Identification of COVID-19 Vaccines Concerns in Health-Related French Web Forums: A Topic Modelling Approach

Pierre Karapetiantz, Bissan Audeh, Cédric Bousquet

Informatics and Technology in Clinical Care and Public Health
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210887

Identification of COVID-19 Vaccines Concerns in Health-Related French Web Forums : A Topic Modelling Approach

Pierre KARAPETIANTZ[a,1], Bissan AUDEH[a] and Cédric BOUSQUET[a]
[a] Inserm, Sorbonne Université, université Paris 13, Laboratoire d'informatique médicale et d'ingénierie des connaissances en e-santé, LIMICS, F-75006 Paris, *France*

Abstract. Since December 2019 and the first reported cases of COVID-19 in Wuhan, China, there have been 199,466,211 confirmed cases of COVID-19 in the World. The WHO defined vaccination hesitancy as one of the top ten threats to global health in 2019. Our objective was thus to identify topics and trends about COVID-19 vaccines from French web forums to understand the perception of the French population on these vaccines before the vaccination campaign started. We performed a topic model analysis on 485 web forums' posts. 10 topics were found. We reviewed 120 posts from 6 of these 10 topics. One topic was about vaccine hesitancy, refusal, and mistrust, and two topics were related to what the users think about the government, the political and economic choices made towards this epidemic.

Keywords. Pharmacovigilance, Social Media, Internet, Forum, Side effects, COVID-19 vaccines

1. Introduction

Since December 2019 and the first reported cases of COVID-19 in Wuhan, China, there have been 199,466,211 confirmed cases of COVID-19 in the World including 4,244,541 deaths according to the World Health Organization (WHO) [1]. Vaccines have been developed, and now a total of 3,984,374,918 vaccine doses have been administrated [1]. However, vaccine hesitancy is a hot topic that should be taken into consideration by vaccination campaigns, especially in France [2]. The WHO defined vaccination hesitancy as one of the top ten threats to global health in 2019 [3].

Social media are potentially interesting to investigate such hesitancy, and any information may be useful to health authorities to promote vaccine use, adherence, and confidence. Moreover, as the web 2.0 can affect the decision about getting vaccinated [4], exploring personal experiences and opinions from social media users is desirable to measure how such media may modify the perception of the population. Our objective was thus to identify topics and trends about COVID-19 vaccines from French web forums to understand the perception of the French population on these vaccines before the vaccination campaign started.

[1] Corresponding Author. Pierre Karapetiantz Limics, 15 rue de l'Ecole de Médecine, 75005 Paris, France; E-mail: pierre.karapetiantz@gmail.com.

2. Material and Methods

2.1. Material

We used the open-source tool Vigi4Med Scraper [5] to extract posts from web forums. The study period was from 1st February 2020 to 31st January 2021, before the vaccination campaign in France which started in February 2021.

2.2. Method

The preliminary data processing to reduce noise and incoherence [6] and the identification of COVID-19 vaccine related posts were done in 7 steps:

- Conversion to lower case: the text of all posts was converted to lower case text as R software (The R Project for Statistical Computing, Vienna) discriminates between lowercase and uppercase words.
- Identification of COVID vaccines related posts: These posts were selected according to the following decision rule:
    - The post contained the words coronavirus or COVID
    - The post contained the word vaccine, or any word related to COVID vaccines such as BioNTech, Cominarty, Pfizer, Johnson & Johnson, Moderna, AstraZeneca, Vaxzevria or Covishield. Spelling variations were tolerated.
- Punctuation and stop words removal
- Removal of certain identified words specific to web forums (e.g., users names)
- Removal of posts dealing about influenza
- Removal of overrepresented tokens that appear in more than 95% of the posts and in half of the posts or more and rare tokens that are present in less than 5% of the posts, less than 2 times and less than 5 posts. Words used to build the corpus (eg, covid) were thus removed as they were systematically present in all posts and did not carry any useful information.
- Multiple whitespace characters were collapsed to a single blank.

We generated *document-term matrix* (DTM) from processed users' posts. This matrix shows the frequency of terms that occur in the collection of posts: rows correspond to posts, and columns to terms. If a term occurs in a particular post, then the matrix entry is 1, if not it is 0. We decided to keep unigrams and bigrams. This made it possible to retain frequent contiguous sequences of two items, such as adverse effects (AEs).

We applied topic modeling with Latent Dirichlet Allocation (LDA) algorithm developed by Blei et al with the Variational Expectation-Maximization (VEM) algorithm [7, 8]. A 10-fold cross-validation was performed, testing the models for each number of topics from 2 to 10 topics. First, the data set was split into 10 test data sets (size of 49 posts) with the remaining data as training data. The number of topics for the final model was chosen according to the minimum perplexity in mean obtained by the 10 testing models and we initialized the parameter alpha of the final model considering the mean of the alphas generated for this number of topics [9]. For topics we found possibly relevant, 20 posts with less than 500 words were sampled for a manual review.

The motivation of the use of this method is that it answers our objective of identifying topics from users' posts. Indeed, when users write a post, they refer to a certain number of topics using words with a certain probability from the set of terms that correspond to that topic. Thus, each message contains several topics among all the identified topics, and the probability distribution shows how prominent the identified topics are in this message.

