The covid-19 crisis: an NLP exploration of the french Twitter feed (February-May 2020)

Sophie Balech, C. Benavent, M. Calciu, Julien Monnot

The covid-19 crisis: an NLP exploration of the french Twitter feed (February-May 2020)

Sophie Balech[1][0000−0002−7363−1271], Christophe Benavent[2][0000−0002−7253−5747], Mihai Calciu[3][0000−0002−8375−4136], and Julien Monnot[2]

[1] Picardie Jules Verne University, 10 Venelle Lafleur, 80000 Amiens, France
sophie.balech@gmail.com
[2] Nanterre University, 200 Avenue de la République, 92000 Nanterre, France
christophe.benavent@gmail.com
[3] Lille University, 104 avenue du Peuple Belge, 59000 Lille, France

Abstract. The Covid-19 pandemic offers a spectacular case of disaster management. In this literature, the paradigm of participation is fundamental: the mitigation of the impact of the disaster, the quality of the preparation and the resilience of the society, which facilitate the reconstruction, depend on the participation of the populations. Being able to observe and measure the state of mental health of the population (anxiety, confidence, expectations, ...) and to identify the points of controversy and the content of the discourse, are necessary to support measures designed to encourage this participation. Social media, and in particular Twitter, offer valuable resources for researching this discourse.

The objective of this empirical study is to reconstruct a micro history of users' reactions to the pandemic as they share them on social networks. The general method used comes from new processing techniques derived from Natural Language Processing (NLP). Three analysis methods are used to process the corpus: analysis of the temporal evolution of term occurrences; creation of dynamic semantic maps to identify co-occurrences; analysis of topics using the SVM method.

The main empirical result is that the mask emerges as a central figure of discourse, at least in the discourse produced by certain social media. The retrospective analysis of the phenomenon allows us to explain what made the mask a focal point not only in conversation, but also in behaviors. Its value resides less in its functional qualities than in its ability to fix attention and organize living conditions under the threat of pandemic.

Keywords: Covid-19 · Twitter feed · NLP methods.

1 Introduction

The Covid-19 pandemic that hit the planet offers a spectacular case of disaster management [12]. In this literature, the paradigm of participation is fundamental [7, 16] : the mitigation of the impact of the disaster, the quality of the preparation and the resilience of the society, which facilitate the reconstruction, depend on the participation of the populations. Being able to observe and measure the state of mental health of the population (anxiety, confidence, expectations, ...) and to identify the points of controversy and the content of the discourse, are necessary to support measures designed to encourage this participation. Social media, and in particular Twitter, offer valuable resources for researching this discourse.

In the literature on disaster management, three concepts are key and necessary at all stages of a disaster: the first is the state of preparedness for the consequences of the phenomenon and its aftershocks, the second concerns the mitigation of these consequences, which requires the participation of populations, and the third relates to the capacity of each individual to bounce back and embark on the road to reconstruction: resilience.

The modern conception of disaster management, with its emphasis on the participation of populations, questions the factors that encourage or curb it. Material, cognitive and organizational resources are obvious, their mobilisation and preparation are decisive, but in the end, it is undoubtedly the mental health of populations and their level of commitment that make the difference.

The objective of this empirical study is to reconstruct a micro history of users' reactions to the pandemic as they share them on social networks. Despite the volume of data and the technique used, the methodological approach is descriptive and aims to establish a fact: the central and increasing role of the mask in social conversation as it emerges from observation. The general method used comes from a new paradigm [4] which build itself between abundant data (web, social networks, ...) and new processing techniques derived from Natural Language Processing (NLP).

2 Dataset

The dataset is a corpus of Twitter content developed by [2] based on a set of keywords around covid, corona and associated words. The original corpus with nearly 110 million tweets, collected between the end of January and end of Mai 2020 and available in dehydrated form due to Twitter's terms of service was used in order to extract the French tweets. From the resulting 2.156 million posts, whose production over time is shown in figure 1, 565,662 have been retained as having contributory content that is original, reply and quote tweets. Retweets have been excluded in order to to avoid redundancy.

This corpus is pre-processed by removing account mentions and URLs by the appropriate regular expressions, putting the whole text in lower case, tokenizing, lemmatizing and annotating the text with the "part of speech" with the Udpipe annotator of CleanNLP [1] as well as identifying syntax dependencies. The whole corpus represents 10 million tokens whose only common names were filtered out in order to analyze the topics discussed in the message flow.

