University of Sheffield

When the Devil is in the Sample

Benchmarking Data Sources to Analyse African News Content Computationally

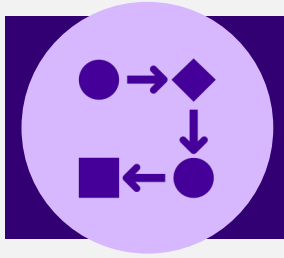
Dani Madrid-Morales

18 June 2024

Overview

Acquiring (news) text data at scale – Challenges of using existing approaches to study (most) Global South countries

Quantifying the variance in data sources – An exploration of whether using different text sources impacts research findings

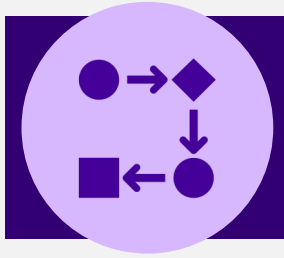


Proposing new workflows for data acquisition – Presentation of two news corpora built to be more comprehensive

Overview

Acquiring (news) text data at scale – Challenges of using existing approaches to study (most) Global South countries

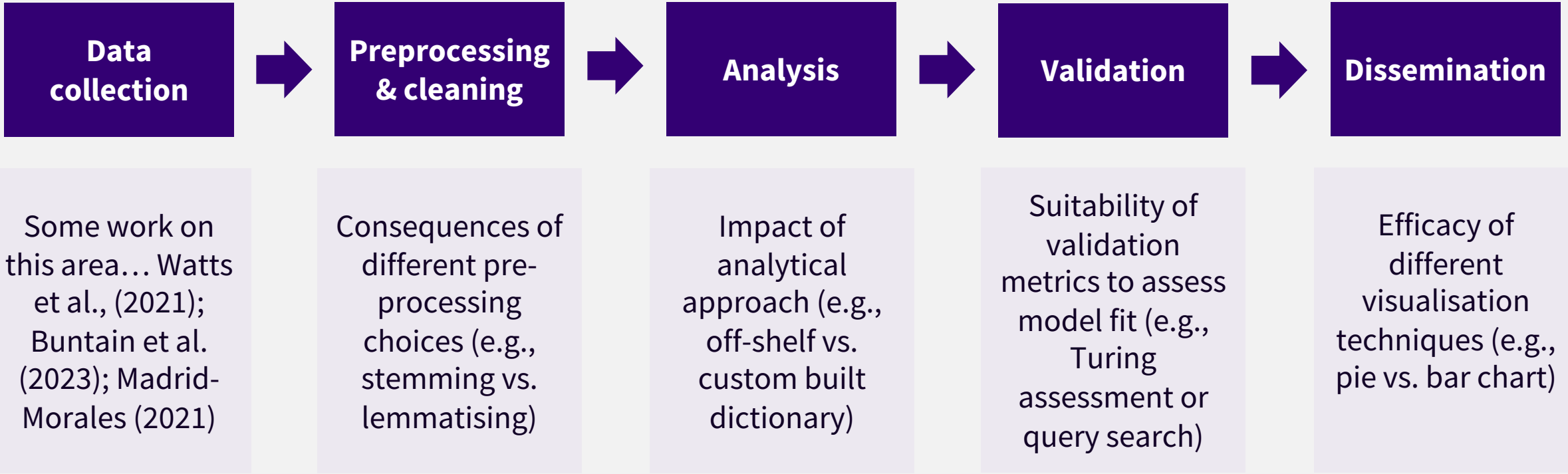
Quantifying the variance in data sources – An exploration of whether using different text sources impacts research findings



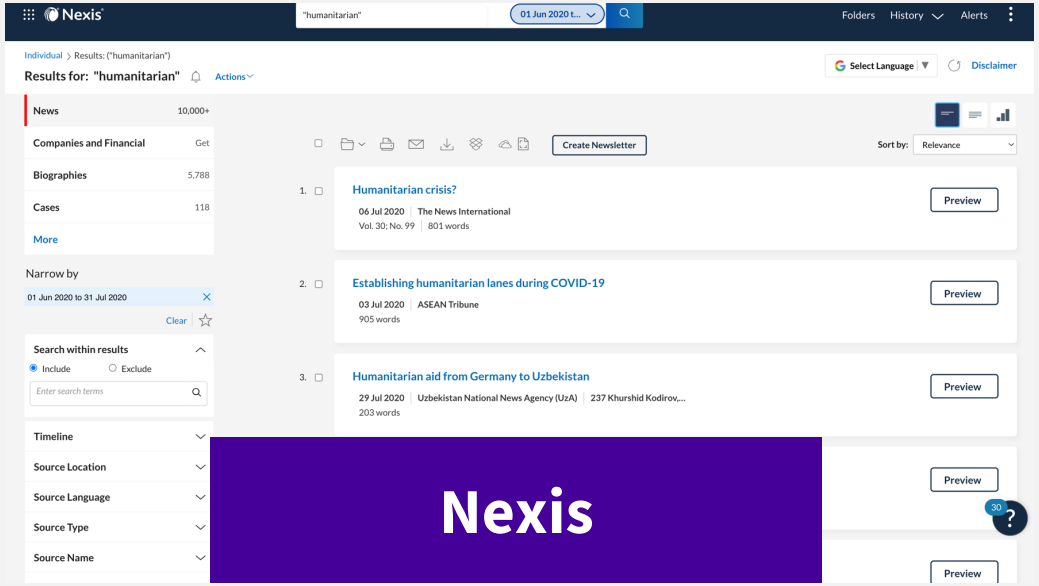
Proposing new workflows for data acquisition – Presentation of two news corpora built to be more comprehensive

