Multilingual topic modelling for tracking COVID-19 trends based on Facebook data analysis

Amina AMARA
Universtity of Sfax
Mohamed Ali HADJ TAIEB (mohamedali.hadjtaieb@gmail.com)
Faculty of Sciences, University of Sfax https://orcid.org/0000-0002-2786-8913
Mohamed BEN AOUICHA
Faculty of Sciences, University of Sfax https://orcid.org/0000-0002-2277-5814

Research Article

Keywords: Social media analysis, Covid-19, Topic modelling, Facebook, Data visualization, Multilingual

Posted Date: July 23rd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-45177/v1

Multilingual topic modelling for tracking COVID-19 trends based on Facebook data analysis

Amina AMARA[a,1], Mohamed Ali HADJ TAIEB[*,a,2], Mohamed BEN AOUICHA[a,3]

[a] *University of Sfax, Tunisia*

Abstract

Social data has shown important role in tracking, monitoring and risk management of disasters. Indeed, several works focused on the benefits of social data analysis to the healthcare practices and curing. Similarly, these data are exploited now for tracking the COVID-19 pandemic but the majority of works exploited twitter as source. In this paper, we choose to exploit Facebook, rarely used, for tracking the evolution of COVID-19 related trends. In fact, a multilingual dataset covering 7 languages (English (EN), Arabic (AR), Spanish (ES), Italian (IT), German (DE), French (FR) and Japanese (JP)) is extracted from Facebook public posts. The proposal is an analytics process including a data gathering step, pre-processing, LDA-based topic modelling and presentation module using graph structure. Data analysing covers the duration spanned from January 1st, 2020 to May 15, 2020 divided on three periods in cumulative way: first period January-February, second period March-April and the last one to 15 May. The results showed that the extracted topics correspond to the chronological development of what has been circulated around the pandemic and the measures that have been taken in the various languages under discussion.

Keywords: Social media analysis, Covid-19, Topic modelling, Facebook, Data

*Corresponding author
[1] amaraamina2290@gmail.com
[2] mohamedali.hadjtaieb@gmail.com
[3] mohamed.benaouicha@fss.usf.tn

visualization, Multilingual

1. Introduction

People around the world need continuously public safety and emergency management services. These services require tools that detect quickly the occurrence of emergencies and create a correct and detailed idea about the current situation. Such tools may help alleviating desolation under harsh conditions related to natural or human-made disasters by fast and semi-automatic identification of the type, extent, place, intensity, and implications of the disaster and the knowledge transfer about the disaster issue. Content provided by the user is growing during disasters in comparison with the normal situation. Therefore, the use of analytics is necessary to identify emergencies and recent disasters, based on social networks and media search, and direct relief proportional to the needs.

Disasters are ranging from earthquakes, floods, hurricanes, droughts, tsunamis, landslides, terrorism and wars to solar flares, cosmic explosions and meteorites. These risks imperil people, populations, civilization, and humankind. Defending against these threats requires various kinds of endeavors supported by varied tools and a great range of technical and human capabilities. Advanced knowledge on the nature of emergencies and effective awareness may help reducing the costs of the disasters. Information and communication tools are vital in modeling emergencies and population response, and in the accurate monitoring of disasters. Advanced technologies can benefit from the development of the new means and methods of data and information transmission, including the Internet, the social networks and social media Sebei et al. (2018). Several researches analyse the benefit of social media in monitoring the defending reaction against the disasters and for making accurate strategies in real time by responding to the urgent population needs Teodorescu (2015); k joseph et al. (2018); Landwehr & Carley (2014). They focused on the use of social media in relation with different disasters kinds such as the earthquakeDoan et al. (2012); Miyabe

et al. (2012); Sakaki et al. (2010a), tsunami PEARY et al. (2012), typhoons Daga (2017), storms Ulvi et al. (2019), flood Kankanamge et al. (2020) and the health risks such as epidemic virus spreading Ahmed (2018), prediction of the contagious population behaviour and accurate detection and identification of professionally unreported drug side effects using widely available public data (open data) Fan et al. (2020); Pizzuti et al. (2020).

The majority of works exploit Twitter data due to its open policies in front of the information extraction despite that the Facebook is ranked as the first social network having active users. The Digital 2019 reports[4] include extensive insights into people's use of the world's top social platforms in more than 230 countries and territories around the world. Worldwide social media users' numbers have grown to almost 3.5 billion at the start of 2019, with 288 million new users in the past 12 months pushing the global penetration figure to 45 percent. Facebook maintains its top platform ranking in early 2019. Facebook's monthly active users numbers grew steadily across the past 12 months, and the platform's latest earnings announcement reports year-on-year user growth of almost 10 percent.

To our knowledge, the present work is the first that extracts insights from Facebook provided centric COVID-19 data. In fact, the majority of works on the pandemic COVID-19 are based on Twitter data. This paper describes the process of building the provided crowdsourcing trends and its evolution within the time since the first January 2020. This study exploits the LDA-based topic modelling method in a multilingual framework and provides a novel representation method based on graph structure and handling it with graph visualization software.

The rest of the paper is organized as follows. Section 2 provides an overview of the research works focusing on studying the utility of social data in relation with different disasters kinds through highlighting those on health domain. Section 3 depicts some research works focusing on COVID-19 related social data

[4] https://wearesocial.com/global-digital-report-2019

analysis. The topic modelling method latent dirichlet allocation is detailed in section 4. Then, the proposed Facebook data-based tracking system of the trends through the time is depicted in Section 5 with its components. The multilingual gathered COVID-19 centric dataset and its statistics and characteristics are presented in Section 6. Section 7 reports on the interpretation of the results (COVID-19 trends) in a multilingual framework. The final section is devoted to presenting our conclusions and future research.

2. Related Work

This section analyzes research works that have studied the use of social data on monitoring and tracking disasters in two dimensions, natural and health disasters.

2.1. Natural disasters

Social media such as Facebook and Twitter have proven to be a useful resource to understand public opinion towards real world events. After the great east Japan earthquake in 2011, numerous tweets were exchanged on Twitter. Several studies have already pointed out that micro-blogging systems have shown potential advantages in emergency situations, but it remains unclear how people use them. Doan et al. (2012) investigated over 1.5 million Twitter messages(tweets) for the period ranging from 9 March 2011 to 3 May 2011 in order to track awareness and anxiety levels in the Tokyo metropolitan district to the 2011 Tohoku Earthquake and subsequent tsunami and nuclear emergencies. Miyabe et al. (2012) gathered tweets immediately after the earthquake and analyzed various factors, including locations. The results showed that the people in the disaster area tend to directly communicate with each other (reply-based tweet). On the other hand, people in the other area prefer spread the information from the disaster area by using retweet.

An important characteristic of Twitter is its real-time nature. For example, when an earthquake occurs, people make many Twitter posts (tweets) related

to the earthquake, which enables detection of earthquake occurrence promptly, simply by observing the tweets. Sakaki et al. (2010b) investigated the real-time interaction of events such as earthquakes, in Twitter, and proposed an algorithm to monitor tweets and to detect a target event. To detect such event, they proposed a classifier of tweets based on features such as the keywords in a tweet, the number of words, and their context. They consider each Twitter user as a sensor. As an application, they construct an earthquake reporting system in Japan.

PEARY et al. (2012) show that during the 2011 east Japan earthquake and tsunami, social media such as Twitter and Facebook served as a lifeline for directly affected individuals and a means of information sharing. Social media was used to perform vital relief functions such as safety identification, displaced-persons locating, damage information provision, support for disabled individuals, volunteer organization, fund-raising, and moral support systems. Their study discusses the potential for public, civil society, and government organizations to utilize social media in disaster preparedness and response.

Daga (2017) implemented content analysis and Social Network Analysis (SNA) on the tweets regarding different situations to be able to understand and depict a visual representation of interaction between users. They studied the user interaction of the Filipino community between two major typhoons that hit the Philippines. Results revealed that users tend to seek and share information from reliable sources such as news websites and verified Twitter users. Determining the interaction of Twitter users in an online community plays a vital role in information dissemination and allows appropriate response during disaster and emergency situations. In this study, SNA was used to help better understand and reveal the social interaction related to typhoon Haiyan[5] and Hagupit[6]. Results of this study have shown that user interaction among Filipino online community both Haiyan and Hagupit was influenced by Twitter-verified

[5] https://fr.wikipedia.org./wiki/Typhon_Haiyan
[6] https://en.wikipedia.org/wiki/Typhoon_Hagupit_(2014)

users having the highest value betweeness centrality score.

Moreover, social media use increases for the natural disasters such as cyclones, hurricanes, or typhoons occur. Ulvi et al. (2019) investigated the roles of social media and mainstream media on hurricanes and how they may potentially have a bigger impact. They studied influences and risk factors of media and their role on the distribution of information were observed. They concluded that social media platforms helped spread awareness, support, and warnings. Social media has shown to have impactful effects during tropical storms around the world. Public health professionals and emergency response team should utilize social media in relief for victims.

Harnessing the crowdsourced information, under social media platforms, has become an opportunity for authorities to obtain enhanced situation awareness data for efficient disaster management practices. Nonetheless, the current disaster-related Twitter analytics methods are not versatile enough to define disaster impacts levels as interpreted by the local communities. Kankanamge et al. (2020) prepared a data analysis framework, and identifying highly impacted disaster areas as perceived by the local communities. For this, the study used real-time Twitter data posted during the 2010–2011 South East Queensland Floods. The findings reveal that utilising Twitter is a promising approach to reflect citizen interests.

2.2. Health domain

Social media's showed a great capacity to allow public participation in content creation and circulation.

Ding & Zhang (2010) presented their study focusing on the use of social media during the H1N1 flu epidemic in the U.S. and China. This study demonstrates that governmental structures may use social media tools for one-way dissemination of risk decisions and policies. In contrast, the general public may get around the institutional control of risk information through extra-institutional collaborative risk communication to extract truths about the emerging risks.

Achrekar et al. (2011) were interested within reducing the impact of seasonal influenza epidemics and other pandemics such as the H1N1 due to its paramount importance for public health authorities. Studies have shown that effective interventions can be taken to contain the epidemics if early detection can be made. They presented the social network enabled flu trends framework, which monitors messages posted on Twitter with a mention of flu indicators to track and predict the emergence and spread of an influenza epidemic in a population. Based on the data collected during 2009 and 2010, they find that the volume of flu related tweets is highly correlated with the number of influenza-like illness cases reported by the centers for disease control and prevention.

In the context of the seasonal influenza, it can cause various complications, worsen chronic illnesses, and sometimes lead to deaths. In fact, during 2009 H1N1[7] flu pandemic, up to 203,000 deaths occurred worldwide. Early detection and prediction of disease outbreak is critical because it can provide more time to prepare a response and significantly reduce the impact caused by a pandemic. Lee et al. (2017) presented a system that predicts future influenza activities, provides more accurate real-time assessment than before, and combines real-time big social media data streams for predictive models to generate accurate predictions. Prediction of further flu levels can represent a big leap because such predictions provide insights for public health that can be benefit for planning, resource allocation, treatments and prevention.

Sharma et al. (2017) studied the pandemic of Zika virus infection. More publicized and of greater concern is the epidemic of microcephaly in Brazil, manifested by an apparent 20-fold increase in incidence from 2014-2015, believed to be caused by Zika virus infection in pregnant women. The increase in the spread of the disease has caused rapid activity surrounding it in social media. They used Facebook for dissemination of public health information via social media. For them, it is important to spread right information that helps public to preventative guidelines. The use of Facebook is argued to the fact that its

[7] https://www.cdc.gov/flu/pandemic-resources/2009-h1n1-pandemic.html

universal availability, outreach, and substantial influence on the information available to the public.

Pruss et al. (2019) examined Twitter discussion surrounding the 2015 outbreak of Zika. Their study is based on gathered data from Twitter mentioning Zika geolocated to North and South America. Using a multilingual topic model, they automatically identified and extracted the key topics of discussion across the dataset in English, Spanish, and Portuguese. They examined the variation in Twitter activity across time and location, finding that rises in activity tend to follow to major events.

Zarrad et al. (2014) exploited social media such as social networks, blogs and forums to analyze users' opinions, attitudes, and emotions about news or social events. In fact, they presented a case study about MERS[8] virus in Kingdom of Saudi Arabia to evaluate their approach.

Tran & Lee (2016) conducted a study of understanding and mining the spread of Ebola-related information on social media. In particular, they are based on large-scale data-driven analysis of geo-tagged social media messages to understand citizen reactions regarding Ebola, build information propagation models, and analyze spatial, temporal and social properties of Ebola-related information. Their work provides new insights into Ebola outbreak by understanding citizen reactions and topic-based information propagation, as well as providing a support for future public health crises. Missier et al. (2016), also, proposed a system for tracking Dengue epidemics using Twitter content classification and topic modelling.

During the epidemics or pandemics, the potential threat to the society is the propagation of rumors through the social media. Sicilia et al. (2018) proposed a rumour detection system that leverages on newly designed features, including influence potential and network characteristics measures. They tested their approach on a real dataset composed of health-related posts collected from Twitter microblog.

[8] http://www.emro.who.int/fr/health-topics/mers-cov/mers-cov.html

In next section, we revise the research works focusing on the use of social data analysis in relation with COVID-19 pandemic.

3. COVID-19 related social data analysis research works

The novel coronavirus disease, named COVID-19, was identified at the first time in Wuhan, Hubei Province, China by the end of December 2019. It has rapidly outbroken worldwide leading to a global health emergency on 30 January 2020 and it has been announced as a pandemic by the World Health Organization (WHO) on 11 March 2020[9]. This ongoing pandemic puts all societal levels in unprecedented situation and pushed many governments around the world to enforce different measures to contain the spread of this ongoing coronavirus. Distance learning, self-quarantines and social distancing are among the maintained measures.