The heterogeneity of the content is fully shown in the distribution of the number of tweets per account (fig. 2). A very small number of accounts produces a large part of all tweets, even if the number of accounts is high: 202,000 distinct

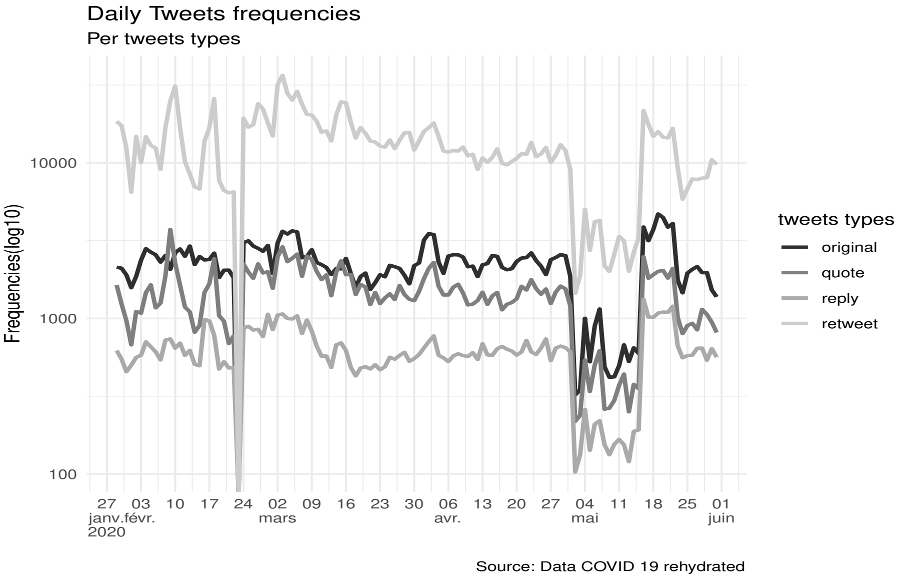

Fig. 1. Number of posts produced per day by type of post (n=2.2 m)

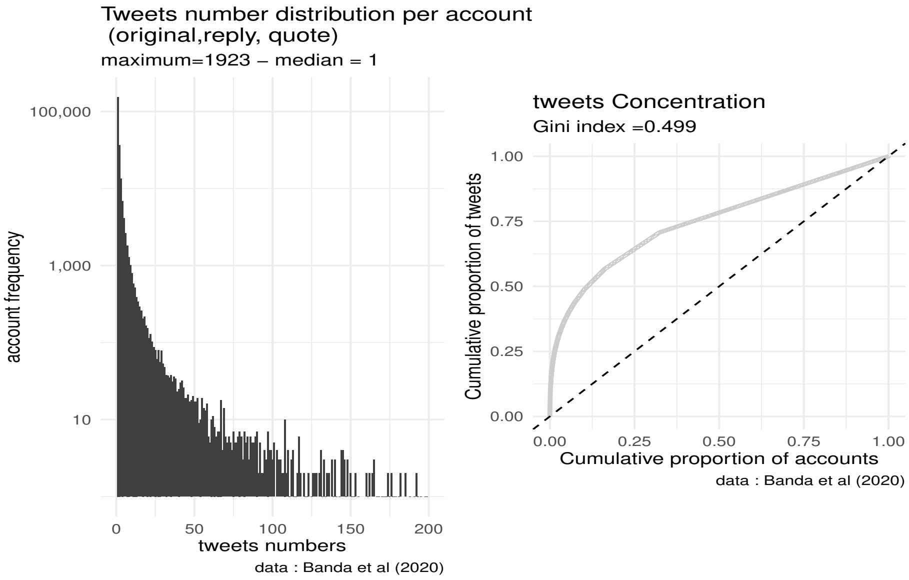

Fig. 2. Tweets distribution per account

contributors. The degree of concentration is high with a Gini index of 0.50. It indicates a huge inequality of contribution: 25% of the 565,662 primary production tweets were created by a minority of about 1000 contributors from a total of 202,000 accounts. We find the major actors: media, politicians, but also activists, journalists, columnists. The extreme heterogeneity of the population and its production is questionable, as is its small size (the number of daily Twitter users is around 4.6 million in France, less than half of whom are active). Although the database claims to be somewhat exhaustive, a certain fragility emerges. On such a scale, small groups of activists can fairly easily find an audience, but at least we have a good reflection of the media arena.

3 Method

We use the tidyverse [15] collection of packages of the R statistical environment [9] to analyze data. Three types of methods are used in a complementary way to highlight the different aspects of chronological variations of the discourse themes.

3.1 Daily evolution of the study's focal categories

The methods used aim first to capture daily changes in the use of the most frequent vocabulary. A certain number of target terms have been identified, those relating to the epidemic (corona, covid), lockdown and deconfinement, and those naturally linked to barrier gestures (mask, gel, teleworking, ...). To test the consistency of the daily changes, the frequency with which countries are cited is also measured, which gives a coherent picture of the trends. In order to deal with morphological variations in terms and to take into account spelling errors, rather than annotation by lemmas, we use regular expressions. Thus, for the term "hôpital" (hospital), the pattern ".*h([o,ô]—[os,ôs])pital.*" is used, which makes it possible to identify more than thirty variants.

3.2 Dynamic semantic maps

Simple semantic maps are used to explore the evolution of discourse on a weekly scale. These maps are obtained by calculating for each of the 22 weeks the co-occurrences between the most frequent words. We are therefore interested here in the joint distribution of terms in the same documents, their proximity resulting from their use in the same texts.