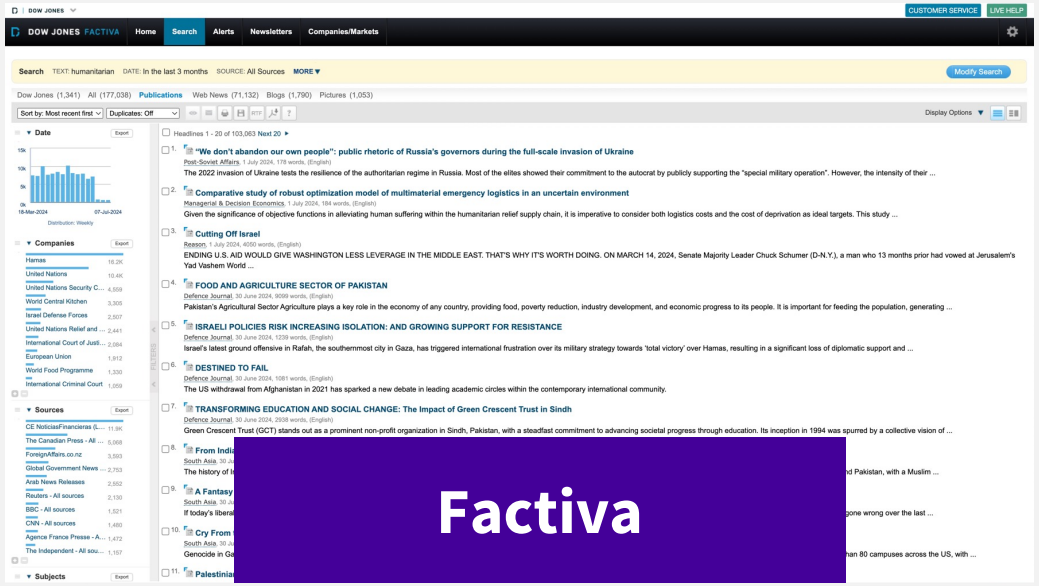
Background & Rationale

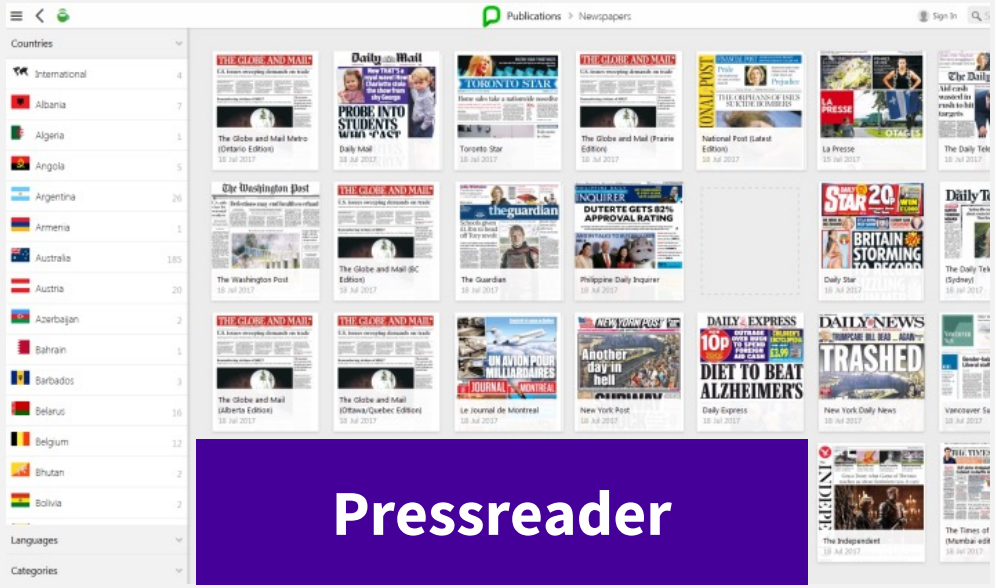
- When considering the different stages in the lifecycle of a computational text analysis project (data collection, cleaning and processing, analysis, validation, and dissemination), most attention has gone into examining the latter four stages.

