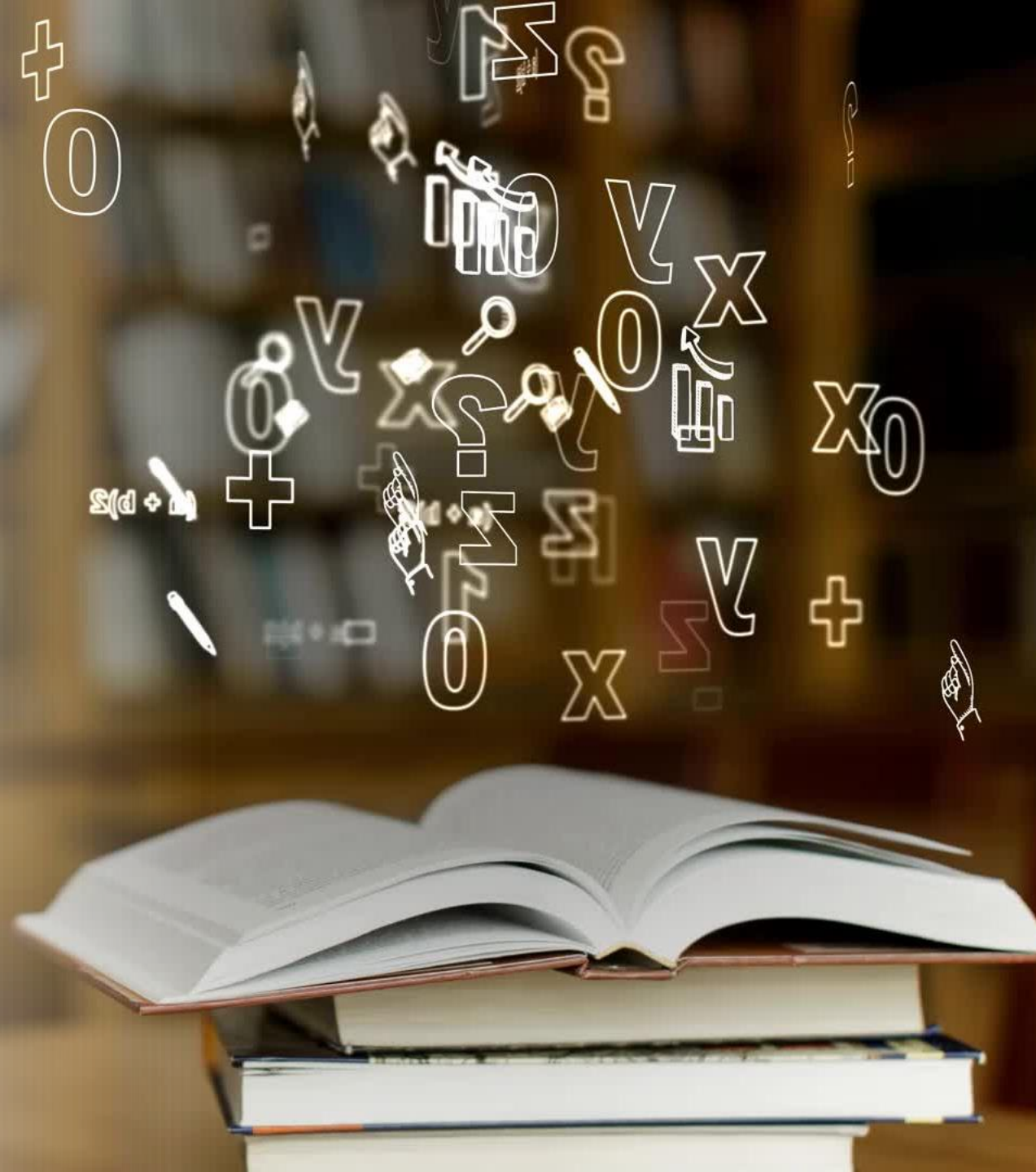# Revue systématique de la littérature et méta-analyse

Laurence Dierickx

DIFEM 2022

SLR

A systematic literature review is a means of collecting and synthesising previous research, providing an overview of areas covered by the research, synthesising studies, and showing evidence on a meta-level (Snyder, 2019).

Its main objective is to answer specific questions by relying on rigorous and explicit methods to reduce bias because of the work's transparency, transferability, and replicability (Thomas and Harden, 2008).

While systematic literature reviews are commonly used in the medical and computer sciences, they can also be used in social sciences to provide an overall picture of the evidence in a topic area to direct future research (Petticrew and Robert, 2006: 21),

Méta-analyse

Meta-analysis refers to statistical techniques that can be descriptive or inferential to obtain overall estimations or a synthesis.
(Shelby and Vaske, 2008)

Together with a meta-analysis, a systematic literature review can help clarify the state of an area of research.
(Davis et al., 2014; Mengist et al., 2020)

En sciences sociales

Key Messages

Implications for Practice

- Conventional subject searching may miss significant articles for inclusion in social science systematic reviews.
- Problems in indexing and abstract content make it difficult to devise a sensitive and exhaustive search strategy; inclusion of alternative search methods such as citation searching, reference list checking and contact with experts thus becomes essential.
- The number of higher quality studies identified by citation searching and reference list checking appears to be greater when compared with the proportion of higher quality studies found by database searching.

Implications for Policy

- Systematic searching of the social science literature requires a range of search techniques including citation searching, reference list checking and contact with experts

Papaioannou, D., Sutton, A., Carroll, C., Booth, A., & Wong, R. (2010). Literature searching for social science systematic reviews: consideration of a range of search techniques. *Health Information & Libraries Journal*, *27*(2), 114-122.

Exemples

Bhaskaran, H., Kashyap, G., & Mishra, H. (2022). Teaching Data Journalism: A Systematic Review. *Journalism Practice*, 1-22.

This study attempts to systematically review the peer-reviewed academic literature on data journalism training in order to ascertain the present status of academic research on the subject. By examining the studies, it brings together insights about the prevalent methods used in data journalism training, the challenges faced by the instructors, the recommended best practices and the students' perception about data journalism training.

Reyes-de-, S., Pérez-Escolar, M., & Navazo-Ostúa, P. (2022). Digital competencies for new journalistic work in media outlets: A systematic review. *Media and communication*, *10*(1), 27-42.

What does the literature suggest about the digital skills that new professional profiles should acquire in the field of journalism? Which dimensions of digital competence are gaining visibility and which dimensions are being neglected?

Danzon-Chambaud, S. (2021). A systematic review of automated journalism scholarship: guidelines and suggestions for future research. *Open Research Europe*, *1*, 4.
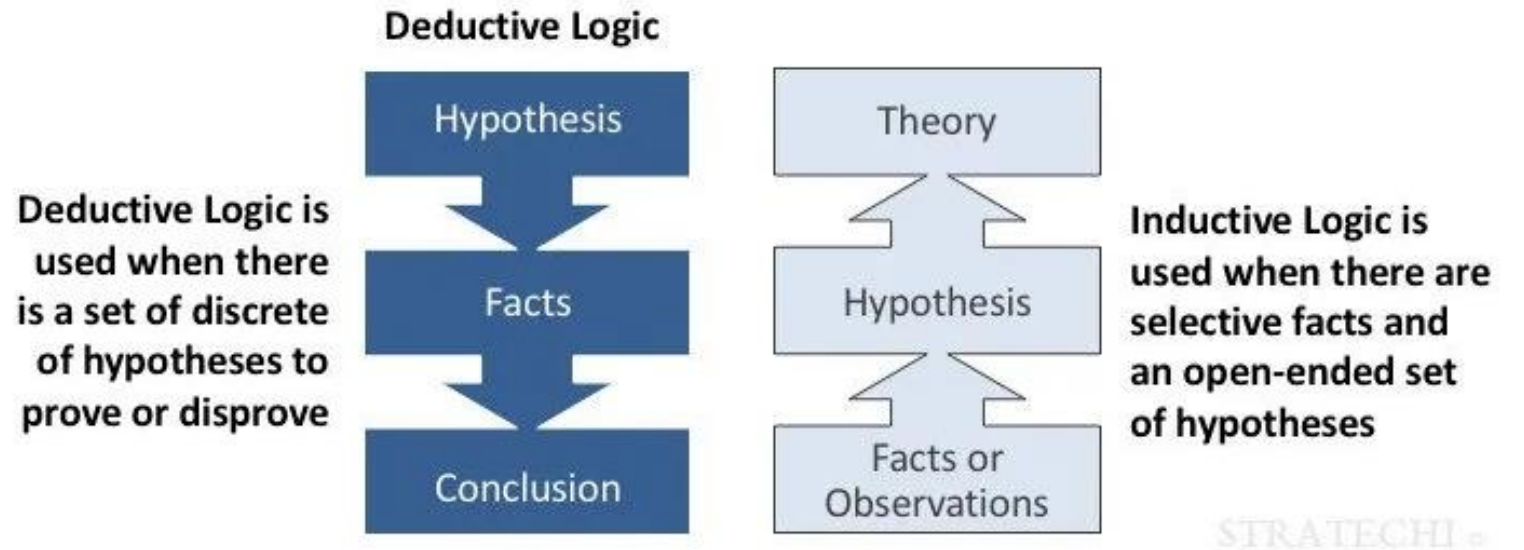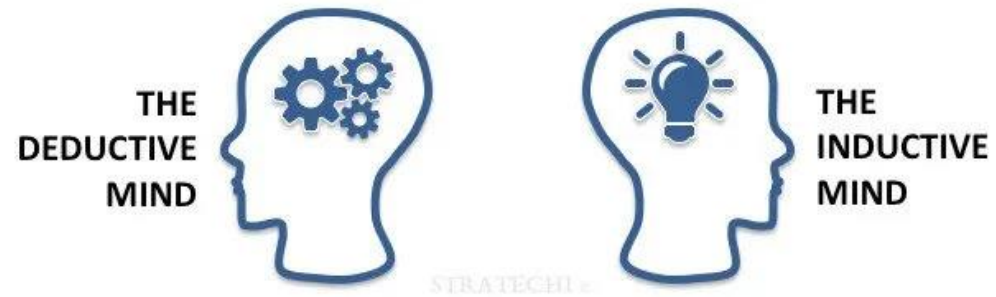
This systematic literature therefore provides researchers with an overview of the main challenges and debates that are occurring within the field of automated journalism studies.