These enforced unprecedented measures, especially "social distancing" the most widely used of such measures, have impacted the lifestyle of people around the globe, and bring them to the frontline social media platforms for both chatting and news. Social media websites like Facebook and Twitter are playing a central and significant role, more than ever, as adequate tools that allow people to stay connected during crises for global social discussions. As more and more social interactions turn online, an important amount of conversations around this ongoing coronavirus are continuously expanding. Researchers are mainly using these online conversations to understand the spread of this novel coronavirus, explore its aspects as well as monitoring people's reactions regarding to the global health emergency and so forth. Although the COVID-19 epidemic's appearance is relatively new, there is a rapid move in the research landscape since more than 24,000 research papers[10] are published online. These researches are distributed over several disciplines such as social science, medicine, public

[9] https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020

[10] https://www.technologyreview.com/2020/03/16/905290/coronavirus-24000-research-papers-available-open-data/Echobox=1585947735

health, and so on. Some of these research papers are preprint and have not yet been peer-reviewed. Table 1 presents some works taking profit from the generated social data to perform analysis related to the pandemic COVID-19.

In the computer science discipline, at the time of our writing (in mid-May 2020) and in the best of our knowledge, main publishers like Springer, IEEE Explorer, Science Direct and ACM published a few preliminary research works Alshaabi et al. (2020); Barkur et al. (2020); Chen et al. (2020b); Li et al. (2020); Limaye et al. (2020); Zhou et al. (2020) about social media's dynamics around the context of COVID-19 as it is illustrated in Table 1. However, there is prosperity of studies Chen et al. (2020a); Alqurashi et al. (2020); Banda et al. (2020); Boberg et al. (2020); Cinelli et al. (2020); Ferrara (2020); Gao et al. (2020); Haouari et al. (2020); Kleinberg et al. (2020); Kuchler et al. (2020); Lopez et al. (2020); Singh et al. (2020); Zarei et al. (2020) that are pre-print papers investigating the evolutionary aspects of the unfolding coronavirus disease. Collected data are used to analyze the behavioral change, track COVID-19 related misinformation and rumors' spreading Chen et al. (2020a), measure the emotional responses and worries[11] about the pandemic Banda et al. (2020); Kleinberg et al. (2020), and so forth. Banda et al.[12] have released a dataset of about 4.4 million of daily tweets related to coronavirus context collected through the Twitter stream API using keywords as 'Covid-19' and 'Coronavirus'. Chen et al. (2020a) proposed a multilingual Covid-19 Twitter dataset, which is made available to the research community[13] , collected since January 2020 using Twitter stream API and Tweepy to follow specific keywords and trending accounts simultaneously. This dataset helps tracking coronavirus unverified rumors, enable understanding of users' sentiment towards this global crisis and more. There are other works focusing on collecting data in specific languages and having similar goals to the latter ones Alqurashi et al. (2020); Haouari et al. (2020). Haouari

[11] https://github.com/ben-aaron188/covid19worry
[12] https://zenodo.org/record/3738018#.XrALQ1UzZ0w
[13] https://github.com/echen102/COVID-19-TweetIDs

et al. (2020)[14] as well as Alqurashi et al. (2020) proposed an Arabic COVID-19 Twitter datasets that present preliminarily analysis on COVID-19 tweets to reveal arabic users' behavior and sentiment during this novel coronavirus. Moreover, Singh et al. (2020) present a work that looks at the conversations surrounding the vast-moving of COVID-19 on the Twitter social network and tackled the problem of misinformation sharing on this social network. Other works Cinelli et al. (2020); Gao et al. (2020) are focusing on data shared on multiple social media platforms related to the COVID-19 pandemic. Indeed, Gao et al. (2020) have released a multilingual dataset of social media posts retrieved from the two micro-blogging websites Twitter and Weibo in 3 different languages: English, Japanese and Chinese language. Last but not least, to the best of our knowledge and at the time of this writing,there are currently very few works focusing on Covid-19 Facebook data analysis Boberg et al. (2020); Kuchler et al. (2020); Perrotta et al. (2020). Kuchler et al. (2020) examine the relationship between the geographic spread of COVID-19 and the geo-location information from users of social networks such as Facebook, in the United States and Italy. Perrotta et al. (2020) proposed a rapid response monitoring system via a continuously run survey across eight countries. They collected key information on people's health status, attitudes and close social contacts by recruiting participants through targeted Facebook advertisement campaigns. Moreover, a Facebook data for Good Mobility dashboard [15] is conceived by a group of infectious disease epidemiologists, at universities around the world, to provide daily updates to decision-makers on how people are moving and where they live, in order to help health organizations to improve the effectiveness of their campaigns and epidemic response.

[14] https://gitlab.com/bigirqu/ArCOV-19

[15] https://visualization.covid19mobility.org/

Table 1: Source of collected data and type of language used in related works.

Collected data	**Group**		**Social media platform**
	Monolingual	Multilingual	
Haouari et al. (2020)	X		Twitter
Alqurashi et al. (2020)	X		Twitter
Alshaabi et al. (2020)		X	Twitter
Singh et al. (2020)		X	Twitter
Chen et al. (2020a)		X	Twitter
Kleinberg et al. (2020)	X		Twitter
Lopez et al. (2020)		X	Twitter
Perrotta et al. (2020)		X	Facebook
Boberg et al. (2020)	X		Facebook
Zarei et al. (2020)		X	Instagram
Gao et al. (2020)		X	Twitter And Weibo
Cinelli et al. (2020)	X		YouTube, Instagram, Twitter, Reddit, Gab

4. Topic modelling: Latent Dirichlet Allocation

Topic modelling is one of the unsupervised machine learning methods, widely used in natural language processing, used for discovering hidden semantic structures, known as "topics", in a text document. "Topics" mean the hidden relations that link words in a vocabulary and their occurrences in documents. Topic modeling Jelodar et al. (2019) seeks to find key concepts throughout the whole corpus and annotate the documents of the corpus based on these concepts. It provides a useful view of a large corpus in terms of individual documents and relationships between them. Latent Dirichlet Allocation (LDA), proposed by Blei et al. (2003), is one of the most popular and recent topic modeling techniques which it is exploited to enhance social media topic coherence Blair et al. (2020). It is an unsupervised probabilistic generative technique for modeling text documents in a given text corpora as mixtures of latent topics based on Bayesian models. Each document is represented as a probabilistic distribution over latent topics and that per-document topic distributions share a common Dirichlet prior. Each topic in the LDA model is defined as a probabilistic distribution over words, and those per-topic word distributions share a common Dirichlet prior as well. Figure 1 shows the LDA topic modeling process.

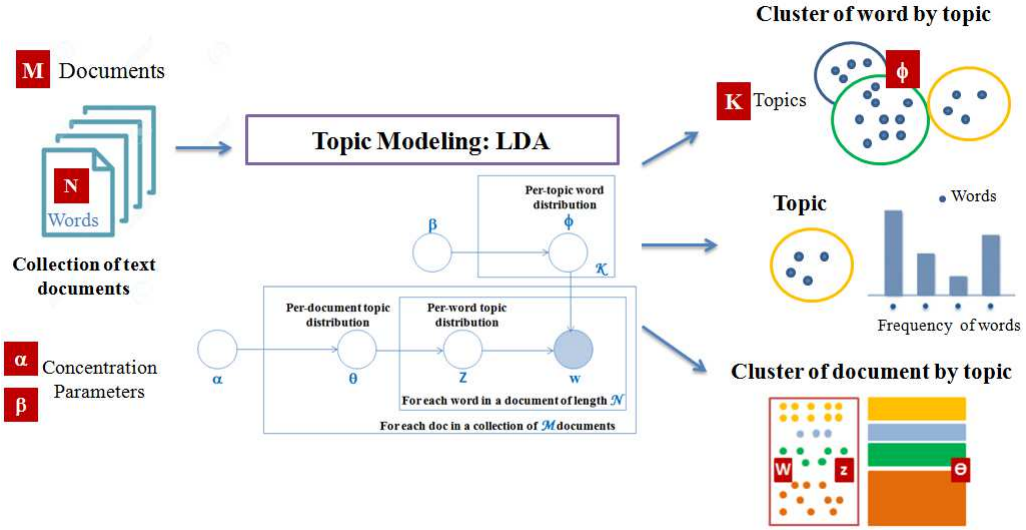


Figure 1: LDA topic modelling process

Given a collection of M documents, each document d is composed of N_d words, with $d \in \{1, .., M\}$. In order to model the collection of documents, the LDA generative process Jelodar et al. (2019) is executed as follow:

- For each topic t ($t \in \{1, .., K\}$), select a multinomial distribution ϕ_t whose hyper-parameter follows the Dirichlet distribution.
- For each document d ($d \in \{1, .., M\}$), select a multinomial distribution θ_d having an hyper-parameter α which follows the Dirichlet distribution.
- For each word w_n ($n \in \{1, .., N_d\}$) in document d,

  i) Choose a topic z_n from θ_d

  ii) Choose a word w_n from ϕ_{zn}

All words included within M documents are observed variables while the other components, composed of topics $\phi_t \forall t \in \{1, .., K\}$ the per-document topic distribution $\theta_d \forall d \in \{1, .., M\}$ and the per-word topic distribution are not known. These latter items are denoted as hidden variables which are predicted from

the analysis of observed variables, i.e. data. The last two variables α and are hyper-parameters. The corpus probability is expressed as follow:

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d \qquad (1)$$

Using LDA-based topic modelling, terms in the set of documents are used to generate vocabulary which is used to discover hidden topics in a large corpus. Document is represented as a mixture of topics where a topic is defined by a probability distribution over the set of terms. In our case, the corpus is composed of a set of Facebook posts as short texts He et al. (2020) so that each post represents a text document and the LDA-based topic modelling is applied on them in order to discover the main discussed topics during COVID-19 disease.

5. Facebook-based COVID-19 tracking trends system

In this section, we describe the proposed system and its different components and tools. As it is already mentioned, the system extract multilingual information. Figure 2 illustrates the different components from the collection and posts extraction to the topic modeling and visualization.

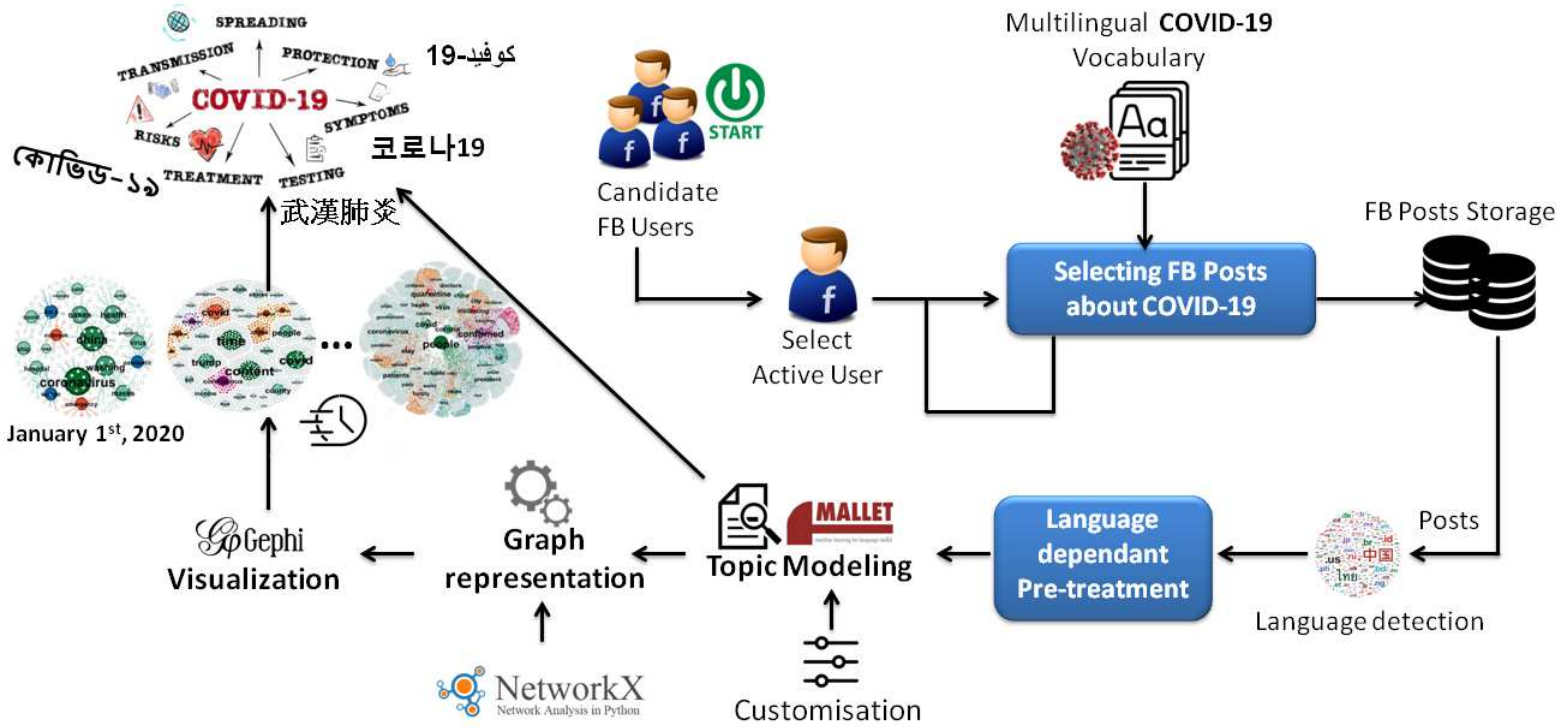


Figure 2: Facebook-based COVID-19 tracking trends evolution system

5.1. Selecting active Facebook users and extracting public posts

The starting dataset covers Facebook users from the whole world. Users are collected during our previous research work Amara et al. (2017). The Facebook users are examined for identifying active users based on the publication rate since first January 2020. The selected users are candidates for having published on COVID-19 context.

The candidate users are explored for scrapping the published posts about the COVID-19 by checking the existence of the built vocabulary. Then, each post is stored after extracting the URL from the post. In fact, the Facebook user can provide his own content, share a content with a comment or in its initial format. So, these posts will be used for building an incremental idea about the COVID-19 along the time period starting from the first January 2020. As it is known, the politic access to the information is so close when it is compared to Twitter as example. Therefore, we design some patterns to locate the public posts with the time related information. The target posts are selected based on multilingual COVID-19 vocabulary built using machine translation.