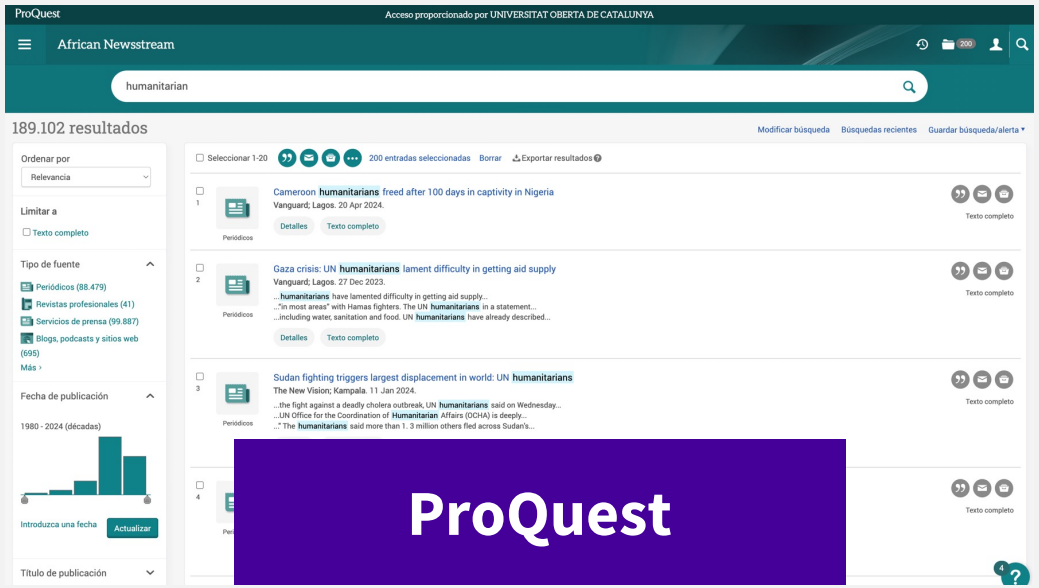


Data collection approaches I – Full text news databases









Potential shortcomings of these approaches – Limited representation

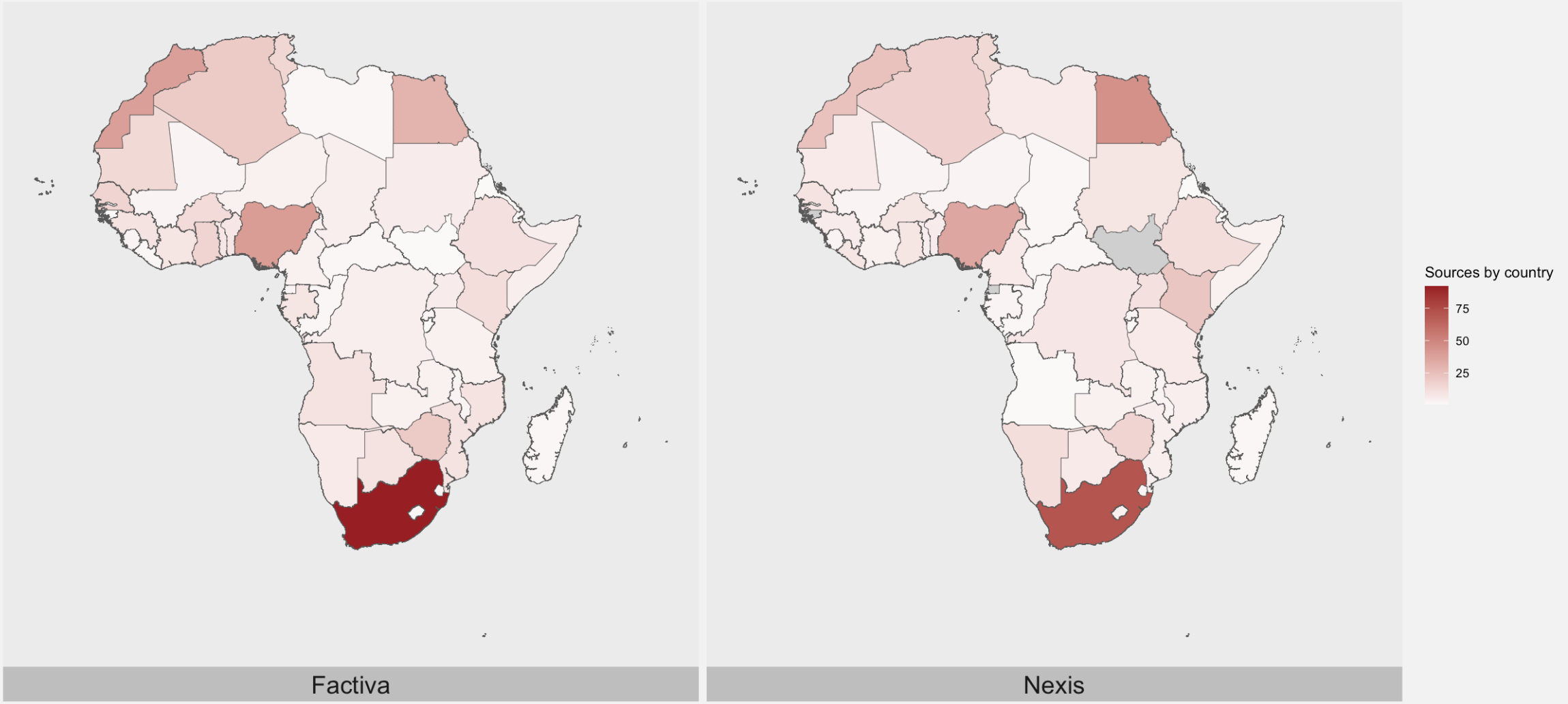
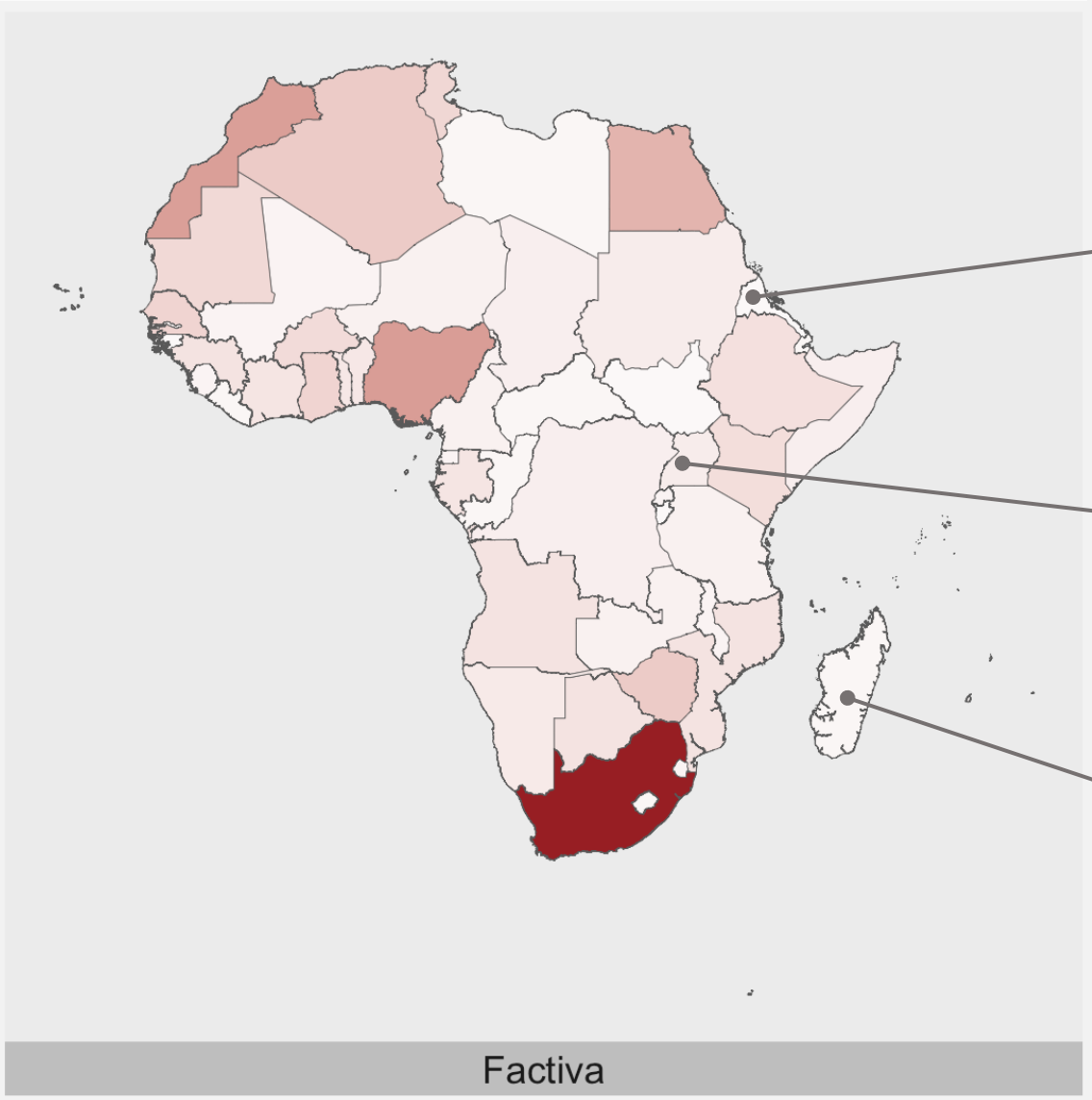


