Introduction à l'analyse de données textuelles dans R

Marc-Antoine Martel, Université de Montréal

Université de Montréal

CÉCD
CSDC

PLAN DE LA PRÉSENTATION

- L’analyse de données textuelles comme méthode de recherche
- Avantages / limites
- Quelques exemples de recherches
- Démonstration dans R (librairie Quanteda)

L'ANALYSE DE CONTENU

- Méthode très utilisée en sciences sociales
- Peut servir à analyser le comportement des acteurs politiques
  - Traitement de l'information par les médias
  - Type de messages diffusés par les partis politiques
  - Sujets débattus à l'Assemblée nationale
  - Désinformation sur les réseaux sociaux

LES ANALYSES AUTOMATISÉES

- Caractéristiques d'un corpus de textes
  - Longueur des textes
  - Occurrence des mots
  - Registres de langue
- Dictionnaires:
  - Lexiques associant une série de mots à des thèmes ou des attributs (enjeux politiques, tonalité, cadrages)
- Combiner des dictionnaires
  - Exemple: tonalité et enjeux.

AVANTAGES

- Larges corpus
- Recherches reproductibles
- Transparence
- Peu couteux
- Importante quantité de données à analyser

LIMITES

- Mise en garde
  - Taxer les riches, ne pas taxer les riches
  - Autobus
  - Tabac
- Métaphores, sarcasme…
- Créer un dictionnaire exhaustif peut être chronophage
- N'est pas un remplacement à l'analyse qualitative. Attention à l'interprétation. Importance de la validation. Attrait des méthodes mixtes.

DONNÉES À ANALYSER

- …tous les écrits disponibles au format numérique!
  - Articles de journaux
  - Plateformes électorales
  - Communiqués de presse
  - Publications sur les réseaux sociaux (Facebook, Twitter, YouTube, etc.)
  - Pages web, blogs, etc.
  - Transcriptions (période des questions à l'Assemblée nationale)

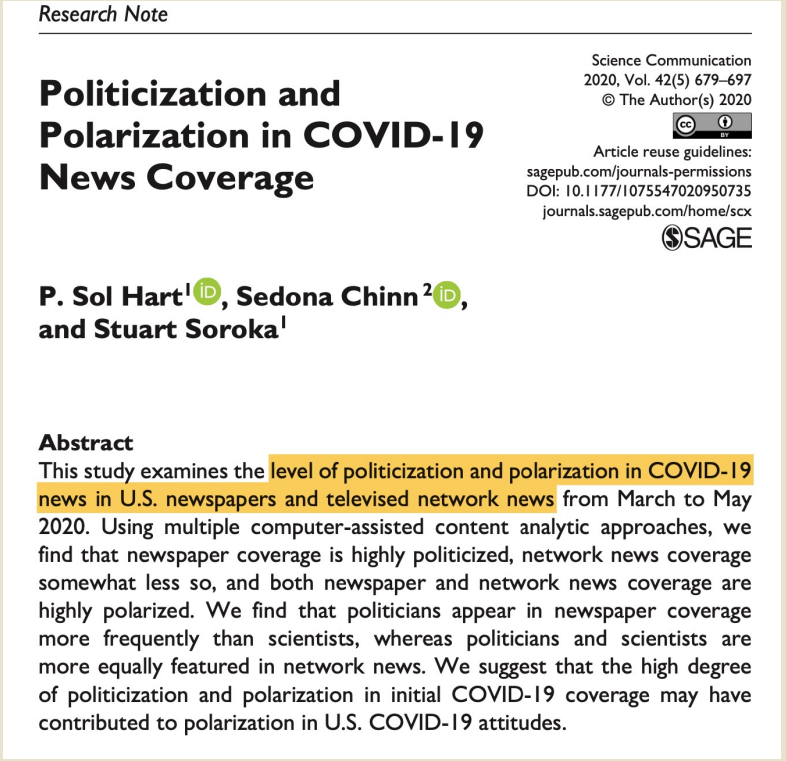
GARDER À L'ESPRIT

- Dictionnaires
  - Critères de sélection des mots à intégrer au dictionnaire
  - Désuétude
- Méthodes supervisées
- Méthodes non-supervisées (topic-models)

EXEMPLES DE RECHERCHES

- Niveau de politisation des nouvelles sur la COVID-19 en fonction de la visibilité des acteurs

- Hart, P. Sol, Sedona Chinn, and Stuart Soroka. « Politicization and polarization in COVID-19 news coverage. » *Science Communication* 42.5 (2020): 679-697.