Approche inductive pour l'analyse de données

An outline of a general inductive approach for qualitative data analysis is described and details provided about the assumptions and procedures used. The purposes for using an inductive approach are to (1) to condense extensive and varied raw text data into a brief, summary format; (2) to establish clear links between the research objectives and the summary findings derived from the raw data and (3) to develop of model or theory about the underlying structure of experiences or processes which are evident in the raw data. The inductive approach reflects frequently reported patterns used in qualitative data analysis. Most inductive studies report a model that has between three and eight main categories in the findings. The general inductive approach provides a convenient and efficient way of analysing qualitative data for many research purposes. The outcomes of analysis may be indistinguishable from those derived from a grounded theory approach. Many researchers are likely to find using a general inductive approach more straightforward than some of the other traditional approaches to qualitative data analysis.

Thomas, D. R. (2003). A general inductive approach for qualitative data analysis.

Logique déductive vs inductive

THE DEDUCTIVE MIND

THE INDUCTIVE MIND

STRATECHI

Deductive Logic

Deductive Logic is used when there is a set of discrete of hypotheses to prove or disprove

Hypothesis

Facts

Conclusion

Theory

Hypothesis

Facts or Observations

Inductive Logic is used when there are selective facts and an open-ended set of hypotheses

STRATECHI

ML & sciences sociales

Abstract

Social scientists are now in an era of data abundance, and machine learning tools are increasingly used to extract meaning from data sets both massive and small. We explain how the inclusion of machine learning in the social sciences requires us to rethink not only applications of machine learning methods but also best practices in the social sciences. In contrast to the traditional tasks for machine learning in computer science and statistics, when machine learning is applied to social scientific data, it is used to discover new concepts, measure the prevalence of those concepts, assess causal effects, and make predictions. The abundance of data and resources facilitates the move away from a deductive social science to a more sequential, interactive, and ultimately inductive approach to inference. We explain how an agnostic approach to machine learning methods focused on the social science tasks facilitates progress across a wide range of questions.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science, 24*, 395-419.

Comprendre le ML

Abstract

Machine learning is a field at the intersection of statistics and computer science that uses algorithms to extract information and knowledge from data. Its applications increasingly find their way into economics, political science, and sociology. We offer a brief introduction to this vast toolbox and illustrate its current uses in the social sciences, including distilling measures from new data sources, such as text and images; characterizing population heterogeneity; improving causal inference; and offering predictions to aid policy decisions and theory development. We argue that, in addition to serving similar purposes in sociology, machine learning tools can speak to long-standing questions on the limitations of the linear modeling framework, the criteria for evaluating empirical findings, transparency around the context of discovery, and the epistemological core of the discipline.

Molina, M., & Garip, F. (2019). Machine learning for sociology. *Annual Review of Sociology*, *45*, 27-45.

https://www.annualreviews.org/doi/full/10.1146/annurev-soc-073117-041106?casa_token=ltPXpKd_6ZEAAAAA:x4zKlcQL13pY9zQrYuXMcSSoobUr06FjfaCl_n8c70C6EhRy7T_GUW_Qc4DIUxHeI33su-jjaJBN

Automated fact-checking to support professional practices
Systematic review and meta-analysis

Global Facts – Academic Track – June 22, 2022

Laurence Dierickx, Carl-Gustav Lindén, Duc Tien Dang Nguyen, Andreas Opdahl
University of Bergen

PRISMA

Preferred Reporting Items for Systematic reviews and Meta-Analyses)

To ensure transparent and complete reporting (Liberati et al., 2009).

Checklist of 27 items that frame the method, the writing of a systematic review report, and a flow diagram that shapes information retrieval and selection (Page et al., 2021).

Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71

https://www.prisma-statement.org/

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	

	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.

RESULTS		
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.
Study characteristics	17	Cite each included study and present its characteristics.
Risk of bias in studies	18	Present assessments of risk of bias for each included study.
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.
	20c	Present results of all investigations of possible causes of heterogeneity among study results.
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.

DISCUSSION		
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.
	23b	Discuss any limitations of the evidence included in the review.
	23c	Discuss any limitations of the review processes used.
	23d	Discuss implications of the results for practice, policy, and future research.

OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	
Competing interests	26	Declare any competing interests of review authors.	
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	

What is a protocol?

A systematic review protocol describes the rationale, hypothesis, and planned methods of the review. It should be prepared before a review is started and used as a guide to carry out the review.

Defining the scope

1. Mapping the field of automated fact-checking

2. Identifying works related to end-users

3. Identifying works related to journalism practices

Are journalists as end-users considered, and how?

How as boundary objects AFC tools connects and disconnects between communities?

Définir
le périmètre

The tool Parsifal, available online, was used to define the research question, a set of keywords, query strings, and inclusion and exclusion criteria (da Silva et al., 2022).

PARSIFAL

https://parsif.al/

PICOC

Méthode pour décrire les cinq éléments d'une question de recherche (evidence-based)

Population   Who?

Intervention   What or How?

Comparison   Compared to what?

Outcome   What are you trying to accomplish / improve?

Context   In what kind of organization / circumstances?

Collecting & selecting the corpus

Queries

- fact-checking AND "machine learning" *(Google Scholar, 205)*
- automated AND fact-checking AND journalism *(Semantic Scholar, 338)*
- ("machine learning" OR "automated") AND fact-checking *(Scopus, 375)*

Inclusion criteria

Published peer-reviewed, open-access arXiv (not PR), book chapters, proceedings
In the scope: AFC systems or tools, end-user perspective, journalism/newsrooms

Exclusion criteria

Books, unpublished papers (outside arXiv), dissertations, duplicates, undated, < 2017

Collecte du corpus: Google Scholar

Scraping Google Search (without getting caught)

A scraping method resilient to IP blocking

> *Disclaimer: use ideas here under your own responsibility.*

The IP blocking problem

If you are into web scraping you probably know that websites don't like automated bots that pay a visit just to gather information. They have set up systems which can figure out that your program is not an actual person and, after a bunch of requests coming from your script, you usually get the dreadful **HTTP 429 Too Many Requests Error**. This message means that your IP address has been blocked from querying the website for a certain amount of time. Your bot can go home and cry.

https://juanluisrto.medium.com/scraping-google-search-without-getting-caught-e43bb91b363e

Collecte du corpus: Semantic Scholar

"fact-checking" AND "automated" AND "journalism" NOT "manual"

SEMANTIC **SCHOLAR**   "fact-checking"

5 results for ""fact-checking" AND "automated" AND "journal

Fields of Study ⌄   Date Range ⌄   Has PDF   Publication Type ⌄   Author ⌄

omated Fact-Checking for Assisting Human Fact-Checkers

av Nakov, D. Corney, +6 authors Giovanni Da San Martino • Computer Science • IJCAI • 13 March 202

R Here, a survey of the available intelligent technologies that can support the human exper
ct-checking endeavor includes identifying claims worth fact- checking, detecting relevant
ns, retrieving relevant evidence to fact-check a claim, **and** actually verifying a claim. Expand

51 PDF • 📄 View PDF on arXiv • 🔖 Save • 🔔 Alert • 66 Cite

nish Fact-Checking Services: An Approach to Their Business Models

el-Ángel Esteban-Navarro, Antonia-Isabel Nogales-Bocio, Miguel-Ángel García-Madurga, Tamara Morte-Nad

ugust 2021

R The findings are that the fact-checking services that depend on media groups are no stra
ity usual in these groups, but in the case of fact- checking services that are born as initiativ
d towards transparency is, in the majority of cases, clear. Expand

PDF • 📄 View PDF • 🔖 Save • 🔔 Alert • 66 Cite

About 6,850 results for ""fact-checking""

Fields of Study ⌄   Date Range ⌄   Has PDF   Publication Type ⌄   Author ⌄

Fake news agenda in the era of COVID-19: Identifying trends through fact-checki

Wilson Ceron, Mathias-Felipe de-Lima-Santos, M. Quiles • Computer Science • Online Soc. Networks Media

66 33 PDF • 📄 View PDF on arXiv • 🔖 Save • 🔔 Alert • 66 Cite

DESINFORMACIÓN Y FACT-CHECKING EN LAS ELECCIONES URUGUAYAS DE 201
VERIFICADO URUGUAY

Juan-Pedro Molina-Cañabate, Raúl Magallón-Rosa • Political Science • Perspectivas de la comunicación

El objetivo de esta investigacion es estudiar los procesos de desinformacion que circularon d
presidenciales uruguayas de 2019. Para ello, se ha examinado el trabajo realizado... Expand

66 2 PDF • ↗ View via Publisher • 🔖 Save • 🔔 Alert • 66 Cite

API Semantic Scholar

https://www.semanticscholar.org/product/api

https://api.semanticscholar.org/api-docs/graph

https://api.semanticscholar.org/graph/v1/paper/search?query=fact-checking

SEMANTIC SCHOLAR Academic Graph API   Recommendations API   Datasets API   Peer Review API

Q Search...