Figure 1. Number of available news sources per country

Potential shortcomings of these approaches – Ideological biases

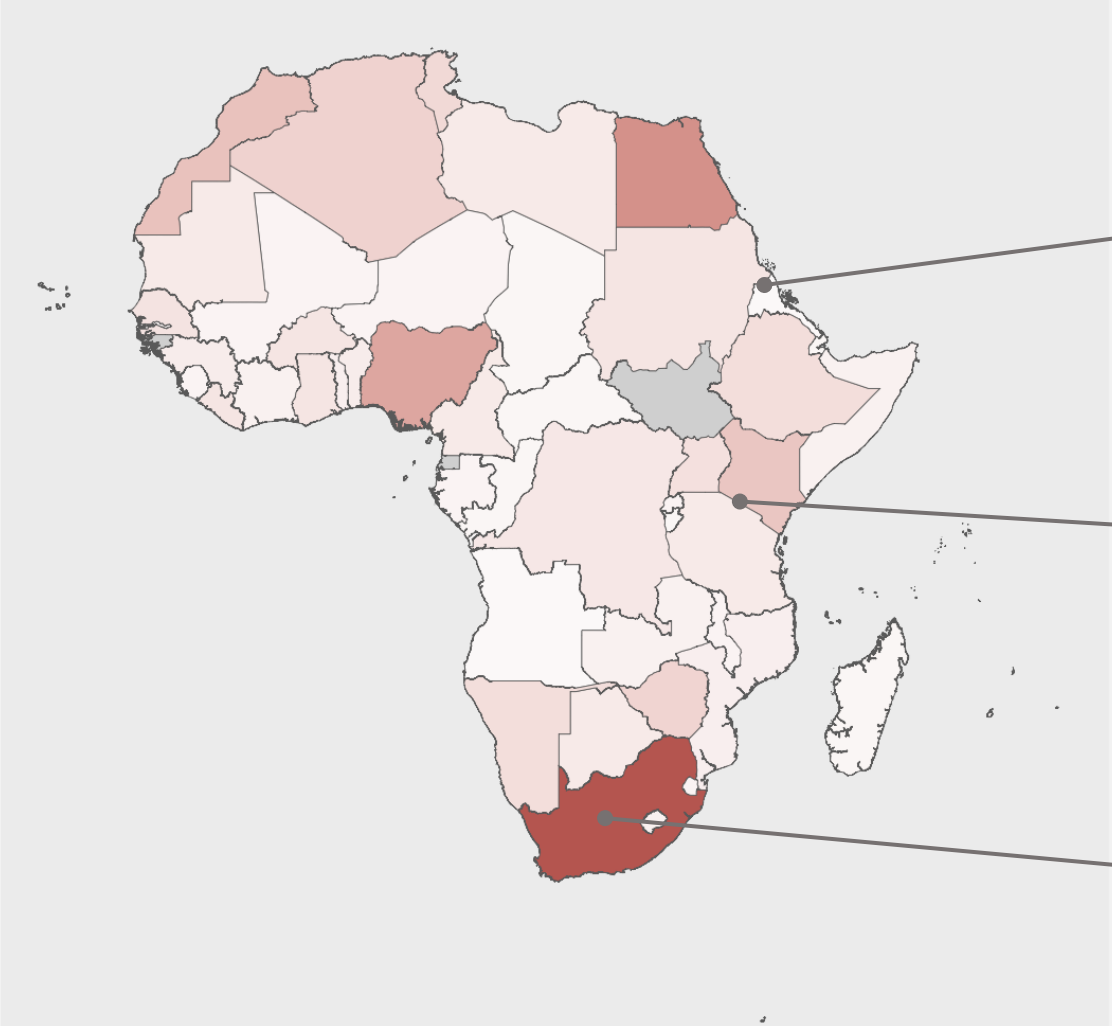


The only available source (Shabait) is actually the website of the Ministry of Information

The only available sources (*Le Renouveau*) are "government owned and controlled"

The only available source (*L'express*) was bought by an opposition politician and businessman

Potential shortcomings of these approaches – No linguistic diversity



Nexis

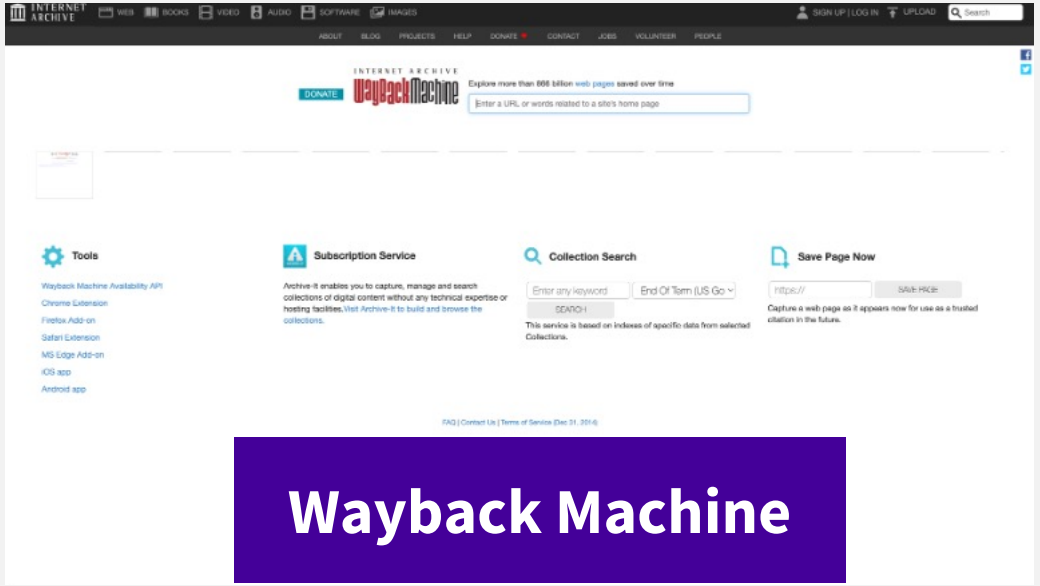
Out of 10+ Ethiopian sources available on Nexis, none are in Amharic