Research Note

Science Communication
2020, Vol. 42(5) 679–697

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1075547020950735
journals.sagepub.com/home/scx
SAGE

Politicization and Polarization in COVID-19 News Coverage

P. Sol Hart[1], Sedona Chinn[2], and Stuart Soroka[1]

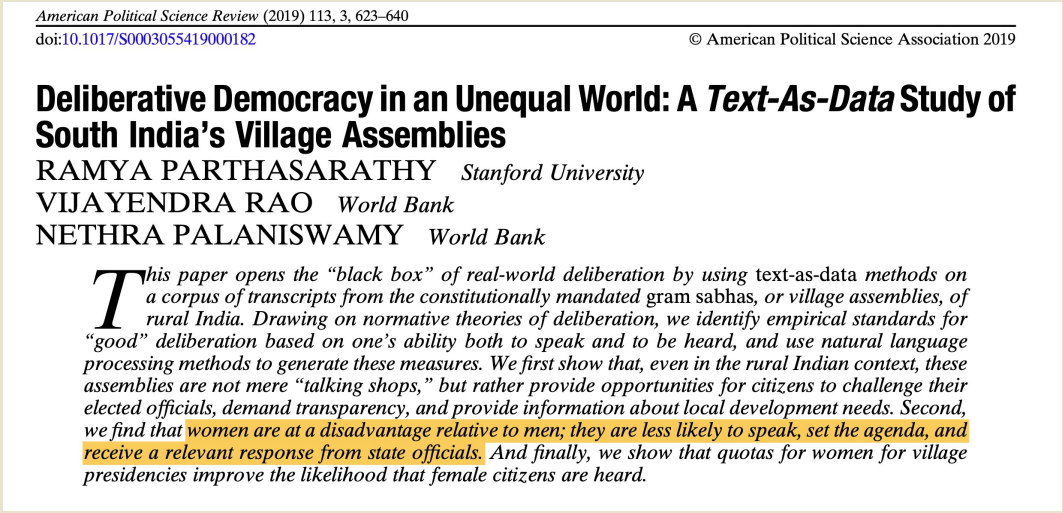
Abstract

This study examines the level of politicization and polarization in COVID-19 news in U.S. newspapers and televised network news from March to May 2020. Using multiple computer-assisted content analytic approaches, we find that newspaper coverage is highly politicized, network news coverage somewhat less so, and both newspaper and network news coverage are highly polarized. We find that politicians appear in newspaper coverage more frequently than scientists, whereas politicians and scientists are more equally featured in network news. We suggest that the high degree of politicization and polarization in initial COVID-19 coverage may have contributed to polarization in U.S. COVID-19 attitudes.

EXEMPLES DE RECHERCHES

- Transcription des interventions de citoyens lors d'assemblées de village.
- Capacité à influencer la discussion:
  - Comparaison du sujet abordé par une personne aux sujets abordés lors des interventions subséquentes.

Parthasarathy, Ramya, Vijayendra Rao et Nethra Palaniswamy. 2019. "Deliberative democracy in an unequal world: A text-as-data study of South India's village assemblies." *American Political Science Review* 113: 623-640.



American Political Science Review (2019) 113, 3, 623–640
doi:10.1017/S0003055419000182

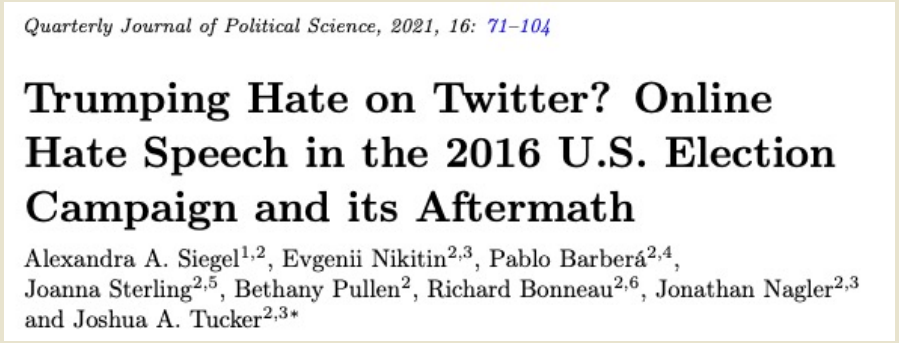
Deliberative Democracy in an Unequal World: A *Text-As-Data* Study of South India's Village Assemblies

RAMYA PARTHASARATHY *Stanford University*
VIJAYENDRA RAO *World Bank*
NETHRA PALANISWAMY *World Bank*

This paper opens the "black box" of real-world deliberation by using text-as-data *methods on a corpus of transcripts from the constitutionally mandated* gram sabhas, *or village assemblies, of rural India. Drawing on normative theories of deliberation, we identify empirical standards for "good" deliberation based on one's ability both to speak and to be heard, and use natural language processing methods to generate these measures. We first show that, even in the rural Indian context, these assemblies are not mere "talking shops," but rather provide opportunities for citizens to challenge their elected officials, demand transparency, and provide information about local development needs. Second, we find that women are at a disadvantage relative to men; they are less likely to speak, set the agenda, and receive a relevant response from state officials. And finally, we show that quotas for women for village presidencies improve the likelihood that female citizens are heard.*