Paper Data >
Author Data >

Academic Graph API (1.0)

Download OpenAPI specification: [Download]

Fetch paper and author data from the Semantic Scholar Academic Graph (S2AG)

Paper Data

Search for papers by keyword

QUERY PARAMETERS

⊣ offset
    integer
    Default: 0
    When returning a list of results, start with the element at this position in the list.
    The sum of `offset` and `limit` must be < 10000

⊣ limit
    integer
    Default: 100
    The maximum number of results to return.
    Must be <= 100

API docs by Redocly

Access Our Continually Updating Corpus

203 Million	2.2 Billion	74 Million
Papers	Citations	Authors

Automatiser la collecte des PDF

https://towardsdatascience.com/scraping-downloading-and-storing-pdfs-in-r-367a0a6d9199

WEB SCRAPING PDFS

Scraping, Downloading, and Storing PDFs in R
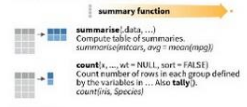
Nesting scrapes to avoid button-clicking

. . .

```
library(tidyverse)
library(rvest)
library(stringr)
```
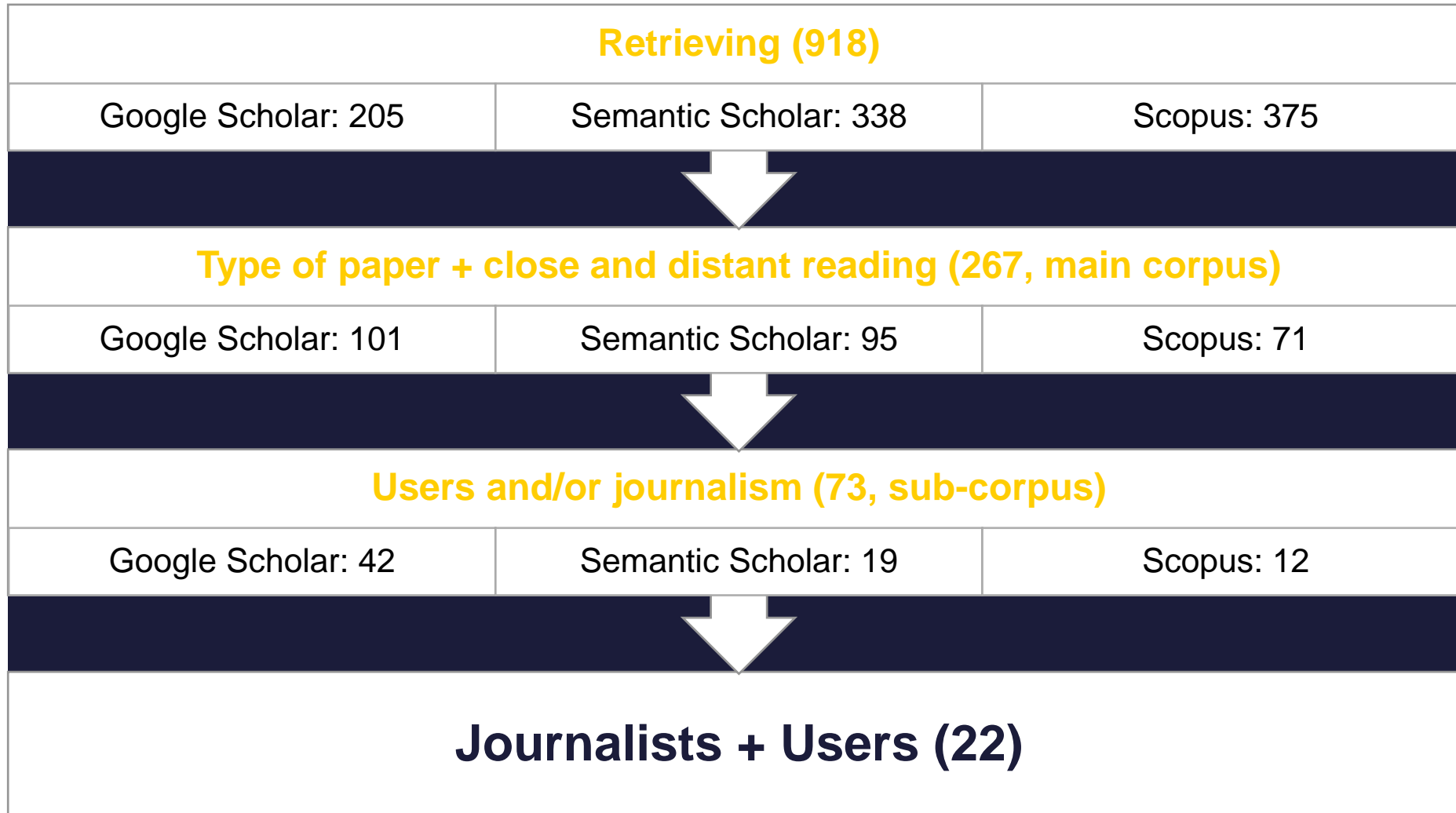
Préparation du corpus

Elimination des doublons

Premier tri basé sur le titre
et le résumé de l'article (lecture)

Close et distant reading pour constituer
le premier sous-corpus (50%).

Extraction du texte (résumé, contenu)

Manuel : nécessite beaucoup de temps mais parfois la seule solution lorsque le PDF est mal encodé

Automatique : suppose un bon encodage du PDF

Nettoyage : les articles académiques comportent de nombreuses mentions non utiles pour l'analyse de contenu (auteur, titre, référence, remerciements etc.)

Importance du contrôle humain !

Organisation des données

Code	Title	Abstract	APA	Field	Type	Year	Journalism	Use related	Citations
ML_DB_002	The current state of fake news: challenges ar	The authenticity of Information has become a longstand	Figueira, Á., & Oliveira, L. (2017). The current state of f	Computer Science	Proceeding	2017 N	N	208
ML_DB_003	An interpretable model to measure fakeness a	Fake news and post-truth are everywhere. The huge nu	Gadek, G., & Guélorget, P. (2020). An interpretable	Computer Science	Proceeding	2020 N	N	5
ML_DB_004	FactRank: Developing automated claim detec	Fact-checking has always been a central task of journa	Berendt, B., Burger, P., Hautekiet, R., Jagers, J., Pleijt	Computer Science	Article	2021 Y	Y	3
ML_DB_005	Fake news agenda in the era of COVID19: Ide	The rise of social media has ignited an unprecedented	Ceron, W., de-Lima-Santos, M. F., & Quiles, M. G. (2	Social Computing	Article	2021 N	N	33
ML_DB_006	Metalearning for fake news detection surrounc	In this article, we pursue the automatic detection of fake	Salem, F. K. A., Al Feel, R., Elbassuoni, S., Ghannam,	Computer Science	Article	2021 N	N	1
ML_DB_007	Fake news detection: A hybrid CNNRNN base	The explosion of social media allowed individuals to spr	Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake r	Computer Science	Article	2021 N	N	75
ML_DB_008	A machine learning based framework for dete	Daily news is one of the primary needs of modern socie	Purevdagva, C., Zhao, R., Huang, P. C., & Mahoney,	Computer Science	Proceeding	2020 Y	N	1
ML_DB_009	Disinformation and misinformation triangle: A	The purpose of this paper is to treat disinformation and	Rubin, V. L. (2019). Disinformation and misinforma	Social Computing	Article	2019 N	N	59
ML_DB_011	A comparison of fake news detecting and fact	Scientific objective of this paper is to analyse how adva	Školkay, A., & Filin, J. (2019). A comparison of fake	Social Science	Article	2019 N	Y	3
ML_DB_012	Can machines learn to detect fake news? A s	Through a systematic literature review method, in this v	Cardoso Durier da Silva, F., Vieira, R., & Garcia, A. C	Computer Science	Proceeding	2019 N	N	34
ML_DB_013	Liar, liar pants on fire: A new benchmark data	Automatic fake news detection is a challenging problem	Wang, W. Y. (2017). " liar, liar pants on fire": A new	Computer Science	Article	2017 N	N	968
ML_DB_014	Automatically identifying fake news in popular	Information quality in social media is an increasingly im	Buntain, C., & Golbeck, J. (2017, November). Autom	Social Computing	Proceeding	2017 Y	N	186
ML_DB_015	Fully automated fact checking using external	Given the constantly growing proliferation of false claim	Karadzhov, G., Nakov, P., Màrquez, L., Barrón-Cede	Computer Science	Article	2017 N	N	101
ML_DB_016	Automated fact checking: Task formulations,	The recently increased focus on misinformation has sti	Thorne, J., & Vlachos, A. (2018). Automated fact ch	Computer Science	Article	2018 N	N	191
ML_DB_017	Predicting factuality of reporting and bias of n	We present a study on predicting the factuality of report	Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., & N	Computer Science	Article	2018 N	N	158
ML_DB_019	Multitask ordinal regression for jointly predictir	In the context of fake news, bias, and propaganda, we s	Baly, R., Karadzhov, G., Saleh, A., Glass, J., & Nakov	Computer Science	Article	2019 N	N	35
ML_DB_020	Open issues in combating fake news: Interpre	Combating fake news needs a variety of defense meth	Mohseni, S., Ragan, E., & Hu, X. (2019). Open issues	Computer Science	Article	2019 N	N	16
ML_DB_021	Explainable fact checking with probabilistic ar	One challenge in fact-checking is the ability to improve	Ahmadi, N., Lee, J., Papotti, P., & Saeed, M. (2019).	Computer Science	Article	2019 N	N	24
ML_DB_022	Automatic fact-checking using context and dis	We study the problem of automatic fact-checking, payir	Atanasova, P., Nakov, P., Màrquez, L., Barrón-Cede	Computer Science	Article	2019 N	N	29
ML_DB_023	A richly annotated corpus for different tasks ir	Automated fact-checking based on machine learning is	Hanselowski, A., Stab, C., Schulz, C., Li, Z., & Gurevy	Computer Science	Article	2019 N	N	45
ML_DB_024	A Context Aware Approach for Detecting Che	In the context of investigative journalism, we address th	Gencheva, P., Koychev, I., Màrquez, L., Barrón-Cede	Computer Science	Article	2020 Y	N	1
ML_DB_028	Linked credibility reviews for explainable misir	In recent years, misinformation on the Web has becom	Denaux, R., & Gomez-Perez, J. M. (2020, November	Computer Science	Proceeding	2020 N	Y	10
ML_DB_029	TRUSTD: Combat Fake Content using Block	The growing trend of sharing news/contents, through sc	Jaroucheh, Z., Alissa, M., Buchanan, W. J., & Liu, X. (	Computer Science	Proceeding	2020 N	Y	5
ML_DB_030	Explainable automated fact-checking for publi	Fact-checking is the task of verifying the veracity of clai	Kotonya, N., & Toni, F. (2020). Explainable automat	Computer Science	Article	2020 N	N	30
ML_DB_031	Fighting an infodemic: Covid19 fake news dat	Along with COVID19 pandemic we are also fighting an	Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, (	Computer Science	Proceeding	2021 N	N	92
ML_DB_032	Automated fact-checking for assisting human	The reporting and the analysis of current events around	Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsaye	Social Science	Article	2021 Y	Y	28
ML_DB_033	Community-Based Fact-Checking on Twitter'	Misinformation undermines the credibility of social medi	Pröllochs, N. (2021). Community-Based Fact-Checki	Social Science	Article	2021 N	N	4
ML_DB_034	NoFake at CheckThat! 2021: fake news detec	Much research has been done for debunking and analy:	Kumari, S. (2021). NoFake at CheckThat! 2021: fake	Computer Science	Article	2021 N	N	2
ML_DB_035	Technological Approaches to Detecting Onlin	The move of propaganda and disinformation to the onlir	Horák, A., Baisa, V., & Herman, O. (2021). Technolo	Computer Science	Book Chapter	2021 N	N	1
ML_DB_036	A survey on automated fact-checking	Fact-checking has become increasingly important due	Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A su	Computer Science	Article	2021 N	N	5
ML_DB_037	Scalable Fact-checking with Human-in-the-Lc	Researchers have been investigating automated solutic	Yang, J., Vega-Oliveros, D., Seibt, T., & Rocha, A. (20	Computer Science	Proceeding	2021 N	N	1