Out of 30 Kenyan and Tanzanian sources available on Nexis, none are in Swahili

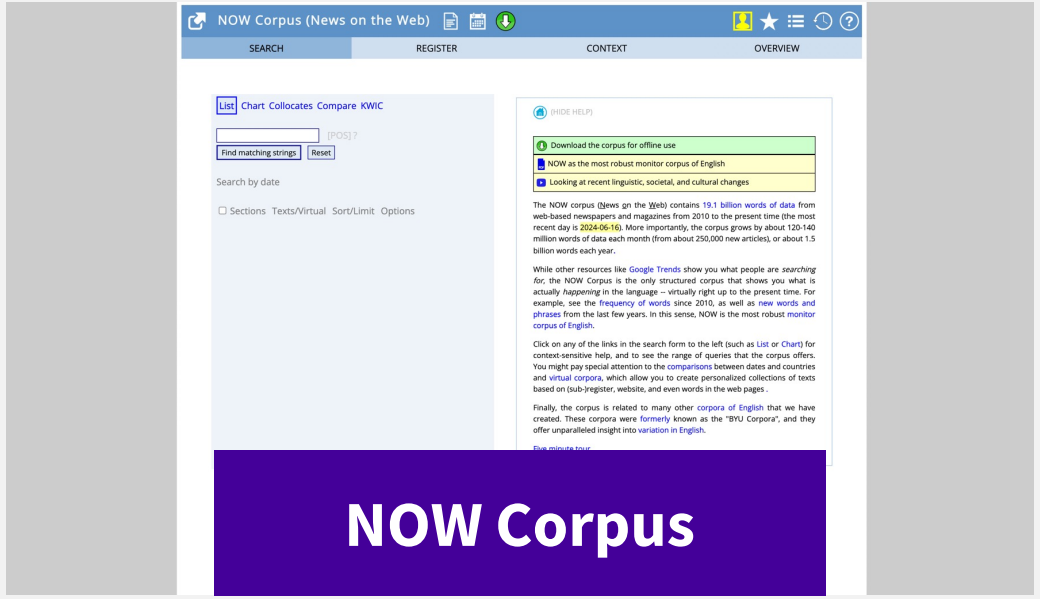
Out of 70+ South African sources available on Nexis, none are in isiZulu or isiXhosa

Data collection approaches II – Large (text) collections

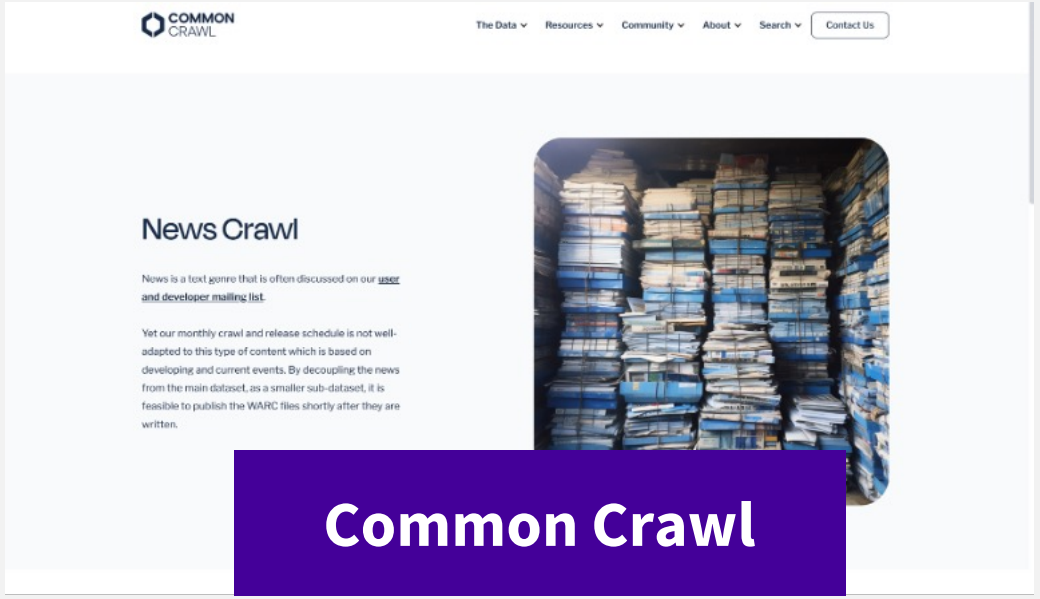
GDELT



Wayback Machine



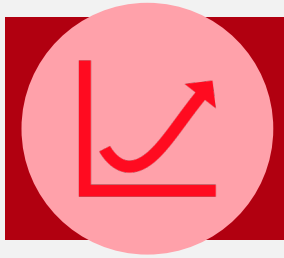
NOW Corpus



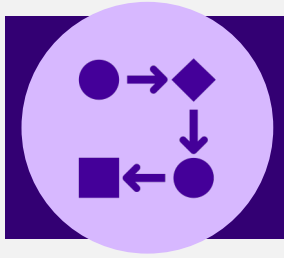
Common Crawl

Overview

Acquiring (news) text data at scale – Challenges of using existing approaches to study (most) Global South countries



Quantifying the variance in data sources – An exploration of whether using different text sources impacts research findings



Proposing new workflows for data acquisition – Presentation of two news corpora built to be more comprehensive

Two benchmarking exercises

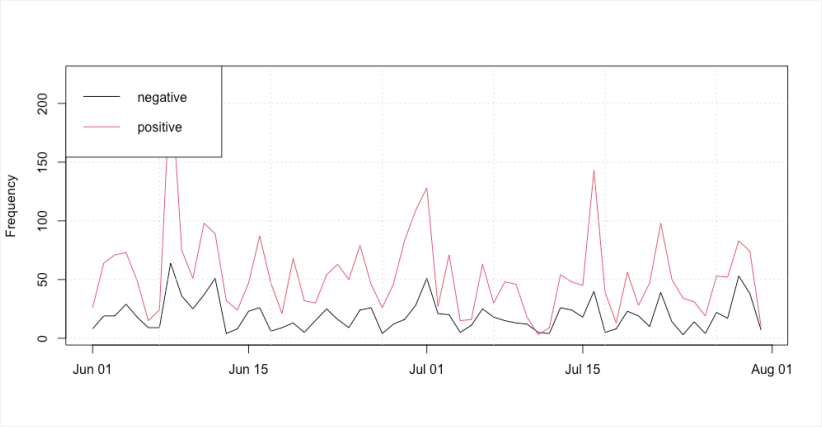
Humanitarian news in English

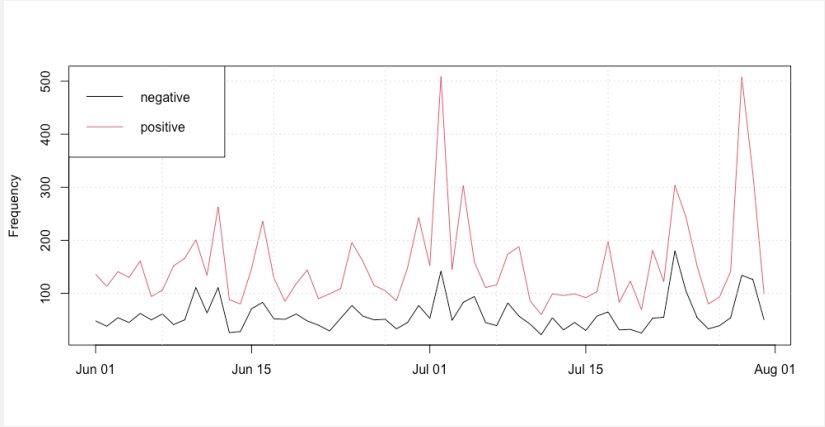
- Compare Factiva & Nexis, to a custom-built multi-source corpus
- From June 1, 2020 to July 31, 2020
  - Factiva: n = 2,092
  - Nexis: n = 1,707
  - Corpus: n = 4,667
- Comparing sentiment around keyword “humanitarian”