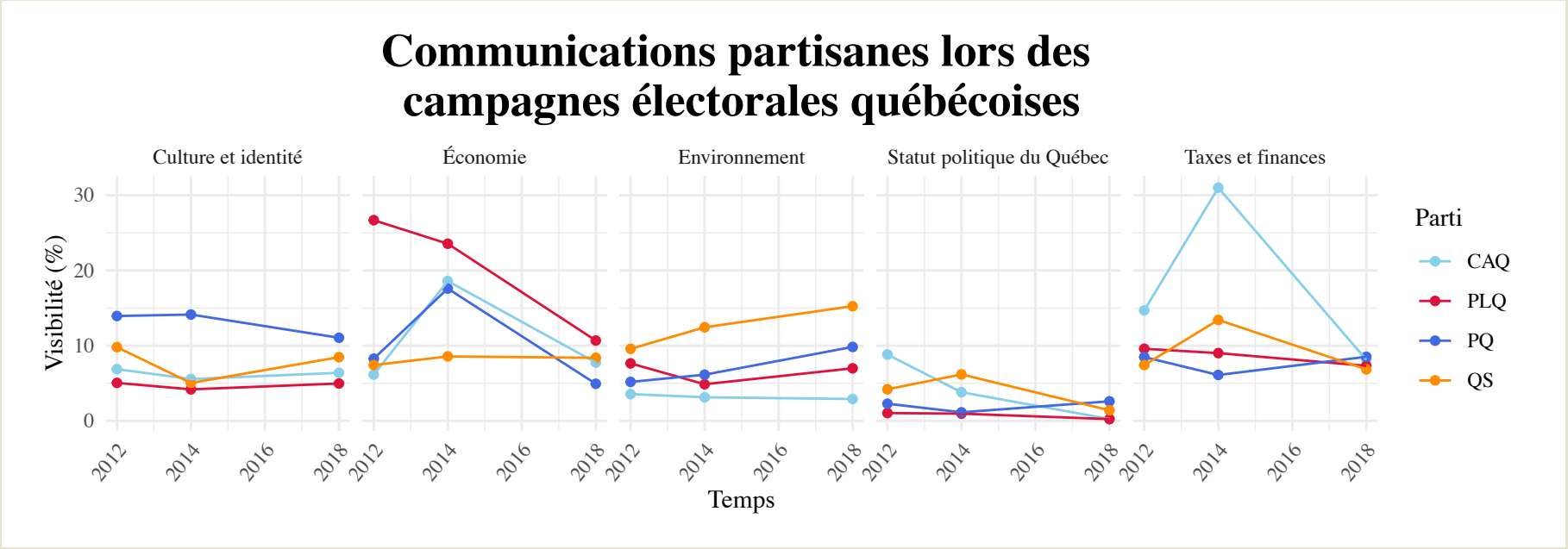
EXEMPLES DE RECHERCHES

- Plus de 750 millions de tweets liés à l'élection, en plus de près de 400 millions de tweets provenant d'un échantillon aléatoire d'utilisateurs américains de Twitter
- Communautés Reddit associées au mouvement alt-right
- Discours haineux au cours de la campagne et dans les six mois suivant l'élection de Trump.



- Alexandra A. Siegel, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler and Joshua A. Tucker (2021), "Trumping Hate on Twitter? Online Hate Speech in the 2016 U.S. Election Campaign and its Aftermath", Quarterly Journal of Political Science: Vol. 16: No. 1, pp 71-104. http://dx.doi.org/10.1561/100.00019045

VISUALISATION DES RÉSULTATS



DÉMONSTRATION

- Pour télécharger les scripts:

https://github.com/ma-martel/Atelier_CECD_2022/archive/refs/heads/main.zip

- Sources:
  - https://quanteda.io/
  - Duval, D. and Pétry, F. (2016) "L'analyse automatisée du ton médiatique : construction et utilisation de la version française du Lexicoder Sentiment Dictionary", *Revue canadienne de science politique*, 49(2), pp. 197–220.

- Partage d'écran

RESSOURCES

- Twitter: https://developer.twitter.com/en/docs/twitter-api
- Facebook: https://www.crowdtangle.com/
- Articles de journaux: http://eureka.cc/fr/
- Archives de sites web: https://archive.org/web/
- Web scraping: https://github.com/benjaminguinaudeau/tidybrowse
- Collection de l'Université Laval: https://www.poltext.org/fr
- Packages R: tm, tidytext, Quanteda (https://tutorials.quanteda.io/)

RECOMMANDATIONS DE LECTURE

- Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press, 2022.
- Grimmer, Justin, et Brandon M. Stewart. 2013. « Text as data: The promise and pitfalls of automatic content analysis methods for political texts ». *Political analysis* 21 (3) : 267-97.
- Young, Lori, et Stuart Soroka. 2012. « Affective News: The Automated Coding of Sentiment in Political Texts ». *Political Communication* 29 (2) : 205-31. https://doi.org/10.1080/10584609.2012.671234.

MERCI ☺

- N’hésitez pas à me contacter pour discuter!
- Twitter: https://twitter.com/_MAMartel
- Courriel: marc-antoine.martel@umontreal.ca