Annotations

Domaine : informatique, journalisme etc.

Inclusion des utilisateurs finaux (oui/non)

Considération du contexte journalistique (oui/non)

ID	Code	Title	Abstract	Field	Type	Year	Journalism	Use related	Citations	Text
AFC_001	ML_DB_130	Supervised Le	A large body	Computer Sci	Article	2019	N	Y	192	SOCIAL MEDIA SYSTEMS have been dr
AFC_002	ML_DB_098	Investigating t	The creation a	Computer Sci	Article	2019	N	Y	31	When a headline asks a question, the an
AFC_003	ML_DB_056	A Conceptual	With the incre	Social Scienc	Proceeding	2021	N	Y	0	A Conceptual Model for Approaching the Des

Pre-processing

Préparation des données en six étapes

Elimination des doublons, harmonisation orthographique (US vs UK English)

Eliminating duplicates → Data cleaning → Data **standardization** and spelling **normalization** → Tokenization → Filtering by term frequency → Listing and removing stop word

Stop Words

En traitement automatique de la langue, les stop words sont des termes dont la signification manque de pertinence et qui peuvent donc être ignorés (Wilbur et Sirotkin, 1992 ; Lo et al, 2005). En d'autres mots, il s'agit d'un dictionnaire négatif (Fox, 1989).

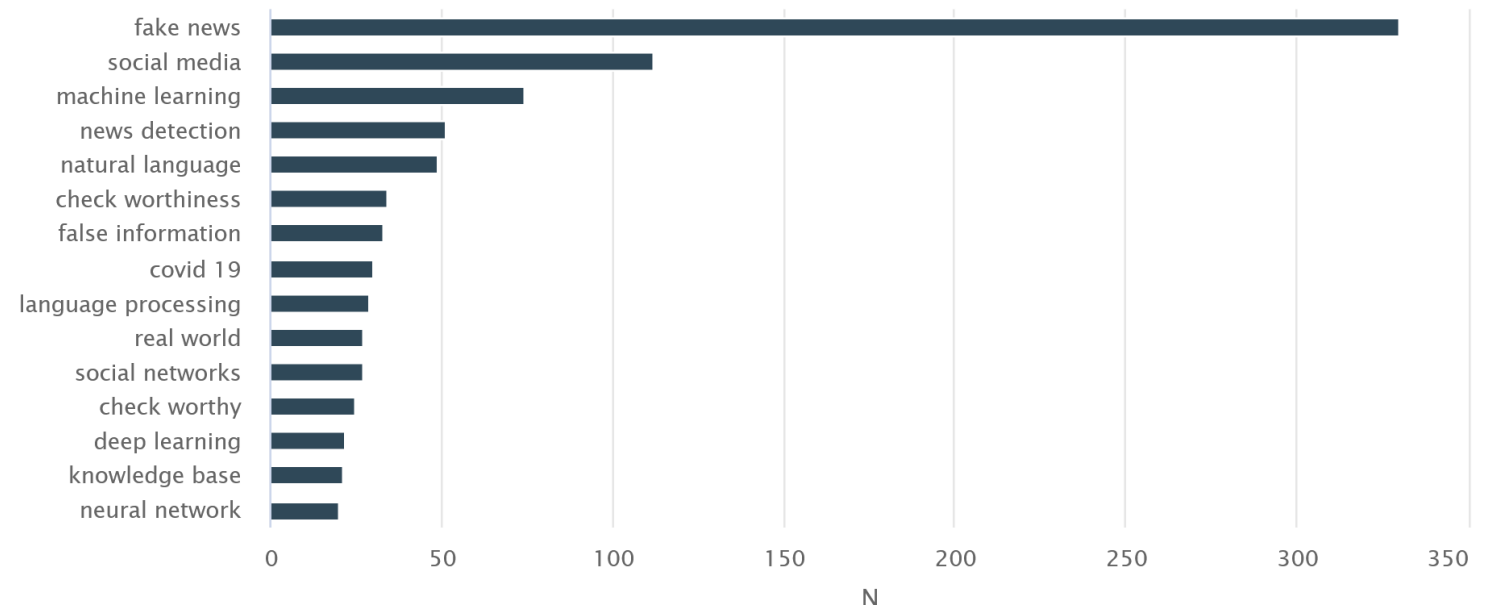Il existe des listes prêtes à l'emploi mais pas toujours adaptées.

Brown stop list (Fox, 1992) et la Van stoplist (Rijsbergen, 1979) sont considérés comme des standards (Saif et al., 2014)