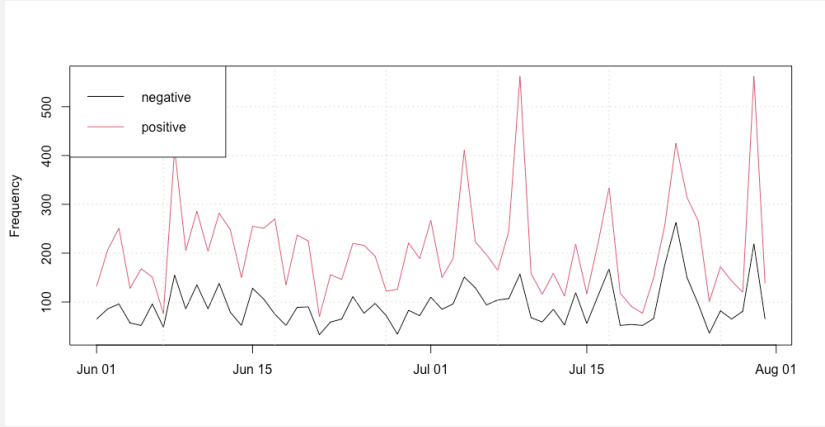
Content from African news websites

- Compare GDELT & Wayback Machine, to a custom-built corpus
- From June 1, 2021 to July 31, 2021
  - GDELT: n = 15,168
  - WBM: n = 121,995
  - Corpus: n = 254,450
- Overall topic diversity using LexiCoder Policy Agendas.