These co-occurrence tables are binary recoded according to the following dualism: existence of relation / absence of relation, according to a predefined threshold, then represented in a small space using Fruchterman and Reingold's algorithm [6]. The implementation of the procedure is carried out with igraph [5]. The elements obtained by the analysis of semantic networks are then the subject of centrality calculations. We retain three centrality indicators:

- the degree of centrality: the number of links each node has;

- the centrality of betweenness: the number of times a node intervenes in a shorter path connecting two other nodes;
- closeness: the sum of the length of the shortest paths connecting a node to the others.

3.3 Analysis of structural themes (STM)

Topic analysis is already well known [3], it aims to identify a number of k topics in a corpus by assuming that each document has a probability of corresponding to one or the other of the k topics (n x k matrix of the theta parameters), and that each term has a certain probability to belong to a given topic (m x k matrix of the beta parameters).

Here we use a method that differs from the initial model, and whose characteristics are, on the one hand, to take into account the longitudinal nature of the data by measuring the prevalence over time of each of the topics (and more generally other co-variables that act as independent variables), and, on the other hand, to assume that the topics are correlated, which is reasonable in this situation. This is the Structural Topic Model proposed by [11]. We use its implementation in the stm package [10].

4 Results

4.1 The mask in first place for attenuation tactics

The first approach to content analysis is simply to count the terms frequency and focus on the most frequent ones. This is what we have represented in figure 3. The different forms of corona and covid dominate the ranking. The mask stands in a good position, it is certainly the first evocation in frequency of a means of defence against the epidemic, of a concrete object. It dominates the other means of attenuation. This raw frequency deserves to be described more qualitatively: do we say the same things about the mask at different phases of the epidemic episode? To this end, syntactic dependency annotations are used to identify which terms are grammatically associated with the common noun "mask": adverbs, adjectives, other nouns (some universal Stanford dependencies are used: acl, amod, nmod and appos). We compare the most frequent ones among the 4 months of observations. For each of the terms obtained their density is calculated and the spectrum of meanings is obtained, represented in figure 4. The results are clear: 1) the mask protects it' s trivial 2) its shape is surgical rather than ffp2 despite hesitation 3) its distribution and commercialization become more important with time. But the main thing is that it is associated with the " mandatory " aspect. The mask used to be a means, it is now as much a legal norm as a social norm.

4.2 From Coronavirus to Covid-19

From the frequencies of occurrence of a series of terms representative of the debates and topics of interest, a daily density is calculated, with a smoothing