Généralement, les pronoms, interjections et conjonctions peuvent être considérés comme des stop words. Etablir une liste de stop words doit à la fois tenir compte du domaine et de l'évolution de la langue. La fréquence des termes ou des tokens d'un corpus permet de travailler sur une liste de stop words, mais il n'y a pas de consensus à propos de celle à privilégier : les termes les plus fréquents doivent-il être exclus ou, au contraire, faut-il ignorer les termes à plus basse fréquence ? La meilleure méthode est celle qui permet d'aboutir aux meilleurs résultats (Fox, 1989 ; Saif et al., 2014).
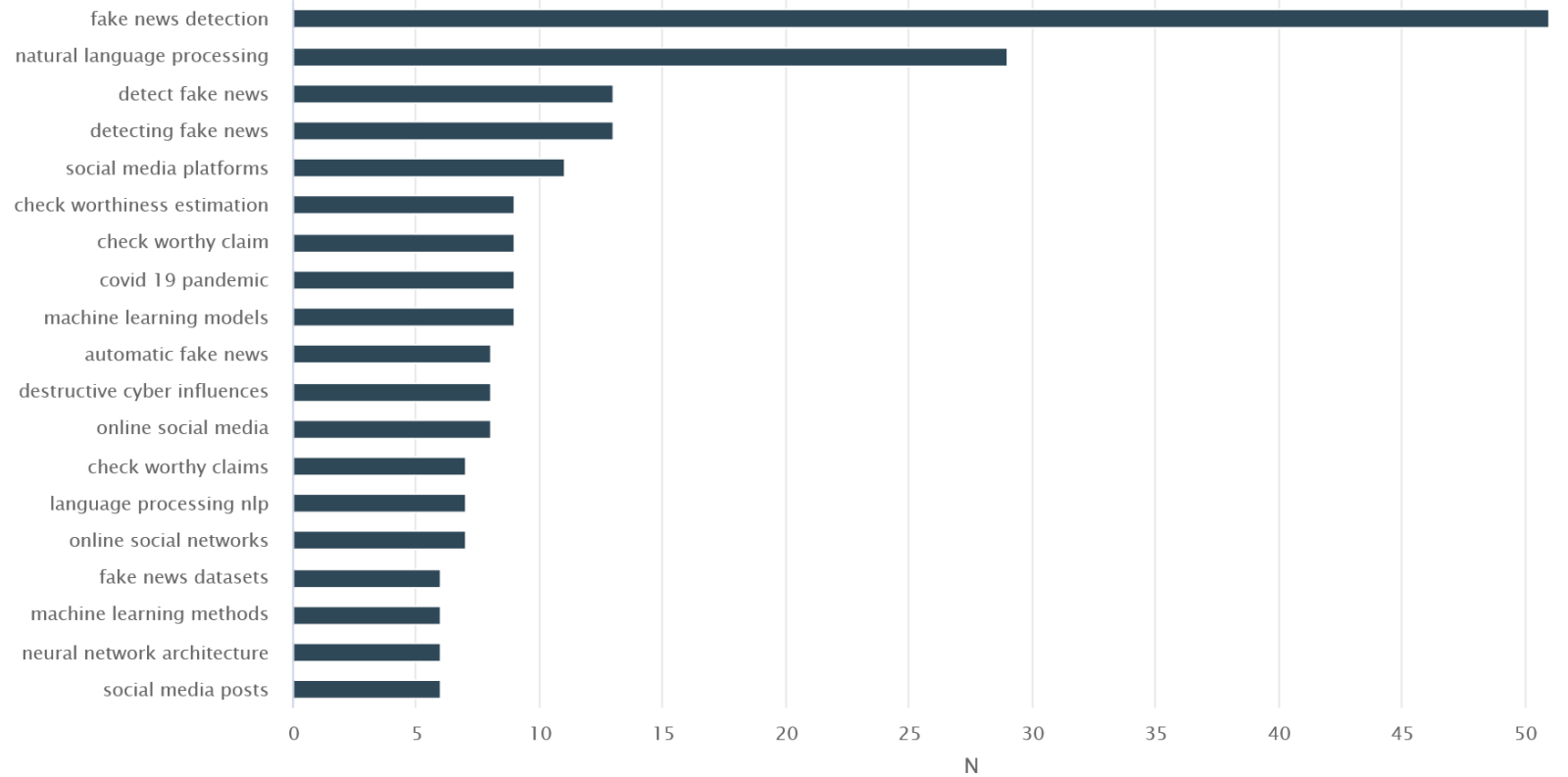
Fréquence des termes

feature	frequency	rank	docfreq	group
news	552	1	150	all
fact-checking	364	2	169	all
fake	361	3	111	all
information	233	4	123	all
claims	230	5	103	all
detection	193	6	90	all
media	178	7	102	all
social	166	8	94	all
claim	158	9	78	all
model	153	10	90	all
task	142	11	80	all
research	139	12	91	all
automated	138	13	106	all
data	135	14	83	all
models	126	15	72	all
paper	123	16	111	all
based	118	17	92	all
content	114	18	61	all
misinformation	114	18	65	all
dataset	114	18	63	all
false	112	21	67	all
learning	109	22	77	all

Bigrammes

	N
fake news	
social media	
machine learning	
news detection	
natural language	
check worthiness	
false information	
covid 19	
language processing	
real world	
social networks	
check worthy	
deep learning	
knowledge base	
neural network	

Trigrammes

NORDIS

Word Cloud sub-corpus 1

NORDIS

Word Cloud sub-corpus 2

Co-occurences

R Packages

The application of the exclusion and exclusion criteria was tackled through two complementary methods: close reading of all of the abstracts and the full texts of half of the collected corpus and distant reading of the abstracts through text mining and text analysis techniques using the programming language R (Ramage et al., 2009; Silge and Robinson, 2016; Welbers et al., 2017) and dedicated packages to proceed a meta-analysis and n-grams frequencies (tidyverse, tidytext, TM, quanteda, highcharter), topic modelling (LDA, formattable) and clustering (textmineR).

The R packages previously mentioned were also used for the meta-analysis of the main corpus. The meta-analysis combined a deductive and an inductive approach (Molina and Garip, 2019; Grimmer and al., 2021) in order to support discoveries considering research questions that globally refer to the challenges of automated fact-checking in the social world of journalism.

Iramuteq, une alternative

Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires (logiciel libre)

Implique un pre-processing adapté au logiciel
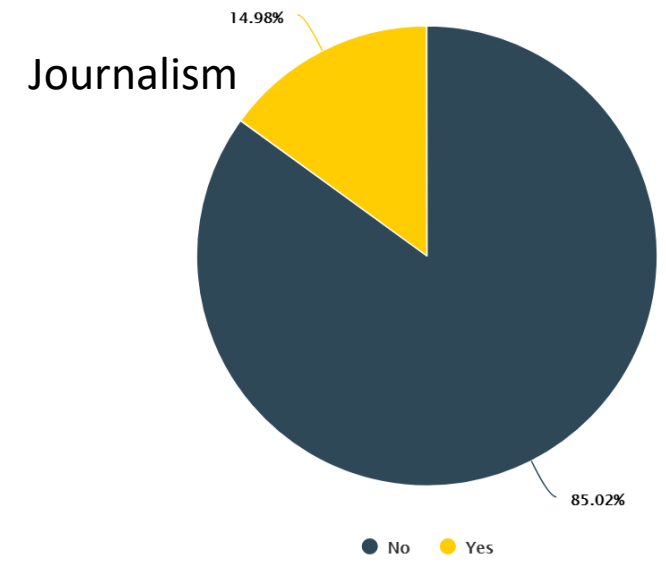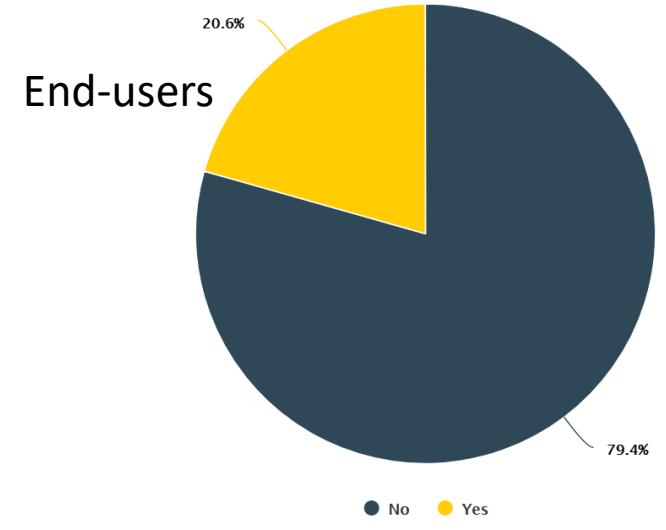
http://www.iramuteq.org/

Traitement quantitatif

- Opérations statistiques
- Par année
- Par discipline
- Par type de texte
- Par finalité
(utilisateurs finaux / journalisme)