Benchmarking “humanitarian news” on Factiva/Nexis I





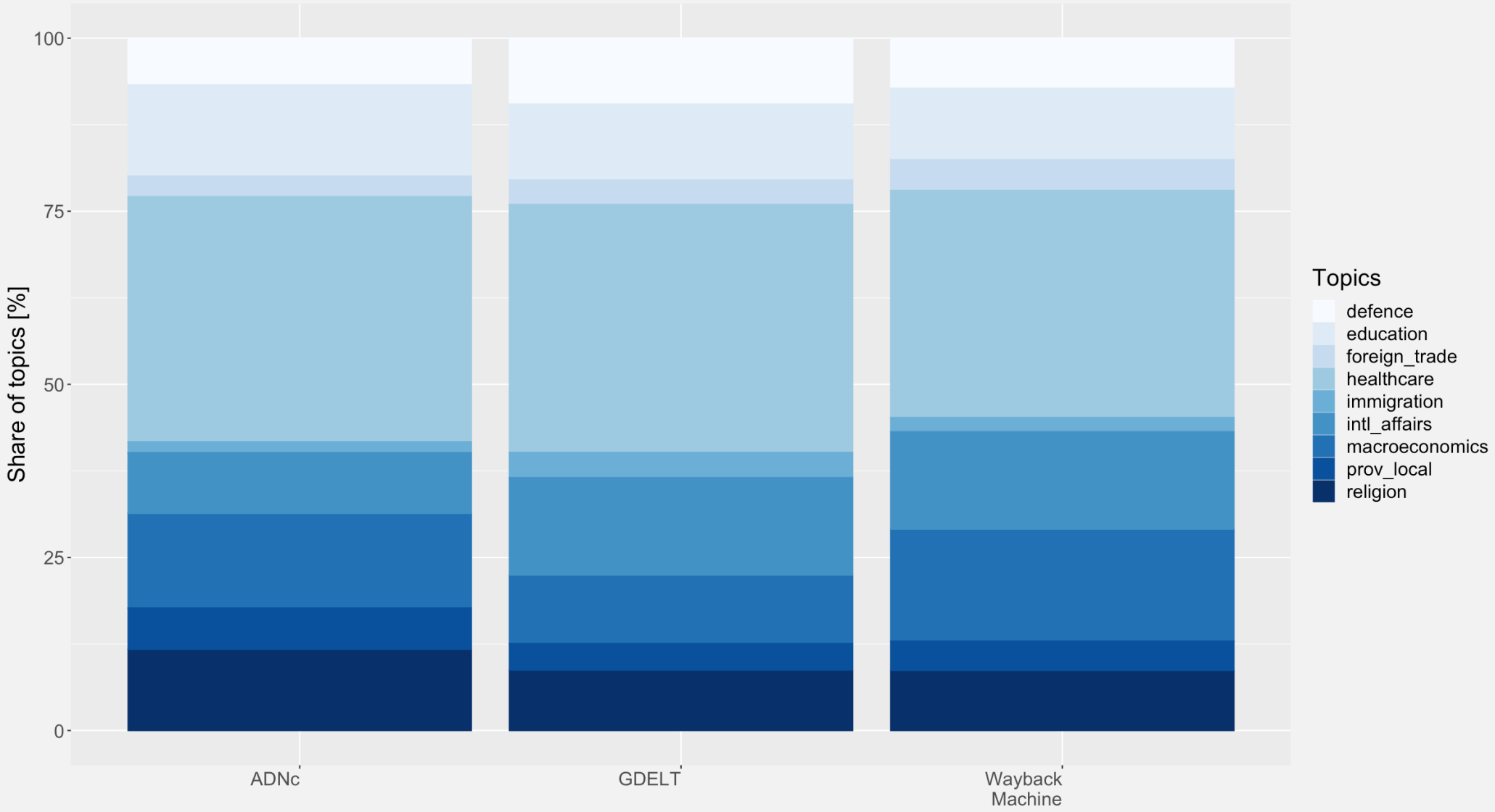


Nexis

Factiva

Custom Corpus

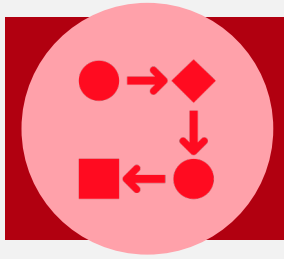
Benchmarking GDELT/Wayback Machine I



Overview

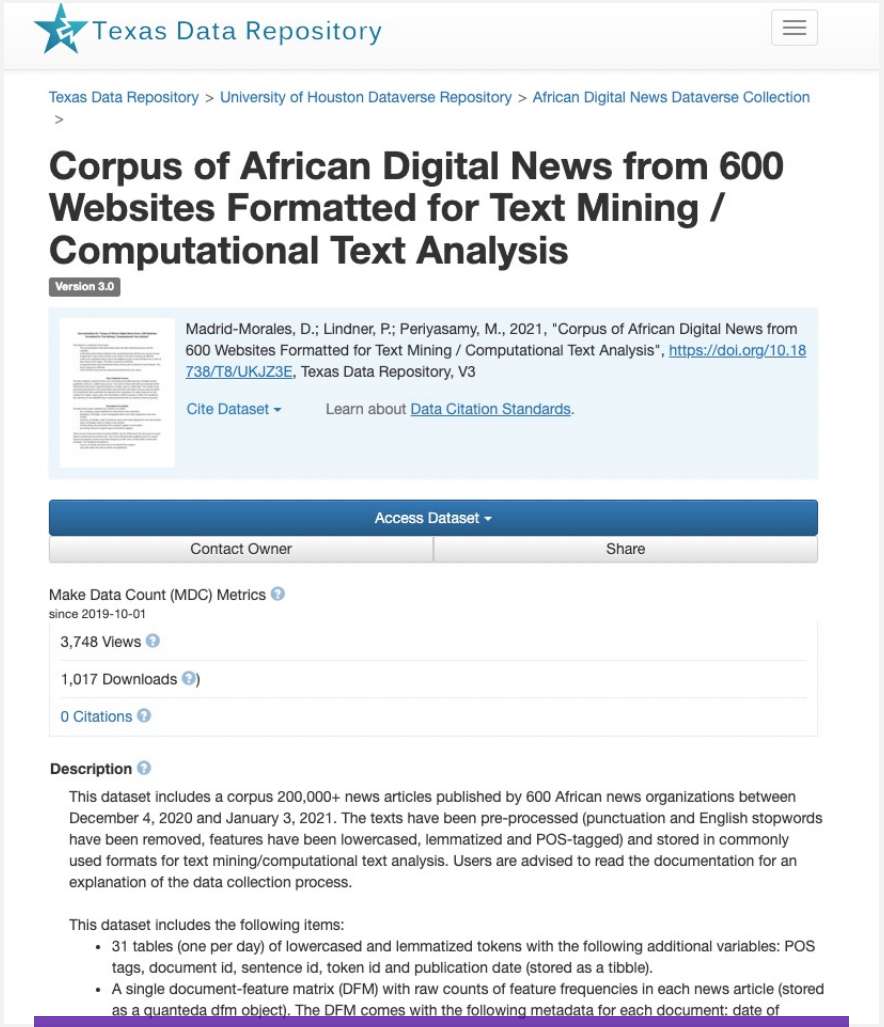
Acquiring (news) text data at scale – Challenges of using existing approaches to study (most) Global South countries

Quantifying the variance in data sources – An exploration of whether using different text sources impacts research findings

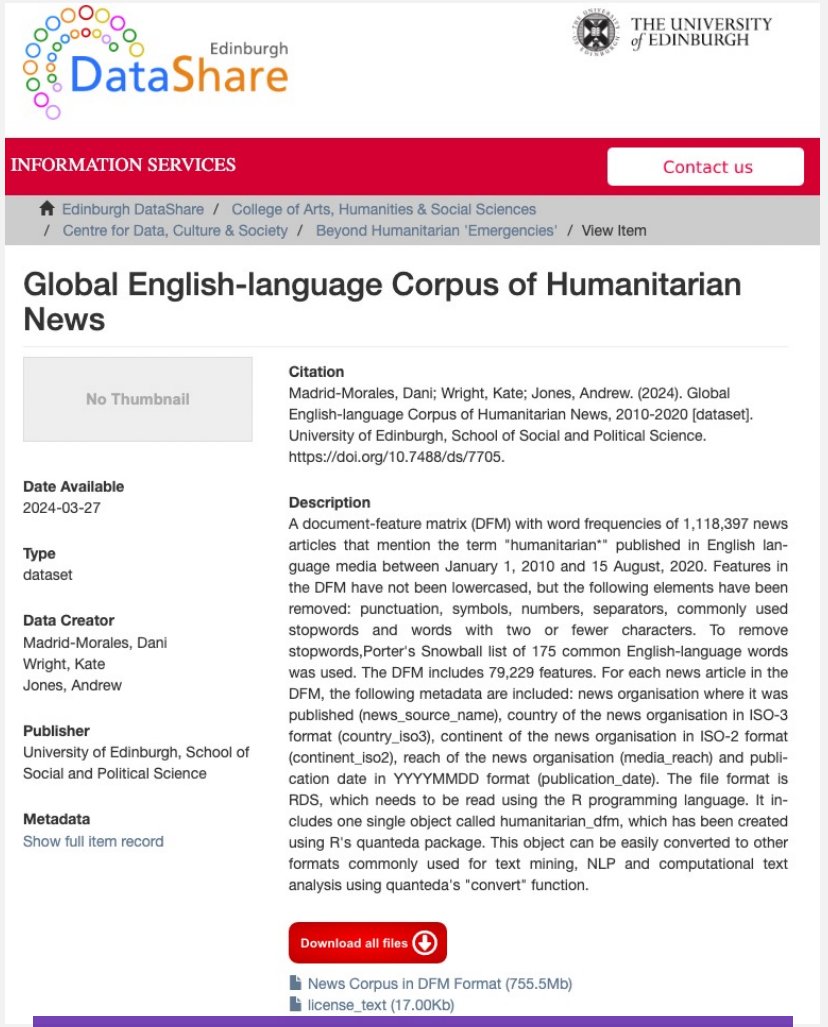


Proposing new workflows for data acquisition – Presentation of two news corpora built to be more comprehensive

Two corpora built with these challenges (?¿) in mind



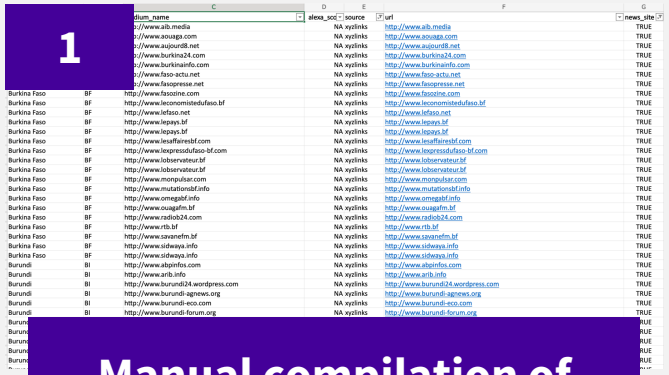
https://doi.org/10.18738/T8/UKJZ3E



https://doi.org/10.7488/ds/7705

Building African Digital News Corpus (ADNC)