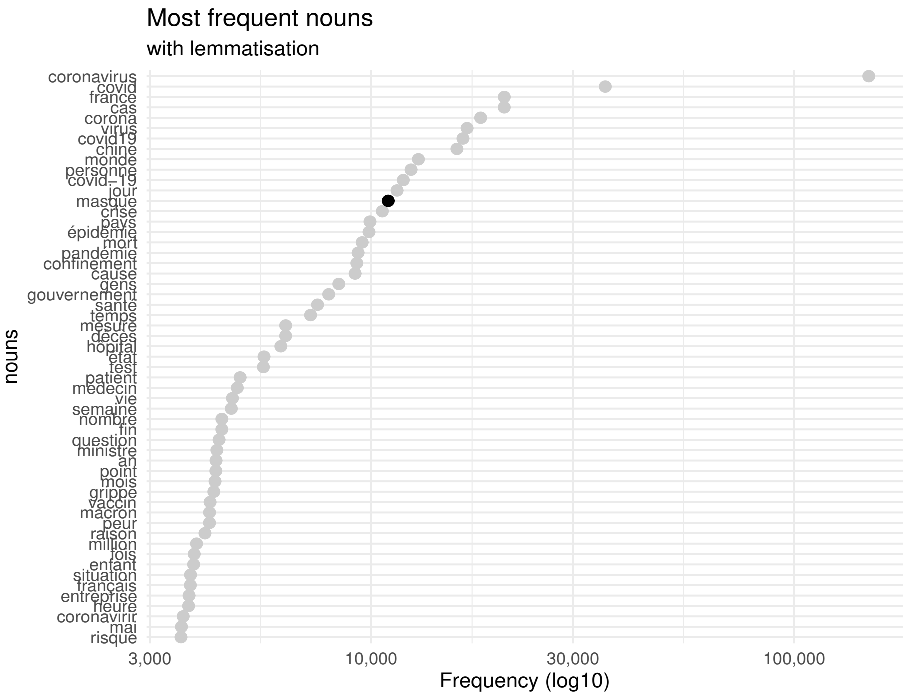

Fig. 3. The mask among the most frequently used words

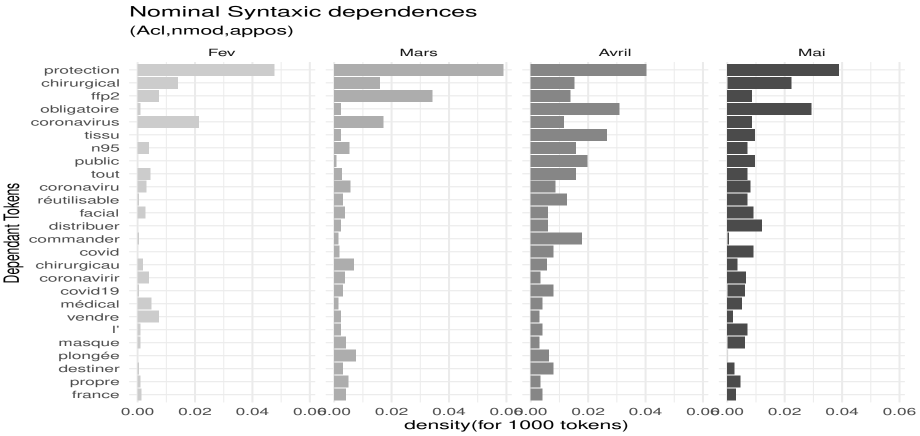

Fig. 4. The most frequent nominal mask dependencies by month (in terms of density $= f_{im}/f_{.m}$ with i = term, m=month). (Universal Dependencies: acl, amod, nmod and appos)

over 7 days. The temporal evolution of the frequency of these terms is shown in figures 5 and 6.

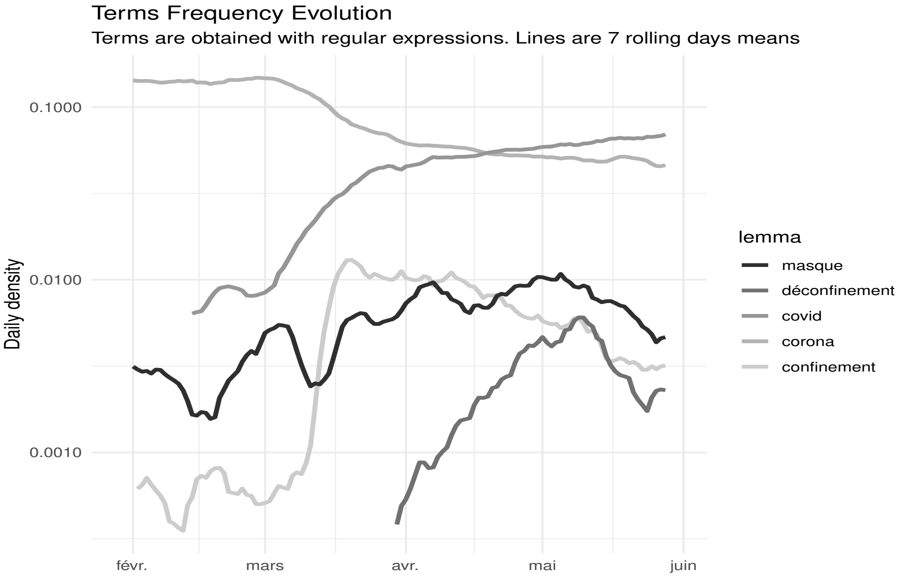

Fig. 5. Changes in the frequency of key terms: corona, covid and lockdown/mask.

The evolution is cristal clear. If from January to February the coronavirus was the star (very dark), the Covid-19 takes over during March and remains in the lead for the rest of the period. In other words, it is the consequences of the epidemic spread of the coronavirus that dominate the debates. The disease weighs directly by the pressure put on public health and by its victims, but also by the measures it provokes to mitigate its impact. In this sense, Covid-19 becomes the social and political incarnation of its biological counterpart the Coronavirus. The change of term marks a change in discourse: the threat that was external is rapidly endogenized, it becomes less the virus than the perturbations it generates (excess mortality, disruption of the health system, redefinition of social relations, economic shock). The moment of confinement is a shift in perspective from outside to inside.

As for the mask, we observe its progressive rise, in successive waves that undoubtedly concern its various polemics, and its quantitative domination over other methods of attenuation: test, hydroalcoholic gel, telework, to mention the most significant ones. The impact of the disaster coming from elsewhere is accompanied by a change of perspective, and a reversal where the mask becomes an increasingly frequent concern.