5.2. Multilingual topic modelling

The proposed work is focused on multilingual context of COVID-19. In fact, the keyword *COVID-19* is exploited in its written format by several languages using Latin letters. The gathered data pertains to users in the whole word (see Figure 3) in different languages. The language and the target url pages are identified. We exploit the language detection library implemented in plain Java and covering more than 50 languages. This library[16] is open source Apache license 2.0 and it calculates language probabilities from features of spelling through naïve bayes with character n-gram. It generates the language profiles from training groups. A profile is the set of probabilities of all spelling in each language.

[16] `https://code.google.com/archive/p/language-detection`

We exploit the Latent Dirichlet Allocation (LDA) Blei et al. (2003) that it is considered as "generative probabilistic model" of a collection of composites made up of parts. Its uses include mainly topic modelling. The composites are documents and the parts are words and/or phrases. In our study, the documents are the published posts.

The computing process of the probabilistic topic model estimated by LDA consists of two tables (matrices). The first table describes the probability of selecting a particular word when sampling a particular topic. The second table describes the chance of selecting a particular topic when sampling a particular post. The LDA algorithm is composed of the following detailed steps in relation with our context:

i. Select the unique set of words.

ii. Select the set of posts according to a specific language.

iii. Fix how many parts you want per posts (sample from a Poisson distribution).

iv. Specify the number of topics as outputs.

v. Affect a number between not-zero and positive infinity to the parameter alpha.

vi. Affect a number between not-zero and positive infinity to the parameter beta.

vii. Build the 'words-versus-topics' table. For each column, the beta is used as input for the Dirichlet distribution (which is a distribution of distributions). Each sample will fill the columns of the table and give the probability of each word per topic (column).

viii. Build the 'posts-versus-topics' table. For each row, the parameter alpha is used by Dirichlet distribution as the input. Each sample will fill the rows of the table and give the probability of each topic (column) per posts.

ix. Build the actual posts. For each of them, a) look up its row in the 'posts-versus-topics' table, b) sample a topic based on the probabilities in the row, c) go to the 'words-versus-topics' table, d) look up the topic sampled, e) sample a part based on the probabilities in the column, f) repeat from step 2

until reaching how many words the post was set to have.

x. Moreover, the number of words representing the topic can be fixed at the beginning of the algorithm. The number of topics can be viewed as a number of clusters and the probabilities as the pertaining degree to the cluster. LDA process plays the role of soft-clustering between posts and the words composing the topics.

Posts follow a pre-processing step for preparing data to the topic modelling process. In fact, this includes mainly the remove of html tags and the stop words according to its own language. Mallet[17] is the tool exploited in this study to realise the topic modelling task McCallum (2002). MALLET is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text. Many of the algorithms in MALLET depend on numerical optimization. In fact, unsupervised topic modelling is useful for analyzing large collections of unlabeled text. The MALLET topic modeling toolkit contains sampling-based implementations of Latent Dirichlet Allocation. The toolkit is open source software, and is released under the common public license.

The topic modelling module provides as output the topic composition of posts P_{Topics} and the top k words for each topic among N topics (where N and k are predefined) T_{words}. This output can be useful for checking that the model is working as well as displaying results of the model. In addition, it reports the Dirichlet parameter of each topic. If hyperparamter optimization is turned on, this number will be roughly proportional to the overall portion of the collection assigned to a given topic.

5.3. Building graph representation

The two matrices provided by Mallet are used for providing a graph representation allowing us in next step to handle the topics-words based graph through graph processing tools. This module is implemented using python and

[17] `http://mallet.cs.umass.edu/index.php`

NetworkX [18] which is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks and it is distributed as open source 3-clause BSD license. This module generates a file GEXF[19] (Graph Exchange XML Format). This graphic data can be exchanged from one application to another due to *.gexf*. Various operating systems can be used to open *.gexf* files and it only requires appropriate applications such as Gephi and gexf4j. This step provides two types of graph representation to the topic modelling output. For the former, it provides only the connection between the post and its corresponding topic based on the highest contribution of the composite to the topic. For the second, an edge is built between the topic and each post according to the participation weight to form the topic. We choose the graphic representation of the topic modelling output to express in meaningful way the evolution of the COVID-19 centric interests since the beginning of the 2020 year and in multilingual framework. The GEXF file is treated by the Gephi[20] as open graph viz platform Cherven (2015) for offering good specialization layout and a range of specific algorithms for graph handling and topological parameters extraction. The topics are represented by one word by customizing the Mallet tool to provide topics with 3 words. Next, to resolve the problem of exploiting the same word in different contexts due to its polysemous nature, we visit the topics from the higher probable to the last one, and we explore the words used to represent the topic until finding not used word by the already visited topics.

Algorithm 1 depicts the different steps for building the graph GEXF using the output of the multilingual topic modelling module. In fact, Algorithm 1 receives as inputs the P_{Topics} representing the distribution of Facebook Posts on the extracted topics and T_{words} containing the N extracted topics and the K words for each topic. The function $selectSingleWordTopic$ at Instruction

[18] `https://networkx.github.io/`

[19] GEXF is a language for describing complex network structures, their associated data and dynamics. This extension is used to describe files containing graphic and visualization data.

[20] `https://gephi.org/`

Algorithm 1 Generate Graph GEXF Topic Modelling

Input: P_{Topics} the contributions of Facebook Posts through the extracted topics, T_{words} where each topic among N extracted topics is represented by K words

Output: *GraphPostsTopics* the graph structure in format GEXF

$T1_{words} \leftarrow selectSingleWordTopic(T_{words})$
$weights \leftarrow read(P_{Topics})$
$topics \leftarrow read(T1_{words})$
$post\_topic\_weights \leftarrow Dictionary()$
for $weight \in weights$ **do**
    $post\_topic\_weights \leftarrow calculate\_edge\_weights(post\_topic\_weights, weight[0], weight[1], weight[2])$
        $//weight[0]$: id of the Post
        $//weight[1]$: path of the Post file after its pre-processing
        $//weight[2]$: series of values representing the participation of the Post to each topic
end for
$GraphPostsTopics \leftarrow networkX.createGraph()$
for $post \in post\_topic\_weights.keys()$ **do**
    GraphPostsTopics.add_node(post)
end for
for $topic \in topics$ **do**
    GraphPostsTopics.add_node(id=topic[0], label=topic[2], vizualisation=topic[1])
        $//topic[0]$: number of the topic among K topics
        $//topic[1]$: topic weight
        $//topic[2]$: word representing the topic
end for
for $post \in post\_topic\_weights.keys()$ **do**
    $idTopicMax \leftarrow 0$ ▷ it contains the *id* of the topic which *post* having the highest participation in it
    for $tid \in post\_topic\_weights[post].keys()$ **do**
        if $post\_topic\_weights[post][tid] > post\_topic\_weights[post][idTopicMax]$ **then**
            $idTopicMax \leftarrow tid$
        end if
        GraphPostsTopics.$add\_edge(idPostMax, post, weight = post\_topic\_weights[post][idTopictMax])$
    end for
end for

1 selects the single word representing the topic among K words. The process is based on ranking the topics according to their weights. Then, it selects the word does not exist among the already selected word by traversing topics from the highest weighted topic to the last one. The entities *weights* (Instruction 2) and *topics* (Instruction 3) are triplets as described within the algorithm. The dictionary (Instruction 4) is a general-purpose data structure for storing a group of objects. A dictionary has a set of keys and each key has a single associated value. When presented with a key, the dictionary will return the associated value. The dictionary *post_topic_weights* is filled using the for loop (instruction 5) such as the Keys are the ids ($weight[0]$) of the posts and Values are the weights contributions of a post to the extracted topics. The building of the graph $GraphPostsTopics$ is based on adding nodes (Instructions 10 and 13) and linking them with edges (Instruction 21).

6. Dataset presentation

This section is devoted to present the multilingual dataset of posts extracted from Facebook. This dataset is in continuous evolving and can be followed through the github web site[21]. This presentation covers several dimensions which are detailed in next paragraphs.

6.1. Dataset statistics

The gathered data covers the period since January 1st, 2020 to May 15th, 2020. Figure 4 shows the reparation of the number of posts in relation with their date of posting. The first two months characterize the beginning confrontation with the new virus and the new disease started from Wuhan. The virus, back then, is not known and the disease is not spread around the world. In fact, it is just localised in China. Therefore, the gathered data is few except in Japan because the disease arrived early. After the first stage, there have been

[21] `https://mohamedalihadjtaieb.github.io/Covid19-based-Facebook-Research/`

an explosion in the number of posts coordinately with the coronavirus crisis in Italy.

Figure 3 presents the geo-location coverage of the users. We tried to cover the maximum of countries where people practice different languages.

Figure 3: Distribution of the COVID-19 related data through countries around the world

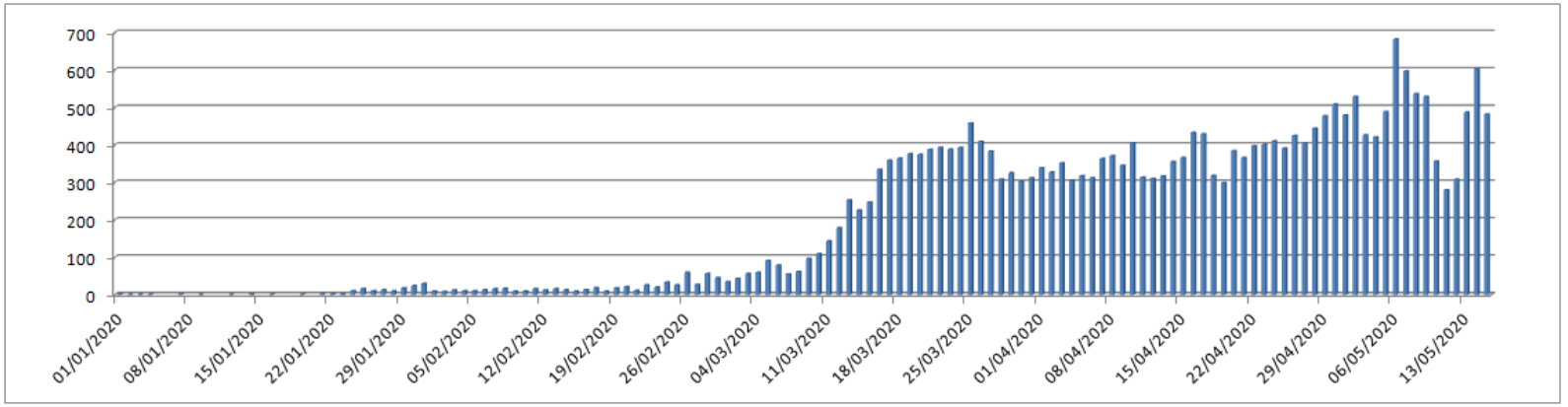


Figure 4: Distribution of the gathered Facebook public posts through the time since January 1st, 2020 to May 15, 2020

Table 2: Distribution of the collected COVID-19 centred Facebook posts through the time and in relation with treated languages.

	January - February 2020	**March 2020**	**April 2020**	**15 May 2020**
EN	261	3870	8541	10294
AR	55	817	2124	2907
DE	7	288	819	1546
FR	12	271	1021	2267
ES	14	338	1012	1594
IT	44	504	1394	1981
JA	230	707	1383	1790

6.2. Multilingual aspect

Table 2 shows the number of gathered posts in relation with each language and in cumulative way. As shown in Figure 3, the languages do not mean a specific country, but the most publications may belong to a particular country such as English from United States, French from France, Italian from Italia, Spanish from Spain and Dutch from Germany. As for Arabic is from several countries. It is important to note that the same users who provided few posts about COVID-19 (except for English and Japanese) during January and February, are the same who provide this respectful number of posts. But then the number of posts increased significantly as the crisis worsened, and this depicts the degree of anxiety that people have reached. It is also important to remember that the collected Facebook posts are those shared with public and not in private way.

The multilingual topic modelling is based on specific lists of stop words for each language: English[22], Arabic[23], Spanish[24], Italian[25], German[26], French[27] and Japanese[28].

[22] https://gist.github.com/sebleier/554280
[23] https://github.com/mohataher/arabic-stop-words/blob/master/list.txt
[24] https://github.com/stopwords-iso/stopwords-es/blob/master/stopwords-es.txt
[25] https://github.com/stopwords-iso/stopwords-it
[26] https://github.com/stopwords-iso/stopwords-de/blob/master/stopwords-de.txt
[27] https://github.com/stopwords-iso/stopwords-fr/blob/master/stopwords-fr.txt
[28] https://github.com/stopwords-iso/stopwords-ja/blob/master/stopwords-ja.txt

6.3. Facebook users COVID-19 related behaviour

Figure 5 shows the percentages of the number of posts on COVID-19 in relation to the percentage of Facebook users who were the subject of the research who achieved a percentage equal to or greater than the abscissa x. In fact, 16.06% of users have 50% of their public posts about COVID-19. We have recorded the highest score of 541 posts on COVID-19 among 1245 posts during the studied period.