1



Manual compilation of seed URLs to crawl

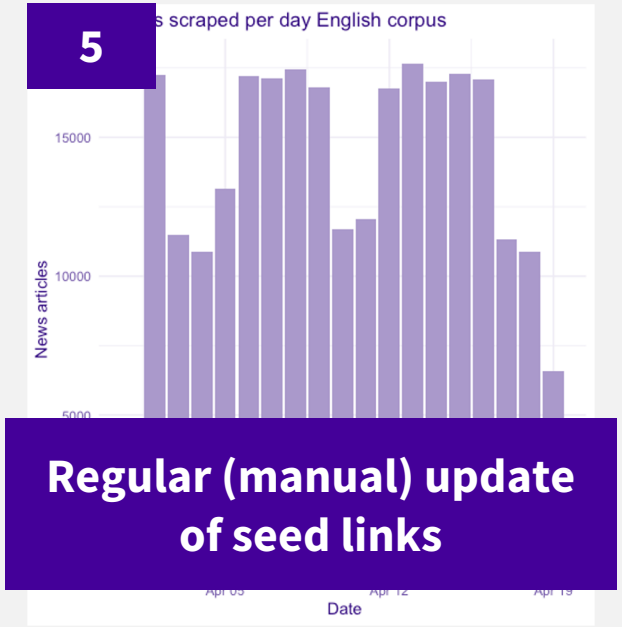
2

Twice daily crawl to collect all new URLs

3

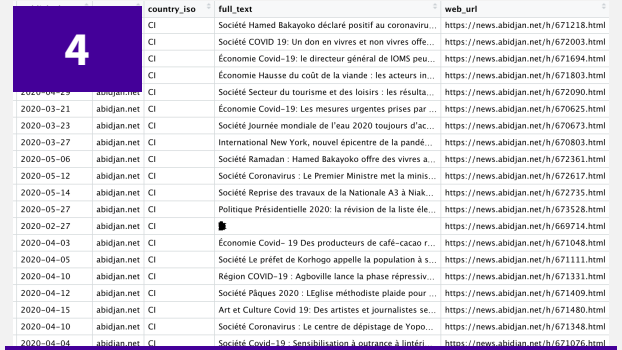
Automated scraping of news content of interest

5



Regular (manual) update of seed links

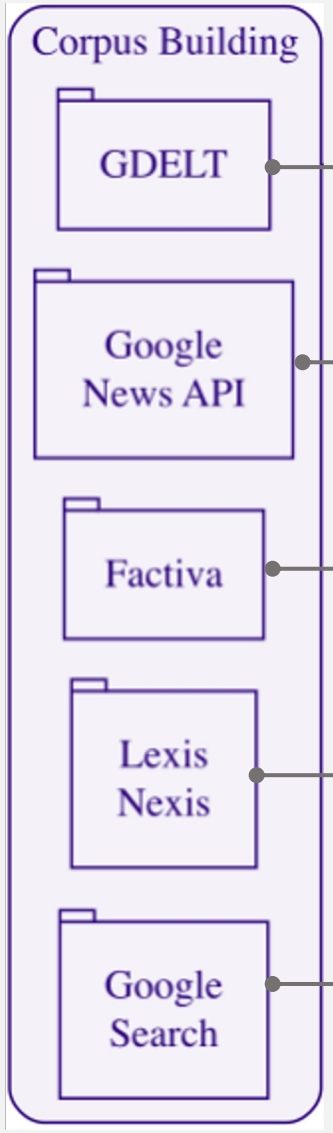
4



country_iso	full_text	web_url
CI	Société Hamed Bakayoko déclaré positif au coronaviru…	https://news.abidjan.net/h/671218.html
CI	Société COVID 19: Un don en vivres et non vivres offe…	https://news.abidjan.net/h/672003.html
CI	Économie Covid-19: le directeur général de IOMS peu…	https://news.abidjan.net/h/671694.html
CI	Économie Hausse du coût de la viande : les acteurs in…	https://news.abidjan.net/h/671803.html
CI	Société Secteur du tourisme et des loisirs : les résulta…	https://news.abidjan.net/h/672090.html
CI	Économie Covid-19: Les mesures urgentes prises par …	https://news.abidjan.net/h/670625.html
CI	Société Journée mondiale de l'eau 2020 toujours d'ac…	https://news.abidjan.net/h/670673.html
CI	International New York, nouvel épicentre de la pandé…	https://news.abidjan.net/h/670803.html
CI	Société Ramadan : Hamed Bakayoko offre des vivres a…	https://news.abidjan.net/h/672361.html
CI	Société Coronavirus : Le Premier Ministre met la minis…	https://news.abidjan.net/h/672617.html
CI	Société Reprise des travaux de la Nationale A3 à Niak…	https://news.abidjan.net/h/672735.html
CI	Politique Présidentielle 2020: la révision de la liste éle…	https://news.abidjan.net/h/673528.html
CI	[redacted]	https://news.abidjan.net/h/669714.html
CI	Économie Covid- 19 Des producteurs de café-cacao r…	https://news.abidjan.net/h/671048.html
CI	Société Le préfet de Korhogo appelle la population à s…	https://news.abidjan.net/h/671111.html
CI	Région COVID-19 : Agboville lance la phase répressiv…	https://news.abidjan.net/h/671331.html
CI	Société Pâques 2020 : LEglise méthodiste plaide pour …	https://news.abidjan.net/h/671409.html
CI	Art et Culture Covid 19: Des artistes et journalistes se…	https://news.abidjan.net/h/671480.html
CI	Société Coronavirus : Le centre de dépistage de Yopo…	https://news.abidjan.net/h/671348.html
CI	Société Covid-19 : Sensibilisation à outrance à Ijntéri…	https://news.abidjan.net/h/671076.html
abidjan.net		
abidjan.net		
abidjan.net		
abidjan.net		
abidjan.net		
abidjan.net		
abidjan.net		
abidjan.net		
abidjan.net		
abidjan.net		
abidjan.net		
abidjan.net		
abidjan.net		
abidjan.net		
abidjan.net		
abidjan.net		
abidjan.net		
abidjan.net		
abidjan.net		
abidjan.net		

Cleaning, indexing and storage in MongoDB

Building Humanitarian News Corpus



Querying database to retrieve links of known domains followed by re-hydration of URLS to scrape news content

Using a commercial service to query a time-limited (2 years) archive of Google News content

Manually downloading news that match search query for all African news sources

Manually downloading news that match search query for all African news sources

Automating time-limited Google Searches by domain to retrieve URLs of interest, followed by scraping content

University of Sheffield

When the Devil is in the Sample

Benchmarking Data Sources to Analyse African News Content Computationally

Dani Madrid-Morales

18 June 2024