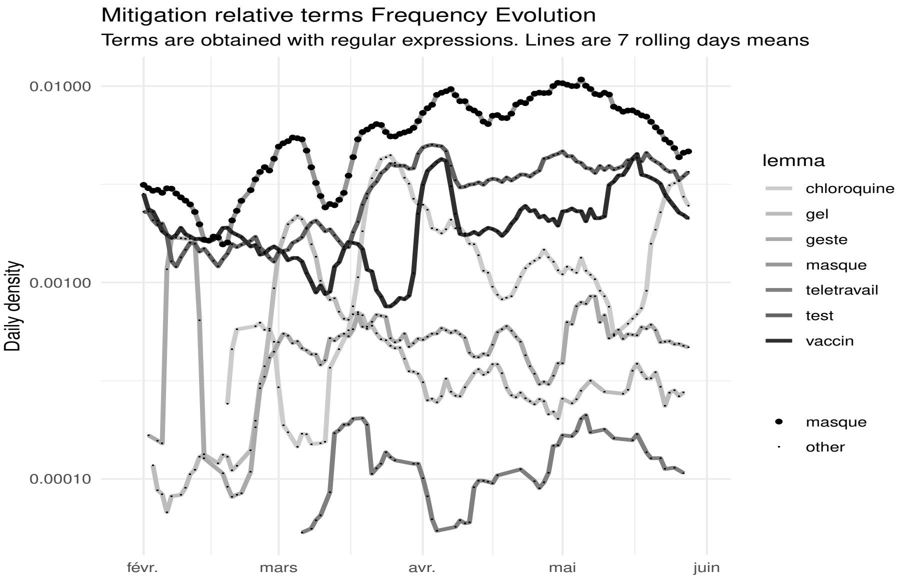

Fig. 6. Means of barrier gestures

4.3 The mask at the center of the debate from the beginning

To report on the conversations and their evolution, a method of semantic maps is applied for each period (22 consecutive weeks). Reading the map is fairly simple: words whose co-occurrences are greater than a frequency determined according to each period's corpus are represented, satisfying a constant proportion across the periods (about 30%). The size of the nodes corresponds to their density in the corpus, that of the arcs to the frequency of the co-occurrence. The relative positions in the plane are calculated by a multidimensional positioning method according to the similarity of the terms. As can be seen in Figure 7 there is a central theme represented by the macro-component appearing as a network of terms surrounded by more specific and disconnected themes positioned at the periphery.

It illustrates the main discussions during the second week of (first) lockdown. During this period, the peripheral conversations dealt with a variety of topics: chloroquine treatment, the health system, the practical issue of travel certification, and a petition addressed to the president for the use of chloroquine. The macro-component is fairly clearly structured. On an almost horizontal axis, lockdown is at the center of a temporal concern: for how long? At the other end, the crisis of the shortage of masks, in particular, for caregivers, is clearly taking shape. At the center, France and the government are bridging the issues. In the north of the component, we find the question of the state of emergency, in the south the factual theme of the scale of the crisis translated into the daily number of deaths.

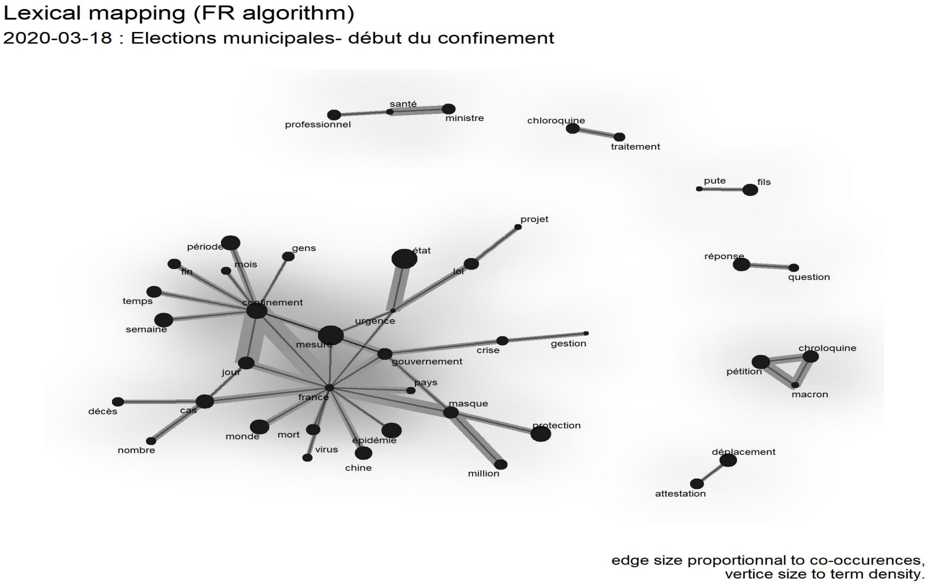

Fig. 7. Example of a semantic map (week 12 (16-22 March 2020) beginning of lockdown and municipal election)

By repeating this interpretative analysis over the 22 weeks of data, and thus 22 maps, we can schematically reconstruct the evolution of the discourse.

To validate the hypothesized evolution of the mask object in the social media discourse, we suggest a centrality test (see fig. 8). One would naturally be curious about the future, but over the observed period there is an undeniable increasing centrality (even if the last few weeks mark a weakening), the mask is connected to an increasing number of conversations. Becoming the lowest common denominator, it becomes the main key through which one can access the different paths of collective thinking. Naturally, the higher citation frequency of the mask compared to other means of attenuation, favours increased variety of associated objects. This is the statistical point of view. We can also consider the hypothesis that being associated with more themes, it is cited more often.