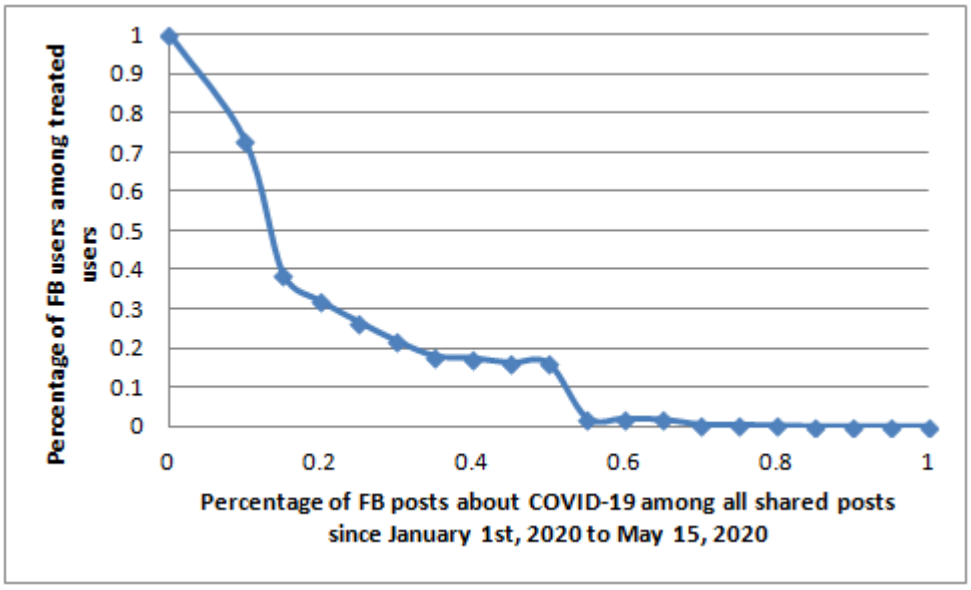


Figure 5: Correlation between percentage of COVID-19 Facebook posts during the period from January 1st, 2020 to May 15, 2020, and the percentage of users having post more than the fixed percentage.

On another point, and a return to the users whom we singled out for the analysis, we find that 2% of them returned to using their Facebook accounts after their last public post before January 1st, 2020 dated back to the period between 2012 and 2018. This indicates the state of alert and suspicion of the pandemic, which prompted a number of users to return to follow and to share the news of the virus on social networking pages, especially with quarantine.

The pearson correlation between the number of COVID-19 centred posts and the total number of published posts since January 1st, 2020 is equal to

0.62 which leads to significant correlation between the two parameters. This means that for the majority of users the number of COVID-19 posts follows the increasing of the number of total posts.

Missier et al. (2017) showed that social media analytics can be used to pinpoint individuals who are actively contributing to social discourse on the specific topic of the Zika virus and its consequences, and are thus likely to be sensitive to health promotion campaigns, by focusing on Twitter content related to the Zika virus and its effect on people. For our case, also, we find some users who can be qualified as COVID-19 publish engine despite that their accounts are for persons. In fact, 1.40% of users, considered in this research, published more than 200 public posts on COVID-19 since January 1st, 2020. They participate with 24.46% among the hole gathered data.

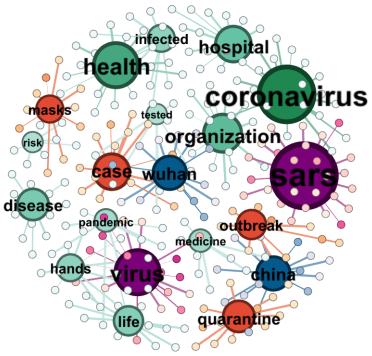
7. Analysis of COVID-19 trends

We present in this section the results of the proposed architecture for tracking the crowdsourcing COVID-19 centred trends based on social data provided by Facebook users about this pandemic. The discussion is focused on 7 different languages (English (EN), Arabic (AR), French (FR), Spanish (ES), Italian (IT), German (DE) and Japanese (JA)) and it explores the evolution over time of the users' interests from the start (January 1, 2020) of the coronavirus crisis until May 15, 2020. The differences figured in the countries represented by the languages will be highlighted. Moreover, the interests differ along with the diversity of stakeholders during this serious crisis having launched an emergency status. The results are visualized in graph structure according to the importance degree of the extracted topics linked to the posts having mostly participated to form a specified topic. The study is divided mainly on 3 periods according to the outbreak of the virus SARS-COV-2 around the world. The first period covers January and February and characterized by the acquaintance stage, the second period along the months March and April which includes the shock stage and radical change in daily life, and for the third until 15 May 2020 with the

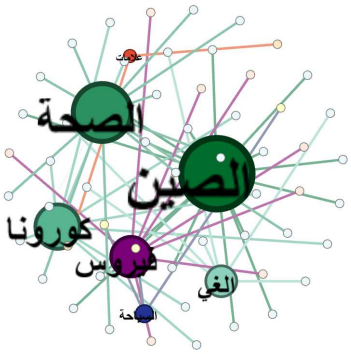
beginning of returning to normal life taking into consideration the coexistence with the virus.

7.1. Analysis of first period: January-February 2020

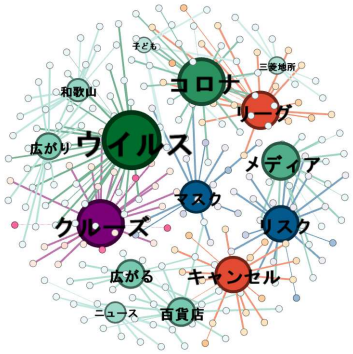
Figure 6 shows the first topics involved in the social network Facebook in relation with different languages. Mainly, users start exploiting the virus name [*Coronavirus* (EN), كورونا (AR) and コロナ (JA), *sars* and *ncov*]. Everyone felt, to varying degrees, that an emergency would affect all countries of the world [*allarmismi* (IT), *urgence* (FR)]. Everyone was afraid of the deadly disease, and this was evident in the results regarding Italy, which has struck the world with the number of dead people before the disease spread to other countries [*murdered* (DE)]. Moreover, they follow the mobilization of the official authorities [*Gobierno* (ES), *autorità* (IT)] especially the health sector [الصحة (AR), *medico* (IT), *hausarzt* (DE)]. They talk also about its origin China and especially Wuhan [الصين (AR), *asie* (FR), *cina* (IT)]. Users are also watching its outbreak [*épidémie* (FR)]. Facebook surfers and parents [*ëltern"* (DE)] were observing the movement of travel for fear of the spread of the disease and the closure of land and sea borders as a result of the pandemic of coronavirus [السياحة (AR), *viajes,vuelve,puerto* (ES), クルーズ (JA)]. Infection has spread between countries through the movement of travelers. This created a global crisis and the desire of the various countries to close their borders and led to the emergence of the problem of the stuck with the lack of freedom of movement. Authorities around the world had resorted to canceling [キャンセル (JA), الغي (AR)] many events. Eastern Asia countries (Japan and South Korea) are the first places where the infection started early which gives explanations to the quite richness of information for the English and Japanese languages. Therefore, we find them touched on the means of protection [マスク (JA), masks, hands (EN)] with the medical procedures [*quarantine* (EN), キャンセル (JA)] that were followed and began to talk about social separation.


infected
hospital
health
coronavirus
masks
tested
organization
risk
case
wuhan
sars
disease
pandemic
outbreak
medicine
hands
virus
china
life
quarantine

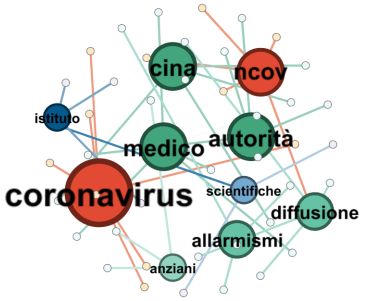
EN


الصحة
الصين
كورونا
فيروس
الغي

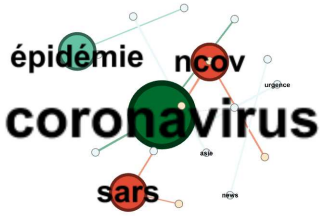
AR


コロナ
和歌山
リーグ
広がり
ウイルス
メディア
リスク
クルーズ
広がる
キャンセル
ニュース
百貨店

JA


cina
ncov
istituto
medico
autorità
coronavirus
scientifiche
diffusione
allarmismi
anziani

IT


épidémie
ncov
coronavirus
sars

FR

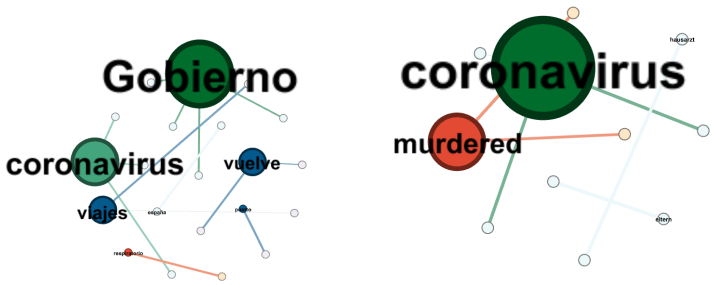


Figure 6: Facebook-based crowd-sourced COVID-19 trends covering 7 languages (EN, AR, FR, ES, IT, DE and JA) for the period January and February 2020.

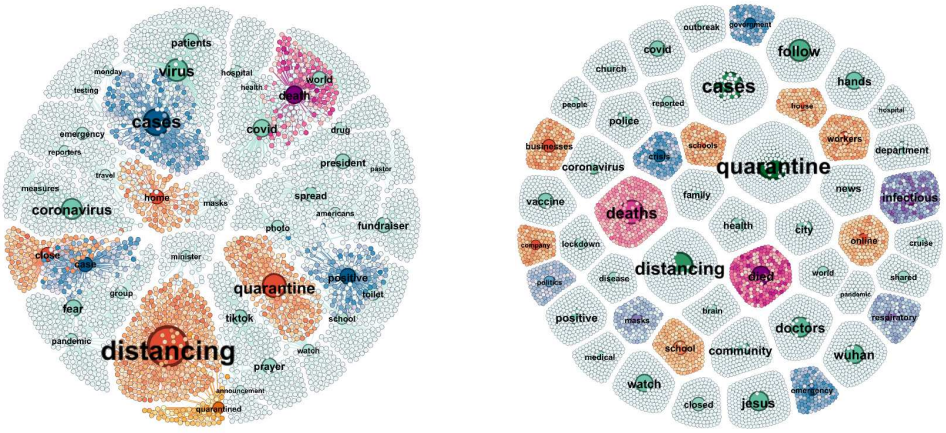
7.2. Analysis of second period: March-April 2020

Regarding the second stage of the pandemic, after the ability of East Asian countries such as Japan and South Korea to control the spread of the disease, the world was shocked by the number of victims in Italy (and in a future stage in the United States of America) [イタリア (JA)] with the beginning of the rapid spread in France and Spain and the emergence of the first cases in most Arabic countries. All of this led to health measures taken by countries through quarantine, the suspension of air traffic and most economic activities, with the cancellation or postponement of most events.

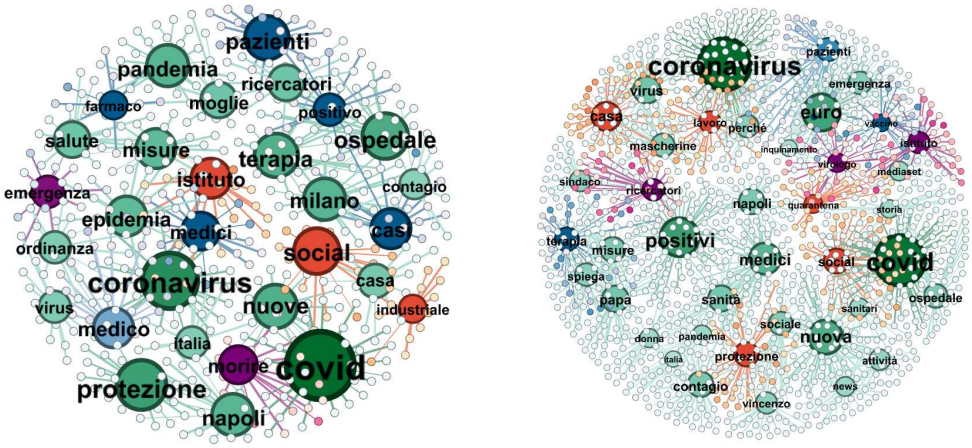
The second stage spanned the months of March and April and was marked by a state of alert with the official authorities, with a general feeling of the seriousness of the situation and the exacerbation of the crisis [リスク (JA), *emergenza* (IT),), *crisis* (ES), *crise* (FR), *ausnahmezustand*, *coronakrise*, *gefährlicher* (DE)]. The most shocked to people is the ability of the virus COVID 19 to inflict casualties[*emergenza* (IT), *fallecido* (ES), وفاة (AR), *morts,décès,deuil* (FR),

deaths (EN)], and therefore we find that this topic was addressed through Facebook browsers with the names of some of the well-known persons on a national or global level. Figure 7 depicts the chronological development of subjects and the trends that people transmit about the virus in the various languages that represent many countries. It is not only interested in the months of March and April, but focuses on the cumulative nature of the discourse extending from the beginning of the year 2020 to the mentioned time limit.

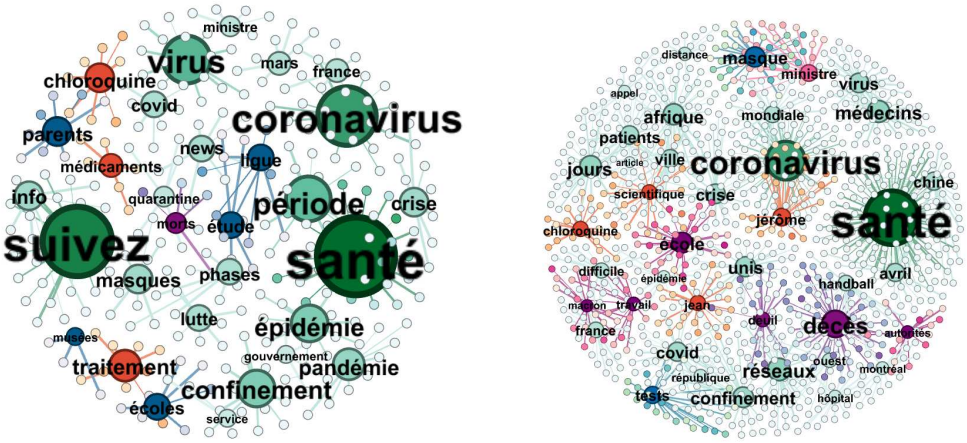
It also appears in our analysis, which is consistent with the reality that we lived through, that countries mobilized their political and financial capabilities to confront the pandemic, and societies felt that they were in a state of war, and this stimulated the use of vocabulary from the glossary of battles, as was shown in various languages [مكافحة، محاربة (AR), に対し (JA), *lucha*, *ejército* (ES) , *lutte* (FR)]. Citizens also pursued their governments, politicians, and leaders, and this justifies the emergence of issues that concern governments and the political class [大統領 (JA),*sindaco* (IT), *presidente*, *gobierno* (ES), *ministre*, *gouvernement, autorités* (FR), *behörde*,*bund*, *bundesregierung*, *regierung* (DE), الحكومة، وزير، وزارة (AR), *president*, *government*, *politics* (EN)].



