Croiser ADT et NLP pour caractériser les commentaires en ligne et détecter les tendances complotistes : le cas des vaccins

Pascal Marchand[1], Pierre Ratinaud[2]

[1]Université de Toulouse– pascal.marchand@iut-tlse3.fr

[2]Université de Toulouse – pierre.ratinaud@univ-tlse2.fr

Abstract

We analyze the comments on articles and reports published on the *France Info* website about vaccination from 07/27/2020 to 09/30/2022. The cleaned, lemmatized, recognized and segmented corpus is subject to a descending hierarchical classification (CDH; Iramuteq software) which makes it possible to define 15 interpretable classes according to a cognitive processing model. We then apply, to the characteristic segments of each of the classes resulting from the CDH, "NLP" models, as well as lexicometric indices (richness, temperature, etc.) or morphosyntactic indexing. The results make it possible to identify hypotheses for the detection of signs of "abnormality" in large textual corpora.

Keywords: Lexical classification, Natural Language Processing, morphosyntax, weak signals.

Résumé

On analyse les commentaires des articles et reportages publiés sur le site de France Info à propos de la vaccination du 27/07/2020 au 30/09/2022. Le corpus nettoyé, lemmatisé, reconnu et segmenté fait l'objet d'une classification hiérarchique descendante (CDH ; logiciel Iramuteq) qui permet de définir 15 classes interprétables selon un modèle de traitement cognitif. On applique ensuite, aux segments caractéristiques de chacune des classes issues de la CDH, des modèles « NLP », ainsi que des indices lexicométriques (richesse, température …) ou des indexations morphosyntaxiques. Les résultats permettent de dégager des pistes pour la détection de signes « d'anormalité » dans les grands corpus textuels.

Mots clés : Classification lexicale, *Natural Language Processing*, morphosyntaxe, signaux faibles.