4.4 Confirmation by topic modeling

The last approach to appreciate the importance of the mask is to apply a topic model on temporal data (STM model). This type of model integrates variables that explain the prevalence of topics. Time, translated here as the order of weeks, is assumed to be associated with the prevalence of certain topics; it represents a proxy for the states of the general environment that affect the content of conversations. The model also allows topics to be correlated with each other.

A satisfactory solution seems to support 20 topics, of which it is difficult to give an exhaustive analysis here, but more easily a synthetic presentation based

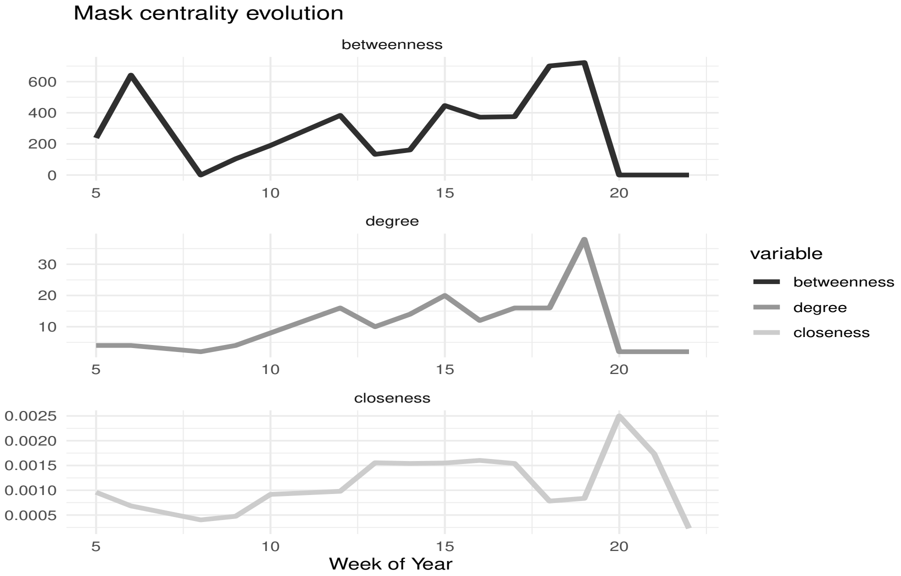

Fig. 8. Evolution of the lexical centrality of the mask (criteria of degree, betweenness and closeness)

on the most visual representation provided by the model. Figure 9 provides the synthesis. The topics are closer the more they are correlated, the correlations (over 0.15) are represented by the thickness of the segments, and the size of the circles is proportional to the frequency of the topics.

This structure is characterized first of all by a kind of duality, a macro segment seems to be articulated around two components. One is centered on the coronavirus and China, it will be characterized as exotic; the other on the covid and the public issue. We find this idea of a change of perspective. What used to be a foreign body becomes an inner pain.

The advantage of the model is that it makes it possible to represent temporal prevalence, which is shown in figure 10. We can clearly identify the topics that are favoured in the first period and then decline, as well as those that rise in the second period.

The beginning of the study period is characterized by the prevalence of questions related to the discovery of the coronavirus, its effects visible through the number of deaths in China, and the situation in France facing an unprecedented virus. Then new themes emerge, related to Covid-19: the development of the pandemic, the case of children, announcements concerning future vaccines and potential treatments (with chloroquine at the heart, of course), to end with the state of crisis in the country and its management by the political power. The general trend is that the discussions are moving towards an endogenization of the epidemic, which starts with an unknown Wuhan virus and turns into a dis-

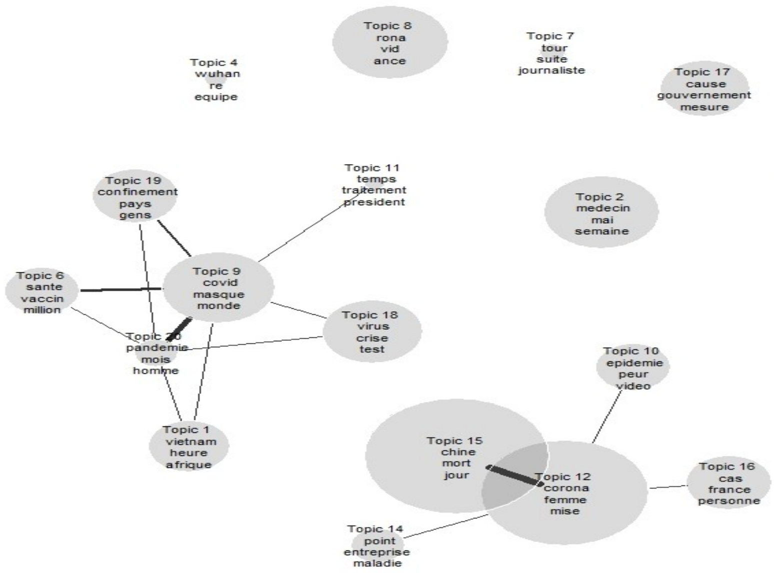

Fig. 9. Correlation network of the 20 identified topics: (r threshold: 0.15 - FR projection))