All statistical analyses were performed with the R language and environment for statistical computing and graphics [10] using the tm, textmineR and topicmodels packages [9, 11, 12].

3. Results

Among the 66 176 posts from 1st February 2020 to 31st January 2021, 704 were identified as concerning COVID-19 vaccines. 211 were excluded as they also referred to influenza, and 8 were excluded after data-management (removal of all the terms in the post). A total of 485 posts were included in the study.

After generating the DTM, we performed the 10-fold cross-validation. The perplexity we obtained is described in Figure 1.

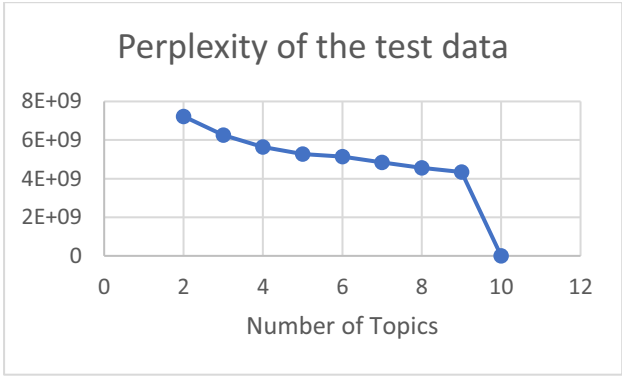

Figure 1. Perplexity according to the number of topics used

For the final model, we determined the number of topics to 10, and alpha was estimated as the mean of the 10 VEM test models for 10 topics.

After analyzing the most related words to each topic, we decided to manually review posts from topics 3, 4, 5 and 6 which seemed to be the most meaningful for our research. Table 1 gives the summary of this analysis.

Table 1. Main themes found for each of the studied topics generated

Topic	Main themes
3	Government-big pharma connivance; vaccination campaign and when it will start; children vaccination; health pass; chloroquine
4	Epidemic in Europe (beginning); vaccine research; vaccine refusal, hesitation and mistrust; civic responsibility
5	Chloroquine; hope for a vaccine, vaccine; Remdesivir; bored with politicians, Chloroquine AE; economic issues
6	Childbirth and COVID; how long before a vaccine?; COVID and children; collective immunity; conspiracy

4. Discussion

We identified 10 main topics. We reviewed 120 posts from 6 of these 10 topics. In future analysis, posts from topic 4 should be the object of a deeper manual analysis to learn more about vaccine hesitancy, refusal, and mistrust. Topics 3 and 5 could be explored to know more about what the users think about the government, the political and economic choices made towards this epidemic.

Our study had some limitations. As social media users may invent neologisms, we plan to reduce noise by selecting only words that are part of the French language, but there is a risk of losing potentially interesting information. Stemming was tried during this study, but results were difficult to interpret. Thus, lemmatization should be privileged. Finally, Dynamic Topic Models should also be explored with this kind of data to evaluate how topics may evolve.

References

[1] WHO, Coronavirus (COVID-19) Dashboard [Internet] 4 August 2021 [cited 4 August 2021] Available from: https://covid19.who.int/.
[2] Le Monde, Les Français sont les plus sceptiques face aux vaccins, selon une enquête mondiale. [Internet] 19 June 2019 [cited 4 August 2021] Available from: https://www.lemonde.fr/societe/article/2019/06/19/les-francais-sont-les-plus-sceptiques-face-aux-vaccins-selon-une-enquete-mondiale_5478259_3224.html.
[3] ANMJ, WHO's top ten threats to global health in 2019 [Internet] 3 May 2019 [cited 4 August 2021] Available from: https://anmj.org.au/whos-top-10-threats-to-global-health-in-2019/.
[4] Betsch C, Sachse K. Dr. Jekyll or Mr. Hyde? (How) the Internet influences vaccination decisions: recent evidence and tentative guidelines for online vaccine communication. Vaccine 2012;30(25):3723-6.
[5] Audeh B, Beigbeder M, Zimmermann A, Jaillon P, Bousquet C. Vigi4Med scraper: a framework for web forum structured data extraction and semantic representation. PloS one. 2017 Jan 25;12(1):e0169658.
[6] Iavindrasana J, Cohen G, Depeursinge A, et al. Clinical data mining: a review. Yearbook of Medical Informatics. 2009:121-133. PMID: 19855885.
[7] Blei DM, Lafferty JD. Topic Models. Srivastava AN, Sahami M, editors. Text Mining: Classification, Clustering, and Applications. CRC Press; 2009. p. 23.
[8] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res [Internet]. 2003;3:993-1022. Available from: http://portal.acm.org/citation.cfm?id=944937.
[9] Grün B, Hornik K. topicmodels: An R Package for Fitting Topic Models. Journal of Statistical Software [Internet]. 2011;40(13):1--30. Available from: http://www.jstatsoft.org/v40/i13/.
[10] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2016.
[11] Feinerer I, Hornik K. tm: Text Mining Package [Internet]. 2012. Available from: http://CRAN.R-project.org/package=tm.
[12] Tommy J. TextmineR: Functions for Text Mining and Topic Modeling. R package version 3.0.5 [Internet] 2021. Available from: https://CRAN.R-project.org/package=textmineR.