Meta-analysis of the main corpus
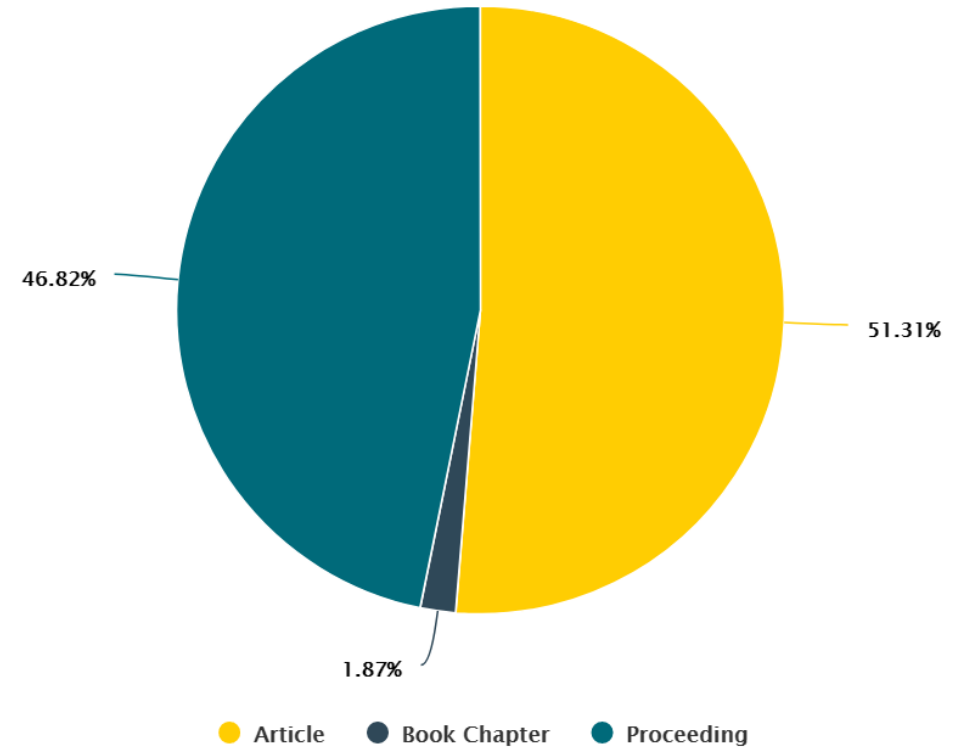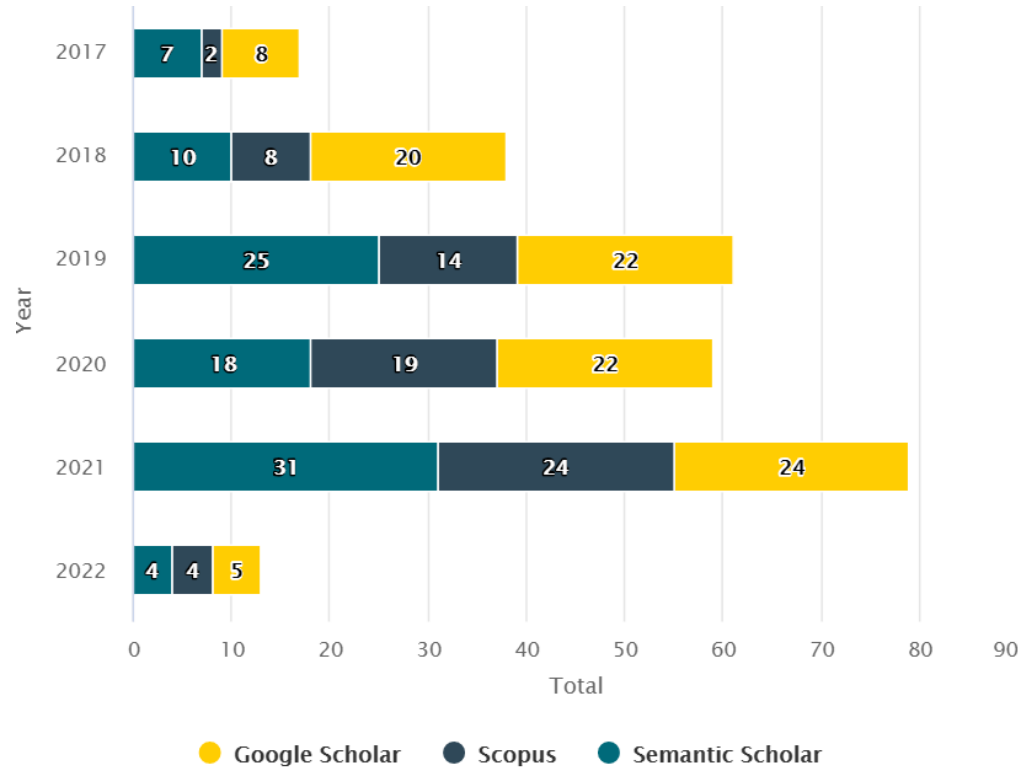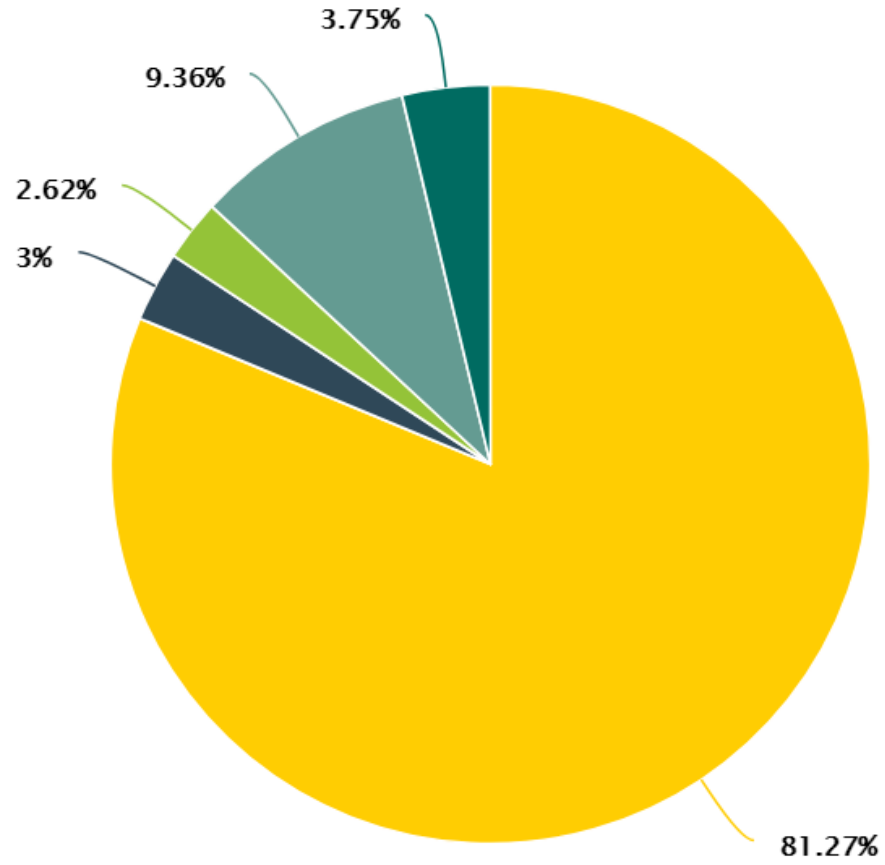
Algorithmes (non-supervisé)

LDA
Latent Dirichlet Allocator, modèle statistique, « bags of words » (ordre des mots pas important)
https://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd

CTM
Correlated Topic Models (hiérarchique)
https://www.datasciencecentral.com/topic-modeling-algorithms-techniques-and-application/