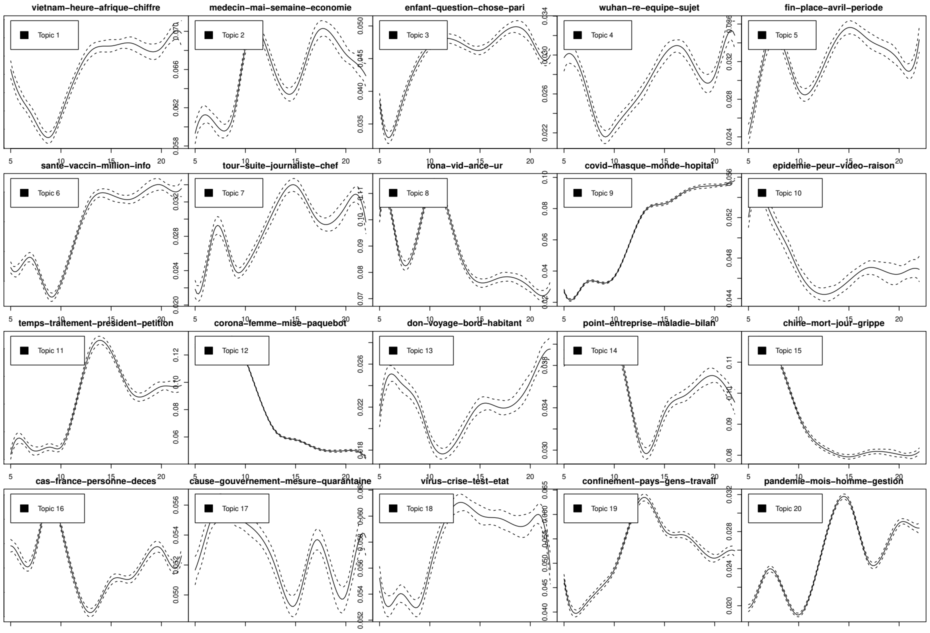

Fig. 10. Time prevalence graphs on the probability of topics (described by the three most distinctive terms)

ease that devastates France and the French (both economically and medically).

As for the mask, a schema is emerging that clarifies its position. It is not found, as one might have expected, in a large number of themes, but as a central theme closely associated with the disease. This topic is itself associated with a few others, and this configuration defines the new paradigm of discourse which is tied up in the first lockdown.

5 Conclusion

Empirically, the main result, obtained by triangulation of methods, is that the mask emerges as a central figure in Twitter discourse as early as April, with the endogenization of the epidemic. This is the term that seems to articulate the polemics (lack of protection for health workers, supply problems, dissemination of the obligation to wear it). Over the period under study, that of the first lockdown, the obligation is limited, and compliance with its wearing is largely voluntary insofar as voluntary behaviour includes the effects of social standardization.

This result was not obvious at first glance, other figures were candidates: videoconferencing, in which millions of workers, students or professors quickly learned to manipulate interfaces with variable happiness but which does not appear in the contents. The hydroalcoholic gel too, whose distributors spread out over shops, stations and administrations and which pharmacists could not manufacture in time. Tests in particular, whose usefulness was questioned before becoming a central argument for government action and no doubt a factor of success in certain Asian countries. It is finally towards the mask that the discourses of the twittosphere converge.

A posteriori analysis of the phenomenon makes it possible to explain what has made the mask a focal point of conversation, but also of behavior. It is worth less for its functional qualities (continuously discussed even from a scientific point of view) than for its ability to fix attention and to organize living conditions under threat of epidemic. The defended hypothesis is that its phenomenological ambivalence nourishes sufficient interpretative flexibility to polarize the questioning of mismatched social worlds (scientists, politicians and citizens), and provides a common framework within which conventions and norms can be renegotiated. Thus, we can consider the mask as a boundary object [14, 13]. Beyond this role of border-crosser, it paradoxically forms the fabric of social relations, the material, ritual and symbolic knots through which social activity is reorganized.

On a practical level, this study underlines the need to consider the means of attenuation (barrier gestures, social distance, use of artefacts ...) not only in terms of their intrinsic and extrinsic effectiveness, respect for compliance and the obligation that provokes resistance, but also from a more anthropological perspective that gives the material objects of (extra) ordinary life a power nourished by their capacity to ritualize social interactions, to symbolically carry the commitment of the actors and even more to give them power over the invisible curse of the epidemic [8]. Even if it does not provide much protection, the mask is as effective as the fetish, it maintains social order when we know nothing about the battles that take place out of our sight. It may well be that getting people to participate requires more than just rational arguments.

Despite the quantitative dimension of our treatments (counting frequencies, densities, probabilities), the analysis conducted in this study is largely exploratory at least in the thematic sense. The ability to quantify the frequency of themes, however, allows us to reconstruct an immediate micro history in a factual manner and to highlight a key phenomenon: the mask is the central figure in twitter discussions. The methodological contribution is straight forward: developing formal methods to deal with massive data contents. Beyond the technical aspects, the general idea is that of building processing procedures that allow the researcher to deal with large volumes of data and to have an objectified representation of the discourse.

