ANIELLO** è indicato nella sezione **Author Contributions** dell'articolo, in coerenza con le ricerche e le tematiche affrontate e desumibili dal curriculum e dal percorso scientifico del candidato.

Data 07/11/2025

Luca D'Aniello