Hclust (R)
Hierarchical clustering
Ward: méthode de la variance minimale

Kmeans clustering
Quantification vectorielle, partitionne n observations en k clusters dans lesquels chaque observation appartient au cluster de moyenne la plus proche

```r
#CTM

K = 6
topicModelctm <- CTM(DTM, K, method = "VEM")
tmResult <- posterior(topicModelctm)
attributes(tmResult)


beta <- tmResult$terms
dim(beta)


rowSums(beta)


nDocs(DTM)


theta <- tmResult$topics
dim(theta)


rowSums(theta)[1:10]


top_terms <- terms(topicModelctm, 20)


top_terms <- as.data.frame(top_terms)


formattable(top_terms)
```
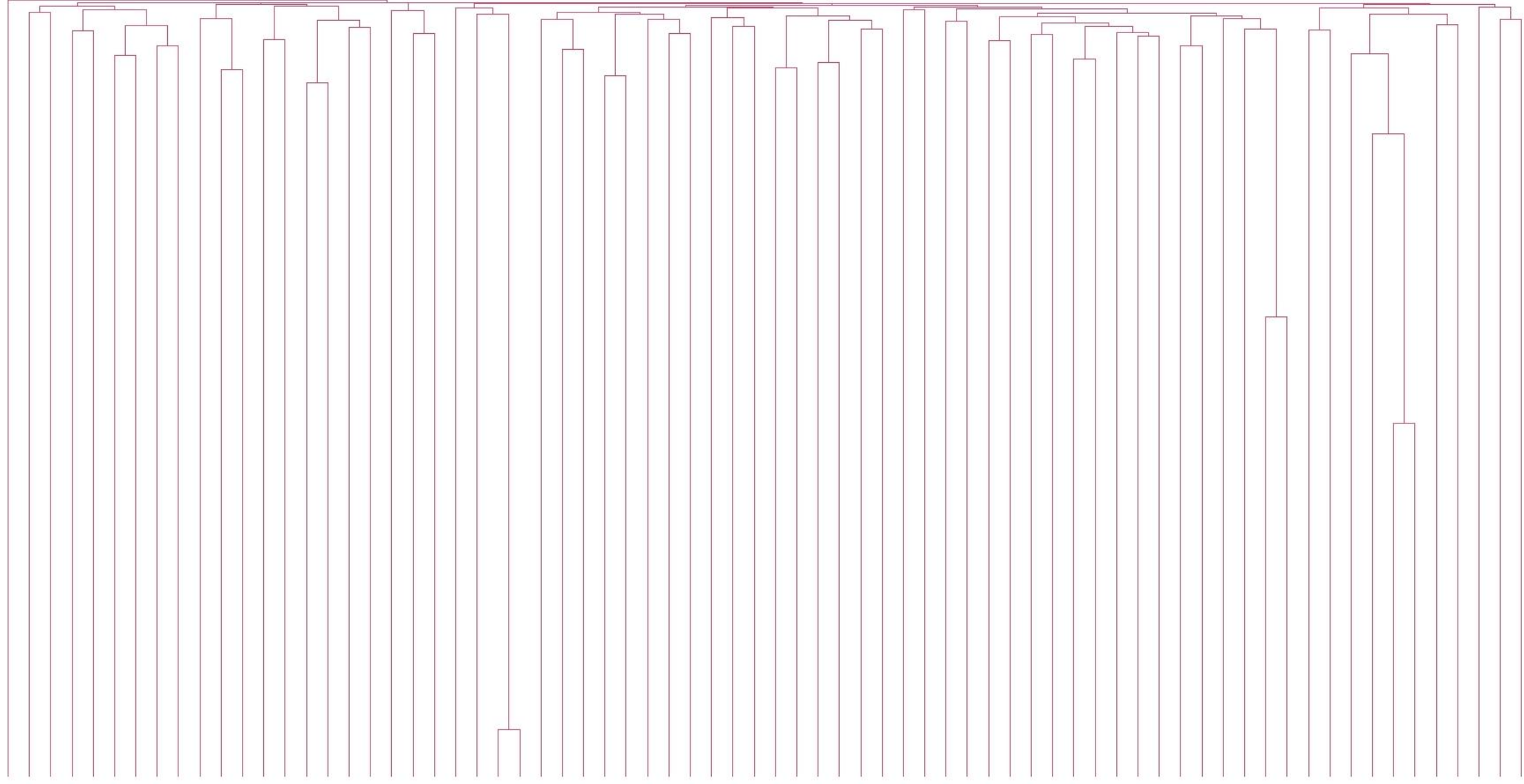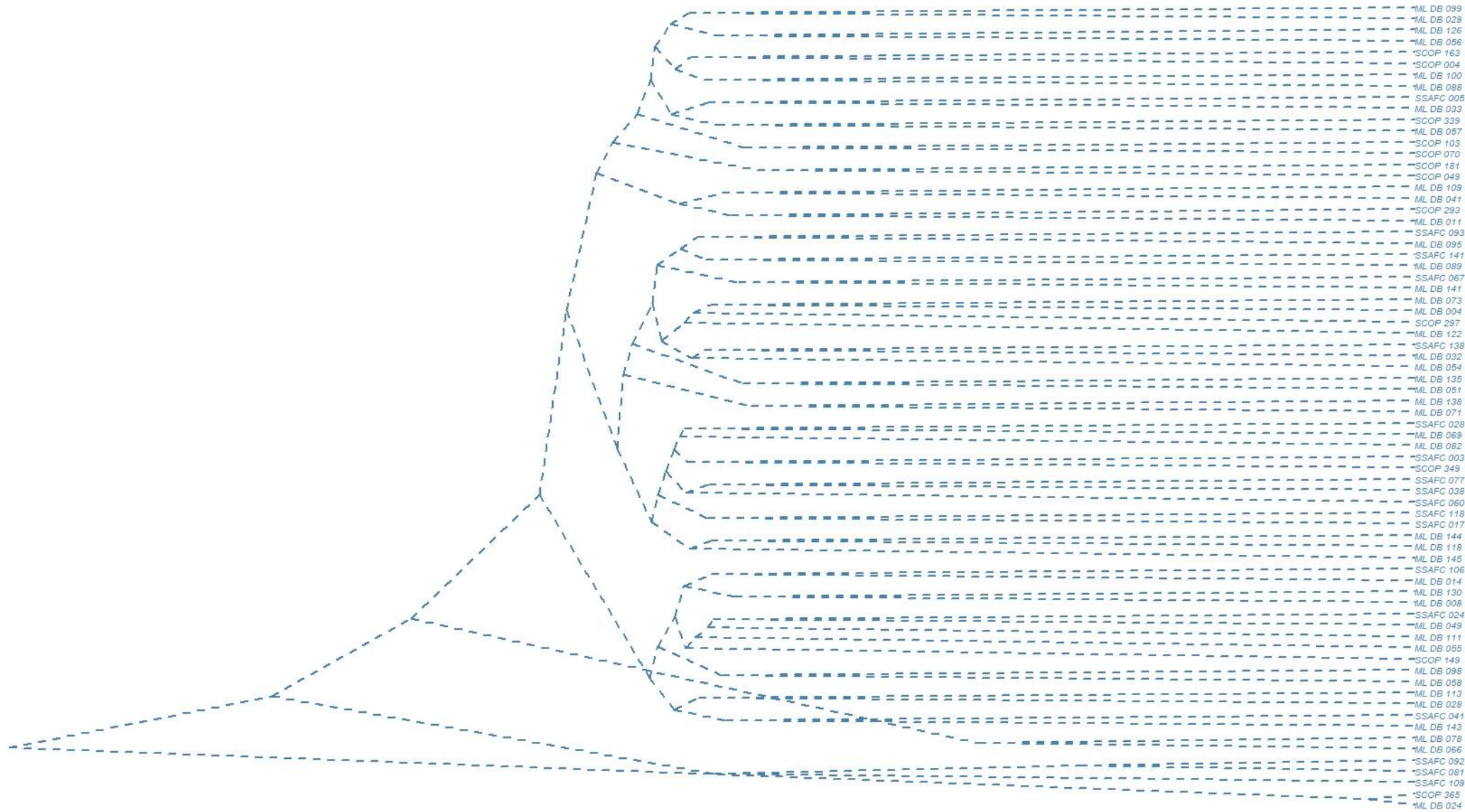
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
model	inform	research	fact-check	task	base	news	social	learn	system
dataset	data	tool	claim	use	use	fake	media	propos	user
perform	approach	public	fals	polit	fact	detect	content	method	autom
articl	process	journalist	evid	evalu	knowledg	featur	misinform	differ	studi
paper	languag	algorithm	retriev	claim	set	type	onlin	classif	predict
train	identifi	develop	autom	check-worthi	web	combin	spread	text	paper
automat	natur	attent	verac	present	verifi	present	disinform	machin	experi
annot	exist	discuss	generat	focus	provid	stori	tweet	problem	work
challeng	becom	new	signific	particip	check	peopl	credibl	use	human
label	sourc	solut	manual	debat	valid	import	sourc	classifi	assess
languag	network	comput	relev	best	graph	various	need	extract	explor
stanc	problem	imag	explan	factual	develop	issu	analysi	avail	time
provid	relat	increas	question	sentenc	framework	network	topic	appli	import
featur	creat	particular	true	lab	pattern	real	digit	neural	investig
stateart	architectur	journal	contain	verif	requir	spread	societi	share	help
new	influenc	futur	answer	score	search	social	sever	represent	accuraci
achiev	techniqu	area	structur	research	demonstr	articl	decept	approach	report
linguist	work	verif	explain	english	qualiti	analyz	twitter	case	find
corpus	produc	aim	promis	speech	support	dissemin	effect	deep	conduct
domain	address	concern	semant	rank	manual	world	covid-	compar	reli

| Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
fact	news	media	claim	tool	fact-check
knowledg	fake	social	task	research	autom
base	detect	content	model	journalist	inform
graph	featur	inform	fact-check	use	process
use	model	onlin	use	news	articl
entiti	dataset	spread	evid	develop	manual
method	use	disinform	system	user	method
fact-check	learn	credibl	dataset	algorithm	data
valid	classif	tweet	retriev	detect	natur
pattern	machin	topic	check-worthi	onlin	sourc
explan	differ	user	evalu	analysi	propos
generat	propos	misinform	predict	imag	languag
set	research	model	research	platform	challeng
search	approach	approach	improv	verif	misinform
check	languag	sourc	annot	public	approach
qualiti	data	societi	automat	fact-check	fals
framework	import	identifi	debat	process	human
present	paper	twitter	text	journal	work
user	present	mislead	lab	futur	problem
perform	text	dissemin	train	time	assess

Cluster plot

Limits

The limits of the research strategy are related to the level of accuracy provided by distant reading. Indeed, it is recognised that topic modelling is not suitable for advanced data relationships and performs poorly when documents do not have a sufficient length (Vayansky and Kumar, 2020).

Clustering is also challenging in finding the similarity between data points and grouping similar ones into the same cluster (Qaddoura et al., 2018).

Therefore, these results were primarily used to support human analysis.
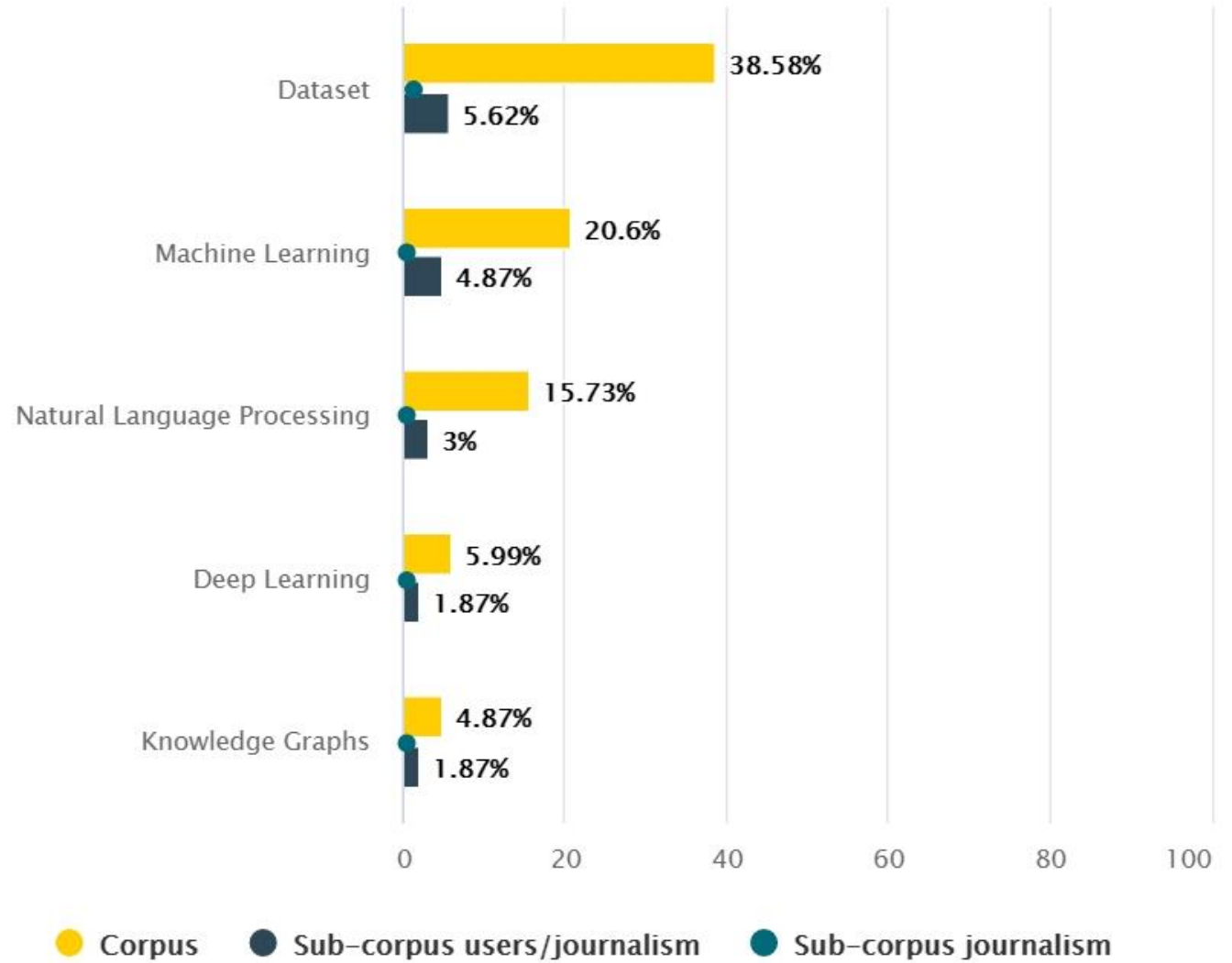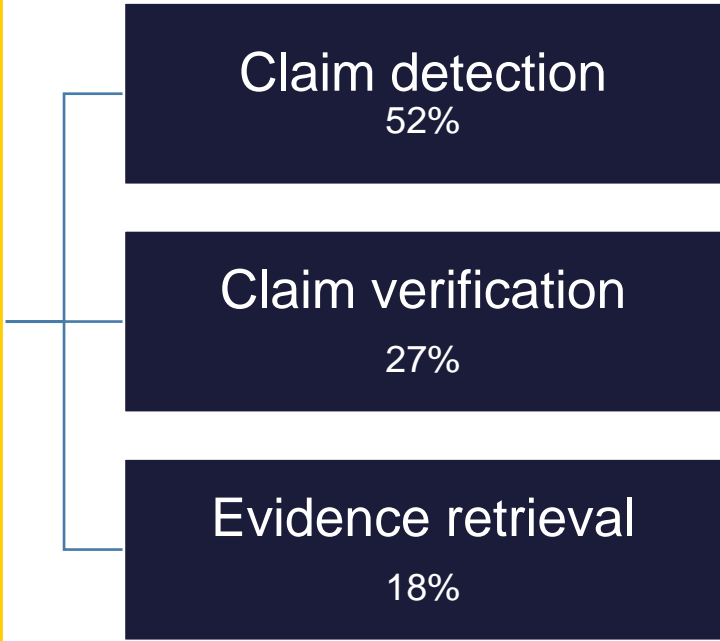
Datasets and characteristics