The methods of analysis presented in this research are particularly appealing to social network managers: they help identify topics of importance and represent the dynamics of discourse. Applications can be found in the management of online reputation, in the identification of irritants and the responses to solutions implemented to deal with them. Detailed code is available upon request.

The limits of this work are numerous and result from the rush of this research which began exactly a few days after the confinement (March 16, 2020) and ends, at least in the form of this contribution, at the end of October 2020, with the empirical material covering the period from February to May 2020. The main limitation lies in the weak explicitation of the social processes that are invoked without having observed them. We observe only the discursive consequences without describing them precisely, on an object of study limited to social media whose profoundly unequal textual production structure gives the most engaged a disproportionate voice.

The research avenues opened up by this work are numerous, but we can cite two main ones: continuing the study (data being available) over the remaining period to December 2020, and extending it to other countries in order to see wether the same phenomena converge and if the role of the mask is maintained depending on the context.

References

1. Arnold, T.: A Tidy Data Model for Natural Language Processing using cleanNLP. The R Journal **9**(2), 248 (2017). https://doi.org/10.32614/RJ-2017-035, https://journal.r-project.org/archive/2017/RJ-2017-035/index.html
2. Banda, J.M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Chowell, G.: A large-scale COVID-19 Twitter chatter dataset for open scientific research – an international collaboration. arXiv:2004.03688 [cs] (Apr 2020), http://arxiv.org/abs/2004.03688, arXiv: 2004.03688
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. J. Mach. Learn. Res. **3**, 993–1022 (Mar 2003), http://dl.acm.org/citation.cfm?id=944919.944937
4. Cambria, E., White, B.: Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. IEEE Computational Intelligence

Magazine **9**(2), 48–57 (May 2014). https://doi.org/10.1109/MCI.2014.2307227, http://ieeexplore.ieee.org/document/6786458/

5. Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJournal **Complex Systems**, 1695 (2006), https://igraph.org
6. Fruchterman, T.M.J., Reingold, E.M.: Graph drawing by force-directed placement. Software: Practice and Experience **21**(11), 1129–1164 (Nov 1991). https://doi.org/10.1002/spe.4380211102, http://doi.wiley.com/10.1002/spe.4380211102
7. Horney, J., Nguyen, M., Salvesen, D., Tomasco, O., Berke, P.: Engaging the public in planning for disaster recovery. International Journal of Disaster Risk Reduction **17**, 33–37 (Aug 2016). https://doi.org/10.1016/j.ijdrr.2016.03.011, https://linkinghub.elsevier.com/retrieve/pii/S2212420915301680
8. Lemonnier, P.: Mundane objects: materiality and non-verbal communication. Critical cultural heritage series, Left Coast Press, Walnut Creek, CA (2012)
9. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2019), https://www.R-project.org/
10. Roberts, M.E., Stewart, B.M., Tingley, D.: stm: An R package for structural topic models. Journal of Statistical Software **91**(2), 1–40 (2019). https://doi.org/10.18637/jss.v091.i02
11. Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., Rand, D.G.: Structural Topic Models for Open-Ended Survey Responses: STRUCTURAL TOPIC MODELS FOR SURVEY RESPONSES. American Journal of Political Science **58**(4), 1064–1082 (Oct 2014). https://doi.org/10.1111/ajps.12103, http://doi.wiley.com/10.1111/ajps.12103
12. Rodriguez, H., Quarantelli, E.L., Dynes, R.R. (eds.): Handbook of disaster research. Handbooks of sociology and social research, Springer, New York (2007)
13. Star, S.L., Griesemer, J.R.: Institutional Ecology, ‘Translations’ and Boundary Objects: Amateurs and Professionals in Berkeley’s Museum of Vertebrate Zoology, 1907-39. Social Studies of Science **19**(3), 387–420 (Aug 1989). https://doi.org/10.1177/030631289019003001, http://journals.sagepub.com/doi/10.1177/030631289019003001
14. Trompette, P., Vinck, D.: Retour sur la notion d’objet-frontière. Revue d’anthropologie des connaissances **3, 1**(1), 5 (2009). https://doi.org/10.3917/rac.006.0005, http://www.cairn.info/revue-anthropologie-des-connaissances-2009-1-page-5.htm
15. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H.: Welcome to the tidyverse. Journal of Open Source Software **4**(43), 1686 (2019). https://doi.org/10.21105/joss.01686
16. Witvorapong, N., Muttarak, R., Pothisiri, W.: Social Participation and Disaster Risk Reduction Behaviors in Tsunami Prone Areas. PLOS ONE **10**(7), e0130862 (Jul 2015). https://doi.org/10.1371/journal.pone.0130862, https://dx.plos.org/10.1371/journal.pone.0130862