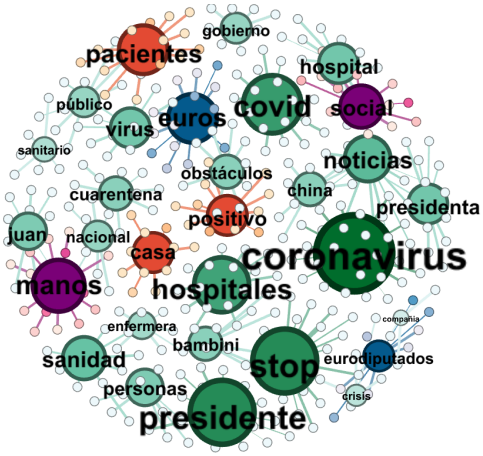
EN

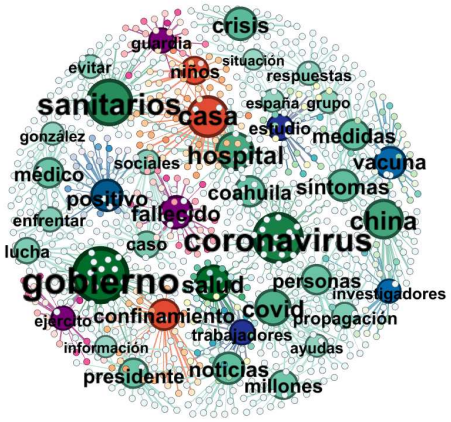

pazienti
pandemia
ricercatori
moglie
farmaco
positivo
salute
misure
terapia
ospedale
istituto
emergenza
contagio
milano
epidemia
medici
casi
ordinanza
social
coronavirus
casa
virus
nuove
industriale
medico
italia
morire
covid
protezione
napoli
coronavirus
pazienti
virus
emergenza
casa
euro
lavoro
perché
vaccino
mascherine
inquinamento
istituto
virologo
mediaset
sindaco
ricercatori
napoli
quarantena
storia
terapia
misure
positivi
medici
social
covid
spiega
papa
sanità
ospedale
sanitari
sociale
donna
pandemia
nuova
protezione
italia
attività
contagio
vincenzo
news

IT

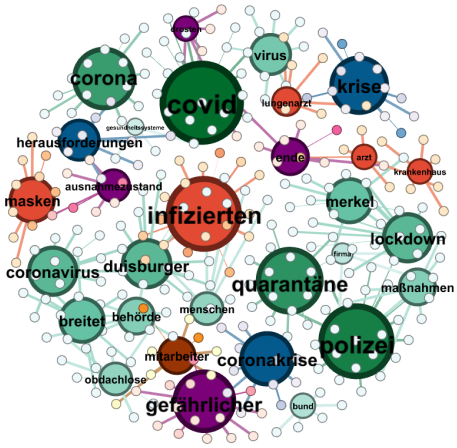

ministre
virus
mars
france
chloroquine
covid
coronavirus
parents
news
ligue
médicaments
info
quarantine
période
crise
mons
étude
suivez
santé
masques
phases
lutte
épidémie
mesures
traitement
gouvernement
pandémie
confinement
écoles
service
distance
masque
ministre
virus
appel
afrique
mondiale
médecins
patients
jours
article
ville
coronavirus
chine
scientifique
crise
jérôme
santé
chloroquine
école
difficile
épidémie
unis
avril
handball
macron
travail
jean
deuil
décès
autorités
france
covid
ouest
république
réseaux
montréal
tests
confinement
hôpital

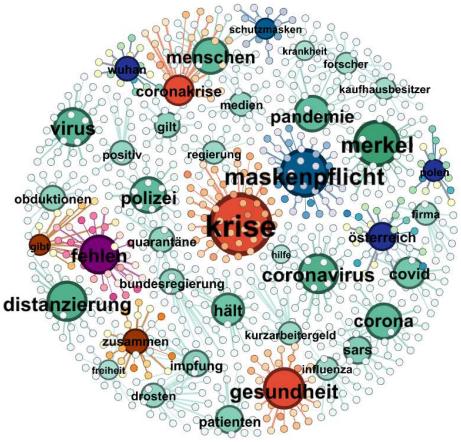
FR


gobierno
pacientes
hospital
público
covid
social
euros
virus
sanitario
noticias
obstáculos
cuarentena
china
presidenta
positivo
juan
nacional
casa
coronavirus
manos
hospitales
enfermera
compañía
bambini
sanidad
stop
eurodiputados
personas
crisis
presidente

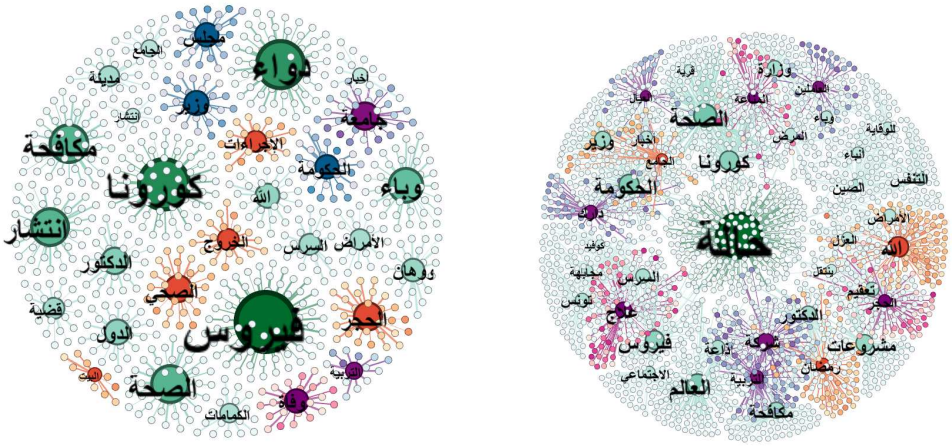

crisis
guardia
evitar
situación
niños
respuestas
sanitarios
casa
españa
grupo
estudio
medidas
gonzález
hospital
vacuna
sociales
médico
síntomas
coahuila
positivo
fallecido
enfrentar
china
caso
coronavirus
lucha
gobierno
salud
personas
investigadores
covid
ejército
confinamiento
propagación
trabajadores
información
ayudas
noticias
presidente
millones

ES

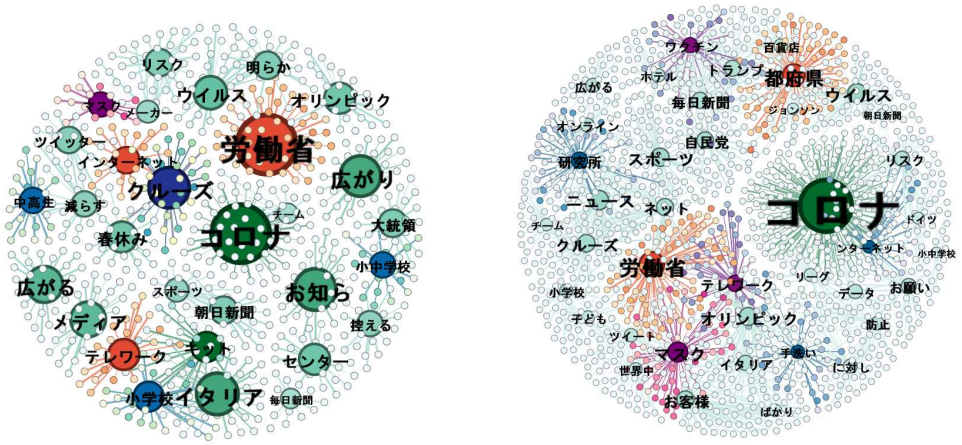

virus
corona
krise
covid
lungenarzt
herausforderungen
ende
arzt
krankenhaus
masken
ausnahmezustand
infizierten
merkel
lockdown
coronavirus
duisburger
quarantäne
maßnahmen
breitet
behörde
menschen
polizei
mitarbeiter
coronakrise
obdachlose
gefährlicher
bund


schutzmasken
krankheit
menschen
wuhan
forscher
coronakrise
kaufhausbesitzer
medien
pandemie
virus
gilt
merkel
positiv
regierung
maskenpflicht
polizei
obduktionen
polizei
krise
firma
österreich
fehlen
quarantäne
hilfe
coronavirus
covid
bundesregierung
distanzierung
hält
corona
zusammen
kurzarbeitergeld
sars
impfung
influenza
freiheit
drosten
gesundheit
patienten

DE



AR



JA

Figure 7: Facebook-based crowd-sourced about COVID-19 trends covering 7 languages (EN, AR, FR, ES, IT, DE and JA) for the period March (left part) and April 2020 (right part).

One of the most important measures taken to combat the virus is quaran-

tine [*confinement* (FR), الحجر، العزل (AR),*quarantäne* (DE), *cuarentena*, *confinamiento* (ES), *quarantena* (IT), *lockdown*, *quarantine* (EN)], which requires people to stay in their homes to limit the spread of the disease. In fact, restricting mobility is a primary method being used to slow down the spread of COVID-19. Therefore, the topics that focus on this axis are considered one of the most exciting topics in all the languages studied in this research [*casa* (IT), *casa* (ES), دارك، البيت (AR), *house* (EN)].

It is also worth noting children [العيال (AR), 子ども (JA), *niños* (ES)] as a common topic between different browsers in different languages. This is mainly related to the quarantine procedure, which resulted in the suspension of studies for all educational levels and making parents busy with the end of the school year [جامعة ، التربية (AR), 小学校 (JA), *estudio* (ES), *écoles*, *étude* (FR), *school* (EN)]. Hundreds of countries have implemented nationwide school closures and many other countries are implementing localized school closures. UNESCO[29] estimates that these closures mean approximately 1.2 billion students, roughly 74% of all enrolled students worldwide, are experiencing a disruption in their education.

In addition to what has been mentioned, we notice the presence of topics in relation to what may be called a stock exchange of cases [*casi* (IT), *caso* (ES), حالة (AR), *tests* (FR)] mainly related to medical analyzes (called tests) that are conducted and whose result is positive due to their pregnancy or disease negative [*positiv* (DE), *positive* (EN), *positivo* (ES), *positivi* (IT)]. These data were linked to several systems that were developed to monitor the spread of the disease around the world. It is noticed that the users of social media platforms, especially the Facebook in question, tend to follow the number of patients and victims and they share this information and comment on it.

In the context of searching for appropriate drug [*médicaments*, *traitement* (FR), *drug* (EN), *farmaco*, *terapia* (IT), دواء (AR)] and vaccination [ワクチン (JA), *vaccino* (IT), *invacuna* (ES), *impfung* (DE), *vaccine* (EN)] to pre-

[29] https://en.unesco.org/covid19/educationresponse

vent the coronary epidemic, an urgent need has emerged for a feasible scientific research [*forscher* , *lungenarzt* (DE), *scientifique* (FR), *investigadores* (ES), *ricercatori* , *virologist* (IT)] and Internet browsers, including Facebook, have shown interest in news of scientific discoveries towards a more understanding of the emerging disease and to follow the progress of research and tests to produce the vaccine. Accordingly, the analytical images showed that a wide range of publications touched on this issue through many topics. For example, *chloroquine* was mentioned especially in publications in the French language, due to the controversy caused by the debate about the effectiveness of this drug or not.

As for the economic impact of the Covid-19 epidemic [euro, *industriale* (IT), *euros* , *compañia* (ES), *firma,* شركة *(AR)* (DE)], the economic activities of many people around the world have been disrupted as an inevitable consequence of quarantine. This is what led to the emergence of many topics about the axis of work [労働省 (JA), *lavoro* (IT), *workers* (EN), *mitarbeiter* (DE), *travail* (FR), *trabajadores* (ES)], which highlights the busyness of surfers with their livelihoods and their fear of entering into a financial crisis as a result of the disruption of their activities or for fear of losing work in the form of prolonging the crisis for a long time. Accordingly, the term "remote work" [テレワーク (JA)] appeared.

And in relation to the users' interest in the prevention axis [防止 (JA), للوقاية (AR), *protezione* (IT)], the focus was on three measures, namely continuous cleaning [手洗い (JA), نغسل (AR), *hands* (EN), *manos* (ES)] and the use of protective masks [マスク (JA), الكمامات (AR), *masks* (EN), *maskenpflicht, schutzmasken, masken* (DE), *masque* (FR), *mascherine* (IT)] with social distancing (including self-isolation) [*distancing* (EN), *distanzierung* (DE), *distance* (FR)]. Focusing on these three points in all publications in different languages is an obligation of the user to follow the advice, but it is also a request from the other to commit to preserving the health of the group that the individual is part of and cannot be active without the participation of others.

As for the axis of communicating with the outside world during the quarantine and the desire to follow all that matters the epidemic as a benefit, many

topics expressed the interest of Facebook users in that. Accordingly, we note the existence of a dictionary of news [ニュース (JA), *news* (EN)(IT)(FR), *noticias* (ES), أنباء ، أخبار] and its sources like social networks [ツイート (JA), *sociales* (ES), *réseaux* (FR)] and others [اذاعة (AR), *mediaset* (IT)], as users tend to share everything new about the disease. Also, the use of the Internet [ンターネット, データ , オンライン (JA)] appeared clearly in order to follow the news, entertainment and work from afar.

There is another dimension common to all languages, which is the religious and faith dimensions [الله، الجامع ، رمضان (AR), *jesus*, *prayer*, *church* (EN),*papa* (IT)], which we have identified in many subjects. This is mainly due to the feeling of fear and distrust of the future with the large numbers of deaths in many countries such as the United States, Italy, Spain and France. Solidarity [*sociale* (IT), *social* , *ayudas* (ES), *obdachlose, gibt, hilfe , zusammen* (DE), الاجتماعي (AR)] between people also had a share of the extracted subjects, and the intention is to help the needy, from the vulnerable groups, health and financially.