LIAR	Based on one decade of short statements collected from the U.S. website Politifact.com.
FEVER	It consists of 185,445 claims generated by altering sentences extracted from Wikipedia.
FAKES	Dataset around the Syrian war, relying on a semi-supervised ML, 804 news labelled "true" or "fake".
PUBHEALTH	Based on 11.8K claims to support fact-check labels for claims.
ClaimBuster	Dataset of 23,533 statements from all U.S. general election presidential debates, annotated by humans.
FakeCovid	Multilingual dataset of 7,623 fact-checked news articles for COVID-19, collected from 04/01 to 01/07/2020.
WikiFactCheck English	Dataset of 124k+ triples consisting of a claim, context and an evidence document extracted from English Wikipedia articles and citations, and 34k+ manually written claims refuted by the evidence documents.
Sentimental LIAR	Extends the LIAR dataset of short claims by adding features based on sentiment and emotion analysis of claims.
AraStance	Multi-country and multi-domain dataset of Arabic stance detection for fact-checking, based on 4,063 claim–article pairs from diverse sources comprising three fact-checking websites and one news website.
ES-Contradiction	It intends to fill the lack of automated contradiction detection systems for the Spanish language and contains examples with two pieces of information classified as Compatible, Contradiction, or Unrelated.
CsFEVER CTKFacts	A derivative of the FEVER dataset in the Czech language and a dataset containing 3,097 claims from a corpus of approximately two million Czech News Agency news reports.

AFC is more about textual than visual content

Multimodal models trained on both texts and images (4)

Datasets collection (2)

Supervised ML to detect deceptive images (1)

Supervised ML for image classification (1)

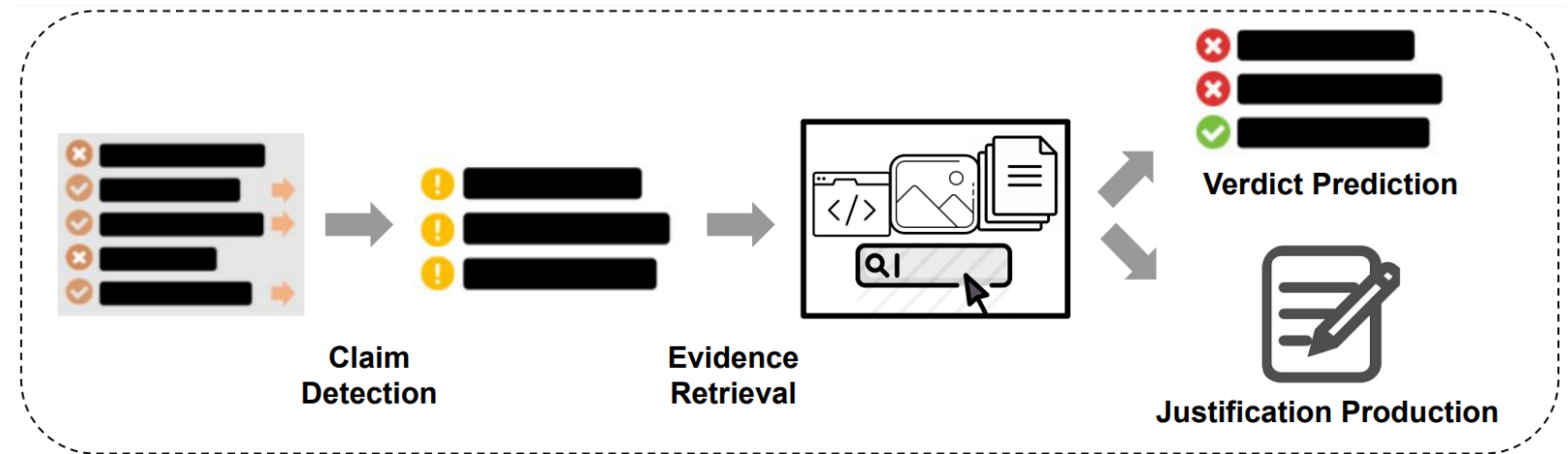Assessment of image forensics services (1)

Social-computing solutions for user-generated content verification (1)

Deepfake detection (1)

Identified user needs

Algorithmic tools specifically designed for the journalistic research process are **rarely or not at all used**.
Journalists are **not aware of hidden research assistant** facilitating their research process.
(de Haan et al, 2022 – The Netherlands)

Trust

Reliability

Accuracy

Relevance

Claim Detection

Evidence Retrieval

Verdict Prediction

Justification Production

Source: https://github.com/Cartus/Automated-Fact-Checking-Resources

Critical issues for using AFC in journalism

- Datasets quality

  Crowdsourcing, Wikipedia, adaptability

- Model performances

  Evaluation/accuracy

- Human-in-the-loop

  Shared expertise, engagement

Transparence de la méthode

Pas seulement dans l'explication de la sélection du corpus, de l'application des critères d'exclusion, de l'application du modèle de rédaction PRISMA

Mise à disposition du corpus, des scripts R développés et des visualisations dans un repo Github

= REPRODUCTIBILITE

laurence001 Create subcorpus2.csv · 079af38 on 12 Jul · 44 commits

R	Create theme.R	2 months ago
data	Create subcorpus2.csv	2 months ago
images	Add files via upload	2 months ago
LICENSE	Initial commit	2 months ago
README.md	Update README.md	2 months ago

README.md

Automated fact-checking to support professional practices: systematic literature review and meta-analysis

Datasets: main corpus, sub-corpus 1, sub-corpus 2, stopwords

Source code: R scripts for analysis

Images: result data visualizations

About

Automated fact-checking to support professional practices: systematic literature review and meta-analysis (data and source code)

- Readme
- MIT license
- 0 stars
- 1 watching
- 0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- HTML 99.8%
- R 0.2%

Quelques conseils

Prendre le temps :

- de travailler sa question de recherche (PICOC)

- pour optimiser la qualité de la collecte

- bien réfléchir à la stratégie d'annotations

- bien connaître les possibilités et limites de l'algorithme utilisé si utilisation du ML

- bien connaître les packages que l'on utilise en ML (R ou Python, affordance du langage)

- relire les articles en cas de doute à propos des résultats automatisés (en particulier non supervisés)

Fully automated fact-checking?

The idea of "a completely automated fact-checking platform that can detect a claim as it appears in real-time, and instantly provide the voter with a rating about its **accuracy**" seems to remain **challenging**, despite some advances in the domain.
(Adair et al., 2017)

A fully automated tool that judges a claim to be true or false is always **limited** in **functionality**, **accuracy** and **understandability**.
(Masood & Aker, 2018)

Evaluating the **authenticity** of news remains very **complex**, even for automated systems.
(Borges et al., 2019)

"Automated fact-checking works well in some cases", but its generalization "still needs **improvement** prior to widespread **use**".
(Lazarski et al., 2021)

Merci pour votre attention !

Twitter
@ohmyshambles