The so-called *rest of the globe* in front of industrial activities largely disrupted, which led to a low level of pollution and improved air quality [30] in countries that followed a general health quarantine for the various productive sectors. This issue appeared on the Italian *inquinamento* language data for April.

7.3. Analysis of third period: End April-15 May 2020

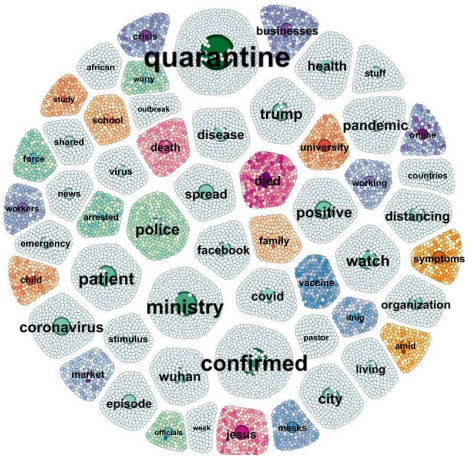
By analyzing the data collected until 15 May, as it is illustrated in Figure 8, we can see the emergence of other issues related to the policy of gradual quarantine lifting (or named Targeted quarantine in Arabic countries such Tunisia) [*déconfinement* (FR), الحجر (AR),*lockerungen*, *öffnen*, *freiheit* (DE), *salir* (ES), *ripartire* (IT)] in many countries that have been affected by the COVID-19 pandemic. This manifests itself by mentioning activities related to the gradual exit

[30] https://www.latribune.fr/opinions/tribunes/quand-les-effets-du-coronavirus-se-voient-depuis-l-espace-841719.html

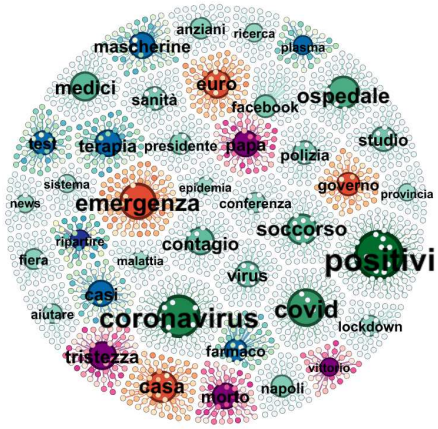
of people to their normal lives [*tierpark*, *maßnahmen*, *schulen*, *Kinos* (DE), *colegio*, *laboral* (ES), المهن (AR), *école*, *plages*, *travail* (FR), *market* (EN)], while maintaining the health measures that have appeared in other topics. As noted, many topics that were in March and April, indicating the closure, with the beginning of May, they have the significance of a gradual opening.

Topics have emerged that support what was discussed in relation to the analytical data for the months of March and April using new terms in different languages indicating meanings closely related to what was discussed above. But what remains is the emergence of topics related to the gradual exit from quarantine or the desire of people to return, albeit in a simple way, to normal life. And this exit is followed by health measures (as already detailed in previous section) that must be applied, which also appeared in the level of many topics and to varying degrees among languages.

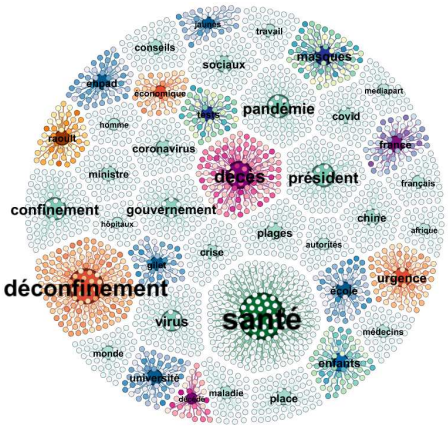
Another issue that has emerged clearly the analysis of the third period is the police [*force* (EN), *polizia* (IT), *policía* (ES), *bußgeldkatalog* (DE)], which monitors law enforcement and deters violators of Salt's official recommendations. It had a great role in many countries to impose quarantine and health measures to be followed with the spread of COVID-19 disease. It should also be mentioned many topics in the Japanese language related to technology, which can be summarized in a word デジタル that means *Digital*.


businesses
crisis
quarantine
health
stuff
african
worry
study
outbreak
school
trump
pandemic
online
shared
death
disease
university
force
virus
dead
working
countries
news
spread
positive
distancing
workers
arrested
police
family
emergency
facebook
watch
symptoms
child
patient
vaccine
covid
ministry
drug
organization
coronavirus
stimulus
pastor
amid
confirmed
living
market
wuhan
episode
city
officials
week
masks

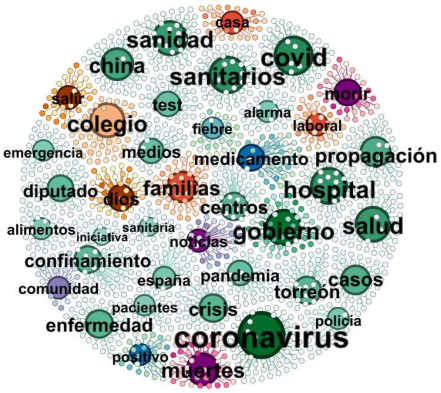
EN


anziani
ricerca
mascherine
plasma
euro
medici
sanità
ospedale
facebook
test
terapia
presidente
papa
studio
polizia
sistema
governo
epidemia
provincia
news
emergenza
conferenza
soccorso
ripartire
contagio
positivi
fiera
malattia
virus
casi
aiutare
coronavirus
covid
lockdown
farmaco
tristezza
vittorio
casa
morto
napoli

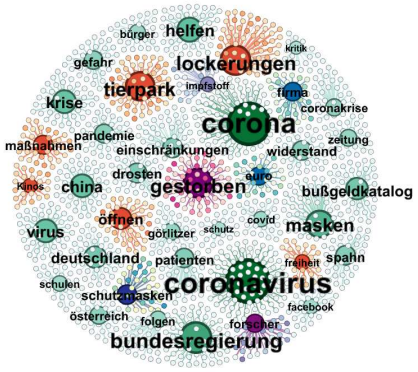
IT


déconfinement
santé
décès
président
pandémie
coronavirus
confinement
gouvernement
virus
crise
urgence
masques
sociaux
conseils
travail
économique
covid
france
ministre
chine
plages
autorités
école
médecins
enfants
monde
maladie
place
hôpitaux

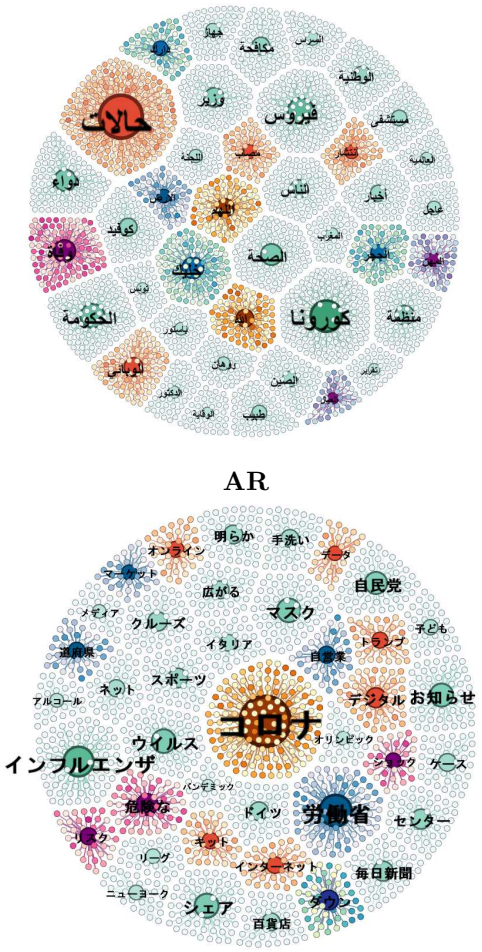
FR


sanidad
covid
china
sanitarios
morir
salir
test
alarma
colegio
fiebre
laboral
emergencia
medios
medicamento
propagación
diputado
dios
familias
hospital
centros
alimentos
sanitaria
iniciativa
noticias
gobierno
salud
confinamiento
españa
pandemia
casos
comunidad
torreón
pacientes
crisis
policía
enfermedad
coronavirus
positivo
muertes

ES


helfen
lockerungen
tierpark
krise
corona
coronakrise
pandemie
maßnahmen
einschränkungen
widerstand
zeitung
china
gestorben
bußgeldkatalog
öffnen
masken
virus
görlitzer
schutz
deutschland
patienten
spahn
coronavirus
schulen
schutzmasken
facebook
österreich
folgen
forscher
bundesregierung

DE



AR

JA

Figure 8: Facebook-based crowd-sourced trends about COVID-19 covering 7 languages (EN, AR, FR, ES, IT, DE and JA) for the period until May 15, 2020.

7.4. Discussion

This section is intended to compare what we have reached in this research with previous research in other issues, despite its limited number. Pruss et al. (2019) work showed how topics can be explored on a more extensive global scale to allow the public health community understand variations in topics across countries in relation with Zika disease. These data contribute importantly to understanding of disease transmission, disease interventions, and public health communication. This work supports the previously reached social media platforms in following up on pandemics. As in Pruss et al. (2019), we have succeeded in simulating the reality of living with the COVID-19 pandemic in many languages, representing many countries. However, we did not rely on translating the publications, but rather we used the publications in the language in which they were written. We have also succeeded in creating a new method for presentation that facilitates the process of analyzing results.

Zhang et al. (2020) investigated the contents of posts about the Zika virus on Yahoo! Answers, identify and reveal subject patterns about the Zika virus, and analyze the temporal changes of the revealed subject topics over 4 defined periods of the Zika virus outbreak. Multidimensional scaling analysis, temporal analysis, and inferential statistical analysis approaches were used in the study.

The present study shows a result closer to the one about Zika directed by Miller et al. (2017) and based on topical analysis concerning what people are tweeting about four disease characteristics and concluding to four main topics (symptoms, transmission, prevention, and treatment) which are existing also for COVID-19.

8. Conclusion and Future Works

This concurrent research with the mobilization of research capabilities across all countries on the COVID-19 pandemic is considered the first of its kind, which depends on the analysis of the data of the social platform Facebook, which is the first platform globally through the number of active users. This research

work focused on multilingualism in seven languages English (EN), Arabic (AR), Italian (IT), Spanish (ES), French (FR), German (DE) and Japanese (JA), representing a large group of countries in order to analyze the cognitive development of people about COVID-19 and the development of topics raised over time from 1 January 2020 to 15 May 2020.

The results of analyzing the data collected and analyzed by adopting three cumulative periods as the first extends from January to February, and the second from January to April with the allocation of the months of March and April each with a graph, while the third period extends 15 May.

This work, which highlighted the ability to simulate the reality of life through the time of COVID-19 in various countries of the world and in many languages. It can have an outlook for emerging issues. This proposed approach also remains valid for exploring other events and gives the possibility for an in-depth analysis of well-selected topics in recursive way. In fact, the LDA output is considered as fuzzy classification affecting the posts to the extracted topics. Moreover, the presentation module is the first attempt for expressing topics in such graph structure providing a meaningful way to interpret and spot main issues such COVID-19 related research.

The topics revolved around several axes, which began to feel the seriousness of the coming corona crisis, with a weakness in the number of publications regarding the COVID-19 virus during January and February, except for the English language as a result of its global use and Japanese due to the early spread of the disease in Japan and East Asian countries in general. The beginnings were characterized by focusing on the origin of the virus with the circulation of several terms related to the virus and everything related to the health sector in the concerned countries. In the second phase, which reaches the months of March and April, the Facebook users of along this target research discussed, in various degrees, everything related to mobilizing countries to confront the pandemic, while mentioning the procedures followed for protection that remain among the most important of the quarantine and the effects that accompanied it on many levels, including the family, social, economic and the politician. It

also touched on the death issue as a result of the high number of victims of this disease. People also showed interest in scientific research through following up on everything new about treatment, especially the discovery of vaccine. In a third stage that extends to the middle of May, we have noticed a gradual emergence of people's desire to lift the quarantine, even if only partially, and their desire to gradually return to normal life appeared with everyone's desire to follow health procedures from everyone in order to preserve the safety of the participants in the public space.

This work, as it depicted the reality experienced by the world in a unique experience with the COVID-19 pandemic, can be proactive in order to sense the changes that are in the process of occurring or that may occur. It is also worth noting that this system can be reproduced to work on any previous or later issue in order to study its temporal development in depth through the publications of social media platforms, especially Facebook.

In future works, we plan to broaden our work to cover other languages and to go deeper in the analysis by developing a recursive process able to zoom in the topics by extracting the sub-topics. Moreover, we will focus on the integration of semantic technologies to analyse the semantic dimension with embedding models in the measurement of the multilingual and cross-lingual semantic similarity/relatedness.

9. Acknowledgment

The work was supported by the Ministry of Higher Education and Scientific Research of Tunisia (MHESR) in the framework of Federated Research Project PRFCOV19-D1-P1.

References

Achrekar, H., Gandhe, A., Lazarus, R., Ssu-Hsin Yu, & Liu, B. (2011). Predicting flu trends using twitter data. In *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 702–707).

Ahmed, W. (2018). Using twitter data to provide qualitative insights into pandemics and epidemics.

Alqurashi, S., Alhindi, A., & Alanazi, E. (2020). Large arabic twitter dataset on covid-19. arXiv:2004.04315.

Alshaabi, T., Arnold, M. V., Minot, J. R., Adams, J. L., Dewhurst, D. R., Reagan, A. J., Muhamad, R., Danforth, C. M., & Dodds, P. S. (2020). How the world's collective attention is being paid to a pandemic: COVID-19 related 1-gram time series for 24 languages on Twitter.

Amara, A., Taieb, M. A. H., & Aouicha, M. B. (2017). Identifying i-bridge across online social networks. In *14th IEEE/ACS International Conference on Computer Systems and Applications, AICCSA 2017, Hammamet, Tunisia, October 30 - Nov. 3, 2017* (pp. 515–520).

Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., & Chowell, G. (2020). A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration, . This dataset will be updated bi-weekly at least with additional tweets, look at the github repo for these updates. Release: We have standardized the name of the resource to match our pre-print manuscript and to not have to update it every week.

Barkur, G., Prabhu, V., & Kamath, G. (2020). Sentiment analysis of nationwide lockdown due to covid 19 outbreak: Evidence from india. *Asian Journal of Psychiatry*, *51*, 102089.

Blair, S., Bi, Y., & Mulvenna, M. (2020). Aggregated topic models for increasing social media topic coherence. *Applied Intelligence*, *50*, 138–156. doi:https://doi.org/10.1007/s10489-019-01438-z.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*, 993–1022.

Boberg, S., Quandt, T., Schatto-Eckrodt, T., & Frischlich, L. (2020). Pandemic populism: Facebook pages of alternative news media and the corona crisis - a computational content analysis.

Chen, E., Lerman, K., & Ferrara, E. (2020a). Covid-19: The first public coronavirus twitter dataset. arXiv:2003.07372.

Chen, Q., Min, C., Zhang, W., Wang, G., Ma, X., & Evans, R. (2020b). Unpacking the black box: How to promote citizen engagement through government social media during the covid-19 crisis. *Computers in Human Behavior*, (p. 106380). doi:10.1016/j.chb.2020.106380.

Cherven, K. (2015). *Mastering Gephi Network Visualization*. Community experience distilled. Packt Publishing.

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. *CoRR*, *abs/2003.05004*.

Daga, R. R. M. (2017). Social network analysis of tweets on typhoon during haiyan and hagupit. In *Proceedings of the 8th International Conference on Computer Modeling and Simulation* ICCMS '17 (p. 151–154). New York, NY, USA: Association for Computing Machinery.

Ding, H., & Zhang, J. (2010). Social media and participatory risk communication during the h1n1 flu epidemic: A comparative study. *China Media Research*, *6*, 80–91.

Doan, S., Vo, B.-K. H., & Collier, N. (2012). An analysis of twitter messages in the 2011 tohoku earthquake. In P. Kostkova, M. Szomszor, & D. Fowler (Eds.), *Electronic Healthcare* (pp. 58–66). Berlin, Heidelberg: Springer Berlin Heidelberg.

Fan, B., Fan, W., Smith, C., & Garner, H. 2020). Adverse drug event detection and extraction from open data: A deep learning approach. *Information Processing Management*, *57*, 102131.

Ferrara, E. (2020). #covid-19 on twitter: Bots, conspiracies, and social media activism. *CoRR*, *abs/2004.09531*.

Gao, Z., Yada, S., Wakamiya, S., & Aramaki, E. (2020). NAIST COVID: multilingual COVID-19 twitter and weibo dataset. *CoRR*, *abs/2004.08145*.

Haouari, F., Hasanain, M., Suwaileh, R., & Elsayed, T. (2020). Arcov-19: The first arabic COVID-19 twitter dataset with propagation networks. *CoRR*, *abs/2004.05861*.

He, J., Li, L., Wang, Y., & Wu, X. (2020). Targeted aspects oriented topic modeling for short texts. *Applied Intelligence*, . doi:10.1007/s10489-020-01672-w.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent dirichlet allocation (lda) and topic modeling: Models, applications, a survey. *Multimedia Tools Appl.*, *78*, 15169–15211.

k joseph, J., Dev, K., A.P., P., & Mohan, M. (2018). Big data analytics and social media in disaster management. (p. 584).

Kankanamge, N., Yigitcanlar, T., Goonetilleke, A., & Kamruzzaman, M. (2020). Determining disaster severity through social media analysis: Testing the methodology with south east queensland flood tweets. *International Journal of Disaster Risk Reduction*, *42*.

Kleinberg, B., van der Vegt, I., & Mozes, M. (2020). Measuring emotions in the COVID-19 real world worry dataset. *CoRR*, *abs/2004.04225*.

Kuchler, T., Russel, D., & Stroebel, J. (2020). *The Geographic Spread of COVID-19 Correlates with Structure of Social Networks as Measured by Facebook*. Working Paper 26990 National Bureau of Economic Research.

Landwehr, P. M., & Carley, K. M. (2014). Social media in disaster relief. In W. W. Chu (Ed.), *Data Mining and Knowledge Discovery for Big Data:*

Methodologies, Challenge and Opportunities (pp. 225–257). Berlin, Heidelberg: Springer Berlin Heidelberg.

Lee, K., Agrawal, A., & Choudhary, A. (2017). Forecasting influenza levels using real-time social media streams. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 409–414).

Li, L., Zhang, Q., Wang, X., Zhang, J., Wang, T., Gao, T.-L., Duan, W., Tsoi, K., & Wang, F.-Y. (2020). Characterizing the propagation of situational information in social media during covid-19 epidemic: A case study on weibo. *IEEE Transactions on Computational Social Systems*, *PP*, 1–7.

Limaye, R., Sauer, M., Ali, J., Bernstein, J., Wahl, B., Barnhill, A., & Labrique, A. (2020). Building trust while influencing online covid-19 content in the social media world. *The Lancet Digital Health*, .

Lopez, C. E., Vasu, M., & Gallemore, C. (2020). Understanding the perception of COVID-19 policies by mining a multilanguage twitter dataset. *CoRR*, *abs/2003.10359*.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. Http://mallet.cs.umass.edu.

Miller, M., Banerjee, T., Muppalla, R., Romine, W., & Sheth, P. A. (2017). What are people tweeting about zika? an exploratory study concerning its symptoms, treatment, transmission, and prevention. *JMIR public health and surveillance*, .

Missier, P., McClean, C., Carlton, J., Cedrim, D., Silva, L., Garcia, A., Plastino, A., & Romanovsky, A. (2017). Recruiting from the network: Discovering twitter users who can help combat zika epidemics. In J. Cabot, R. De Virgilio, & R. Torlone (Eds.), *Web Engineering* (pp. 437–445). Cham: Springer International Publishing.

Missier, P., Romanovsky, A. B., Miu, T., Pal, A., Daniilakis, M., Garcia, A. F., Cedrim, D., & da Silva Sousa, L. (2016). Tracking dengue epidemics us-

ing twitter content classification and topic modelling. In *Current Trends in Web Engineering - ICWE 2016 International Workshops, DUI, TELERISE, SoWeMine, and Liquid Web, Lugano, Switzerland, June 6-9, 2016, Revised Selected Papers* (pp. 80–92).

Miyabe, M., Miura, A., & Aramaki, E. (2012). Use trend analysis of twitter after the great east japan earthquake. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion* CSCW '12 (p. 175–178). New York, NY, USA: Association for Computing Machinery.

PEARY, B., Shaw, R., & TAKEUCHI, Y. (2012). Utilization of social media in the east japan earthquake and tsunami and its effectiveness. *Journal of Natural Disaster Science*, *34*, 3–18.

Perrotta, D., Grow, A., Rampazzo, F., Cimentada, J., Del Fava, E., Gil-Clavel, S., & Zagheni, E. (2020). Behaviors and attitudes in response to the covid-19 pandemic: Insights from a cross-national facebook survey. *medRxiv*, .

Pizzuti, A. G., Patel, K. H., McCreary, E. K., Heil, E., Bland, C. M., Chinaeke, E., Love, B. L., & Bookstaver, P. B. (2020). Healthcare practitioners' views of social media as an educational resource. *PLOS ONE*, *15*, 1–16.

Pruss, D., Fujinuma, Y., Daughton, A., Paul, M., Arnot, B., Szafir, D., & Boyd-Graber, J. (2019). Zika discourse in the americas: A multilingual topic analysis of Twitter. *PlosOne*, .

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010a). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (pp. 851–860). ACM.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010b). Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web* WWW '10 (p. 851–860). New York, NY, USA: Association for Computing Machinery.

Sebei, H., Taieb, M. A. H., & Aouicha, M. B. (2018). Review of social media analytics process and big data pipeline. *Social Netw. Analys. Mining*, *8*, 30:1–30:28.

Sharma, M., Yadav, K., Yadav, N., & Ferdinand, K. C. (2017). Zika virus pandemic—analysis of facebook as a social media health information platform. *American Journal of Infection Control*, *45*, 301 – 302.

Sicilia, R., Giudice], S. L., Pei, Y., Pechenizkiy, M., & Soda, P. (2018). Twitter rumour detection in the health domain. *Expert Systems with Applications*, *110*, 33 – 40.

Singh, L., Bansal, S., Bode, L., Budak, C., Chi, G., Kawintiranon, K., Padden, C., Vanarsdall, R., Vraga, E. K., & Wang, Y. (2020). A first look at COVID-19 information and misinformation sharing on twitter. *CoRR*, *abs/2003.13907*.

Teodorescu, H.-N. (2015). Using analytics and social media for monitoring and mitigation of social disasters. *Procedia Engineering*, *107*, 325–334.

Tran, T., & Lee, K. (2016). Understanding citizen reactions and ebola-related information propagation on social media. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 106–111).

Ulvi, O., Lippincott, N., Khan, M. H., Mehal, P., Bass, M., Lambert, K., Lentz, E., & Haque, U. (2019). The role of social and mainstream media during storms. *Journal of Public Health and Emergency*, *3*.

Zarei, K., Farahbakhsh, R., Crespi, N., & Tyson, G. (2020). A first instagram dataset on covid-19.

Zarrad, A., Jaloud, A., & Alsmadi, I. (2014). The evaluation of the public opinion - a case study: Mers-cov infection virus in ksa. In *2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing* (pp. 664–670).

Zhang, J., Chen, Y., Zhao, Y., Wolfram, D., & Ma, F. (2020). Public health and social media: A study of Zika virus-related posts on Yahoo! Answers. *Journal of the Association for Information Science & Technology*, *71*, 282–299.

Zhou, C., Su, F., Pei, T., Zhang, A., du, Y., Luo, B., Cao, Z., Wang, J., Yuan, W., Zhu, Y., Song, C., Chen, J., Xu, J., Li, F., Ma, T., Jiang, L., Yan, F., Yi, J., Hu, Y., & Xiao, H. (2020). Covid-19: Challenges to gis with big data. *Geography and Sustainability*, .

Figures

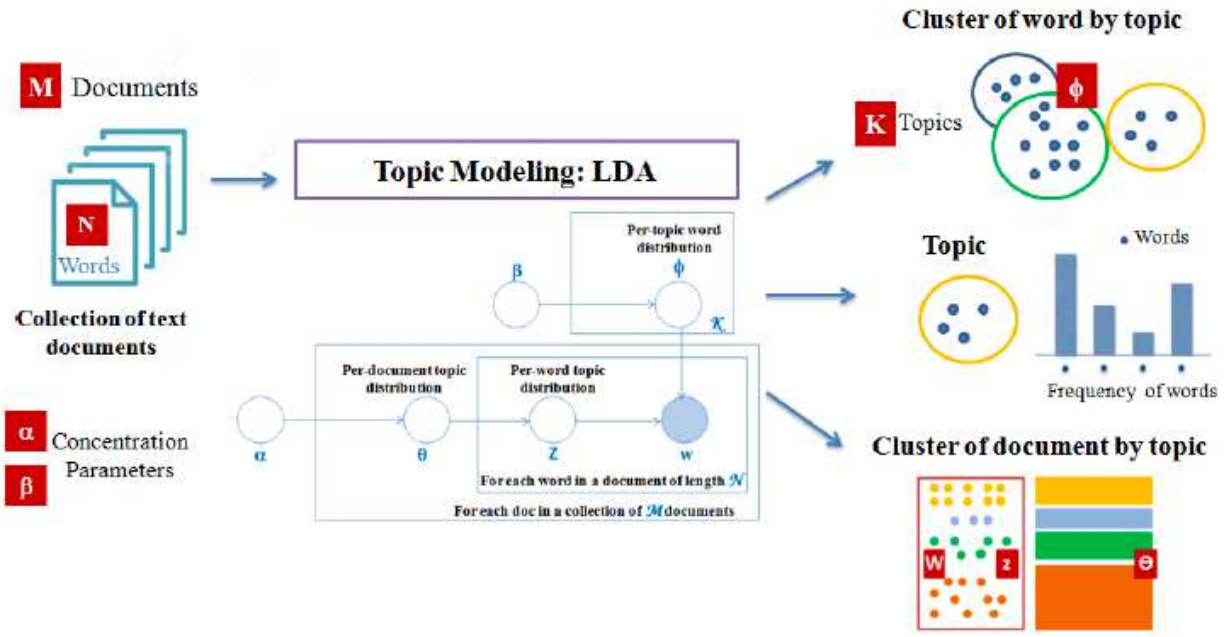


Figure 1

LDA topic modelling process

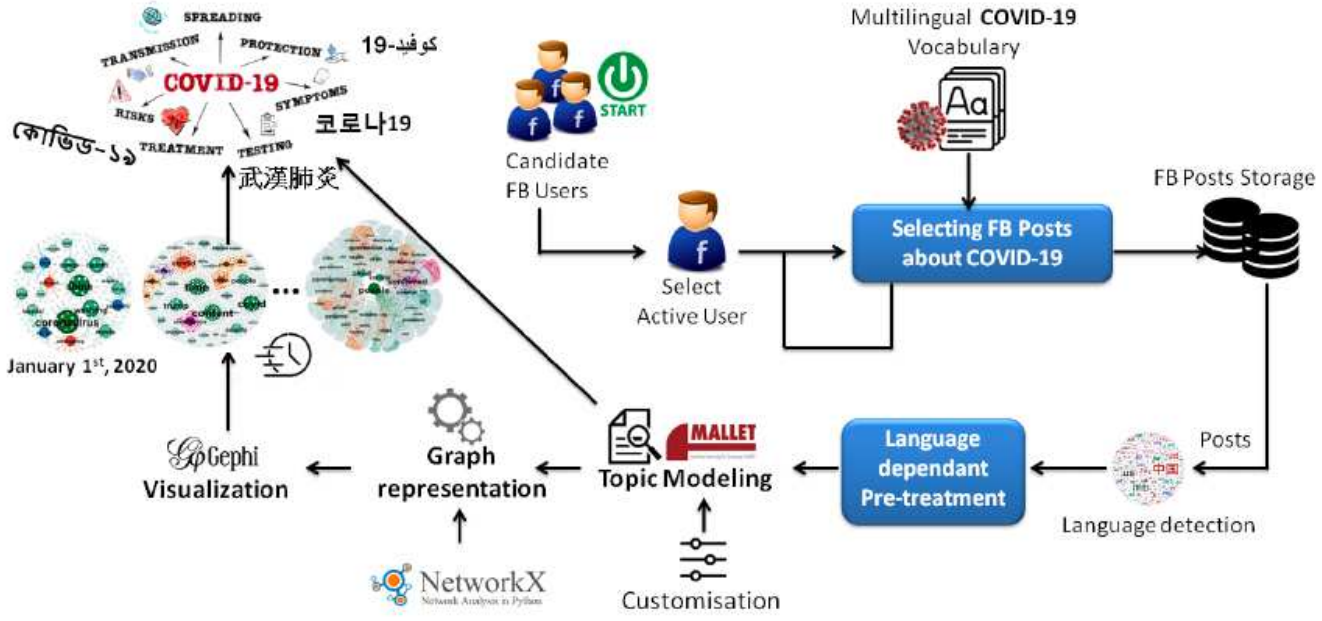


Figure 2

Facebook-based COVID-19 tracking trends evolution system

Figure 3

Distribution of the COVID-19 related data through countries around the world. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

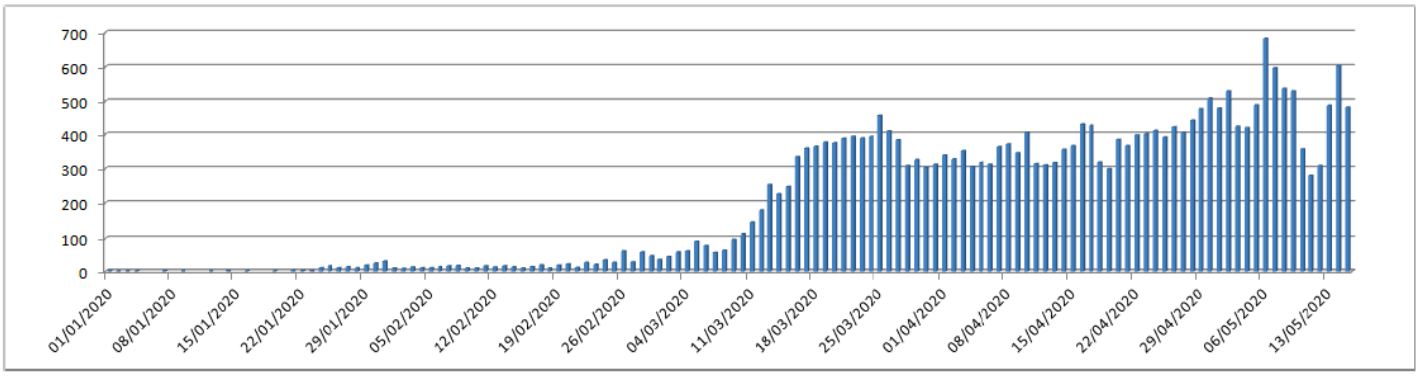


Figure 4

Distribution of the gathered Facebook public posts through the time since January 1st, 2020 to May 15, 2020

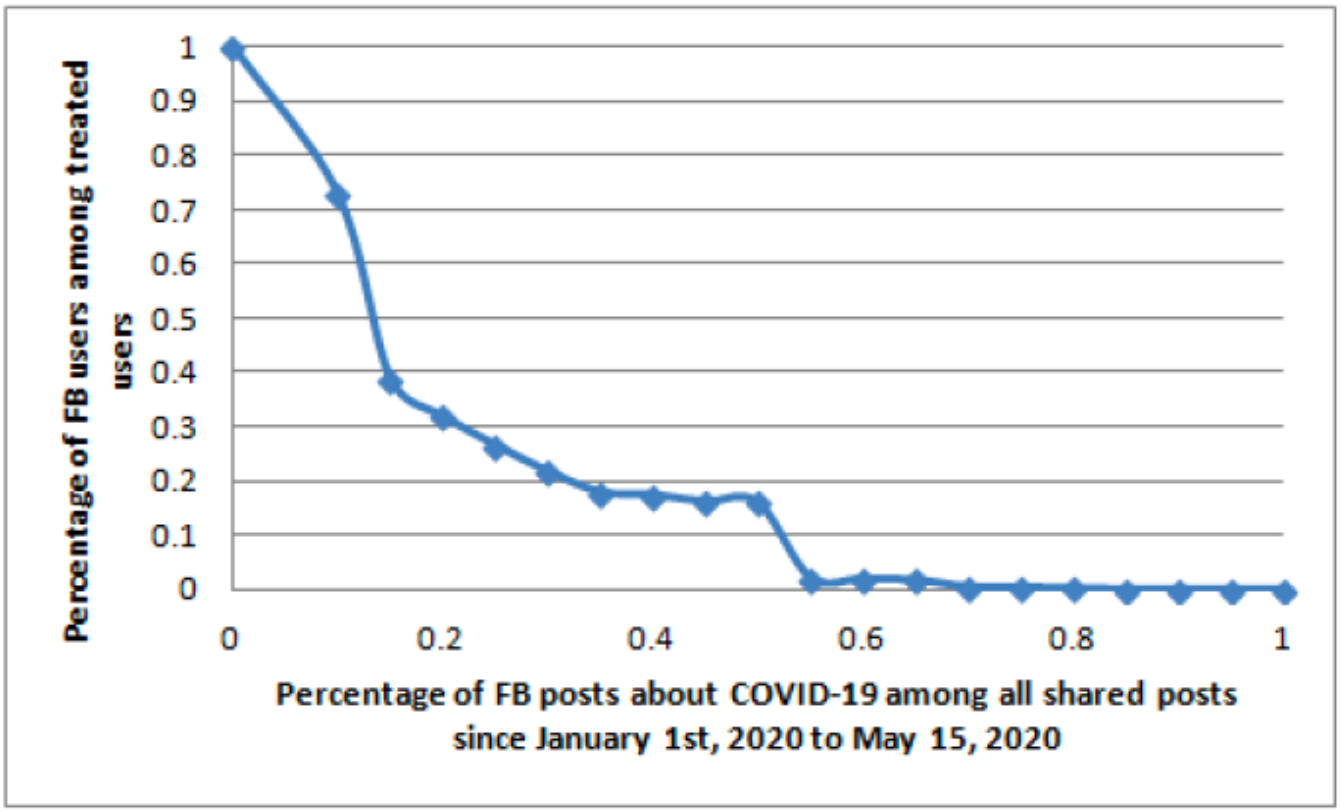


Figure 5

Correlation between percentage of COVID-19 Facebook posts during the period from January 1st, 2020 to May 15, 2020, and the percentage of users having post more than the fixed percentage.

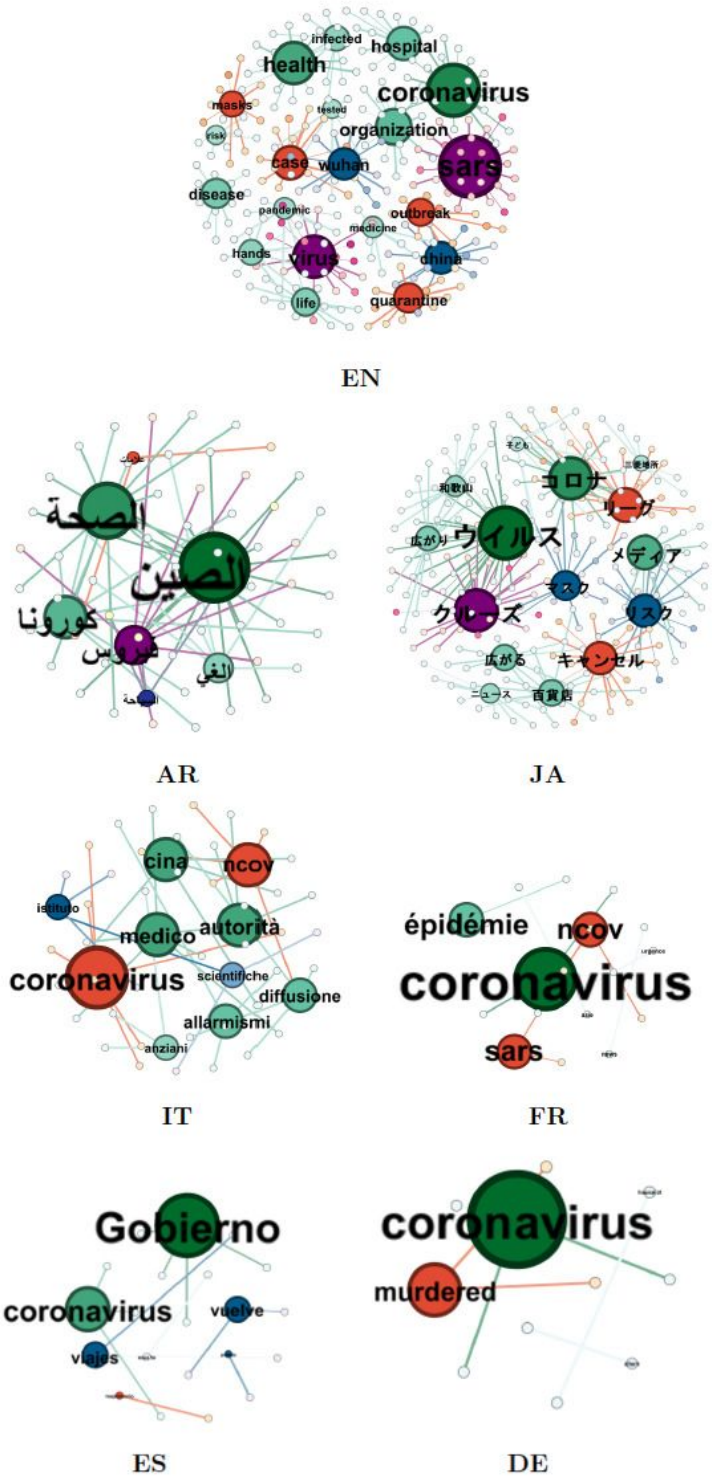


Figure 6

Facebook-based crowd-sourced COVID-19 trends covering 7 languages (EN, AR, FR, ES, IT, DE and JA) for the period January and February 2020.

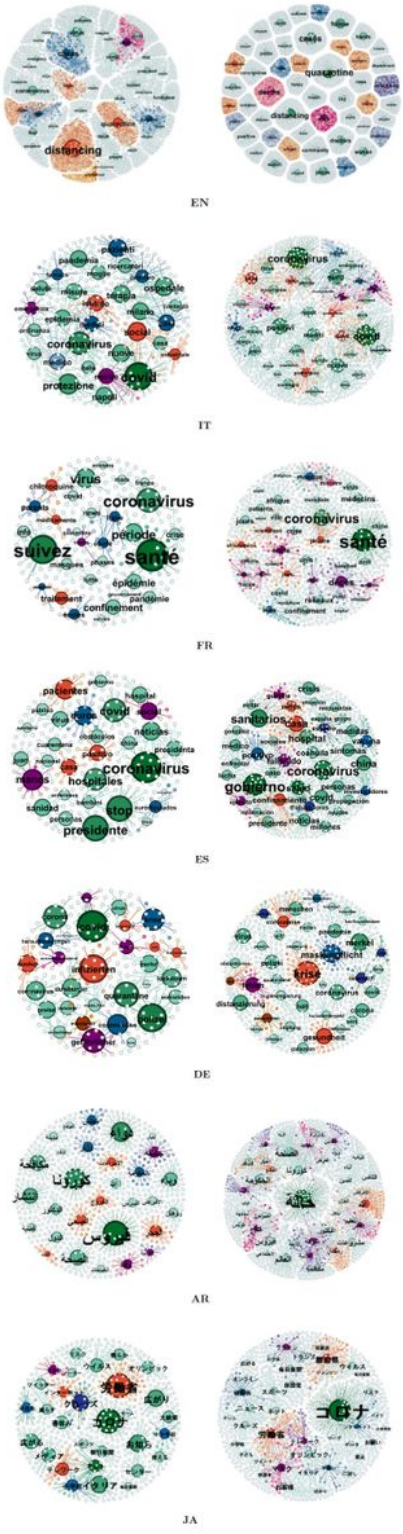


Figure 7

Facebook-based crowd-sourced about COVID-19 trends covering 7 languages (EN, AR, FR, ES, IT, DE and JA) for the period March (left part) and April 2020 (right part).

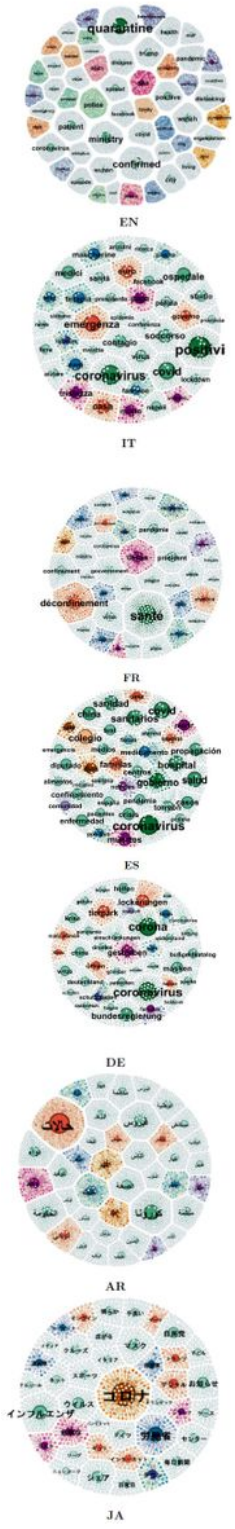


Figure 8

Facebook-based crowd-sourced trends about COVID-19 covering 7 languages (EN, AR, FR, ES, IT, DE and JA) for the period until May 15, 2020.