



Science ouverte au Sud
gestion et l'ouverture des productions de la recherche :
Panorama et perspectives en Afrique



Vers une base de connaissances africaine des données journalistiques (articles et commentaires)

Expérimentation sur le projet Check4Decision

Dr Edouard Ngor SARR

Université Assane SECK de Ziguinchor-UASZ

Check4Decision Project

CERIDES UCAO

Du 25 au 27 Octobre 2022 à COTONOU -BENIN

Extraction & Agrégation & la réutilisation de données journalistiques issues de sources web ?

Projet de recherche scientifique basé à l'Université de THIES au Sénégal : <http://check4decision.univ-thies.sn/>

- Un consortium de 04 universités
- Un financement initial du CEA MITIC

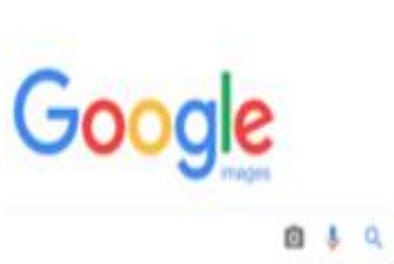
Equipe de Recherche et Partenaires



 <p>Prof. Ousmane SALL Enseignant-Chercheur</p> 	 <p>Dr. Babiga BIRREGAH Enseignant-Chercheur</p> 	 <p>Dr. Edouard Ngor SARR Enseignant-Chercheur</p> 	 <p>Prof. Mamadou Bousso Enseignant-Chercheur</p> 	 <p>Dr. Marie NDIAYE DIOP Enseignante-Chercheure</p> 			
							

Check4Decision Research Project

A l'origine (2017): Fact-checking



Youtube DataViewer



In video veritas



- Journal *Le Monde* en France (<https://www.lemonde.fr/verification/>)

- Du NFS aux Etats Unis
- Un modèle d'IA pour détecter automatiquement les affirmations à vérifier.
- <https://idir.uta.edu/claimbuster/>

InVid et CrossCheck lancés par le réseau de fact-checking First Draft News (<https://fr.firstdraftnews.com/>) et financés par Google News Lab. <http://invid.condat.de/>

	Mode d'utilisation	Type de données en entrée				
		Plateforme en ligne	Image	Audio/Vidéo	Email	Texte
VERA	Extension navigateur					
FactMinder		✓				✓
Claimbuster		✓				✓
Truth Teller		✓		✓		✓
MapChecking		✓	✓			
CrossCheck		✓				✓
Hoaxbuster		✓			✓	
Decodex	✓					✓
Check News		✓				✓
Google Images		✓		✓		
TinEye		✓		✓		
YouTube DataViewer		✓		✓		
FullFact	✓		✓	✓	✓	✓
LazyTrusth		✓			✓	
Holy Grail		✓		✓		✓
Trooclick	✓					✓



Full Fact fights bad information



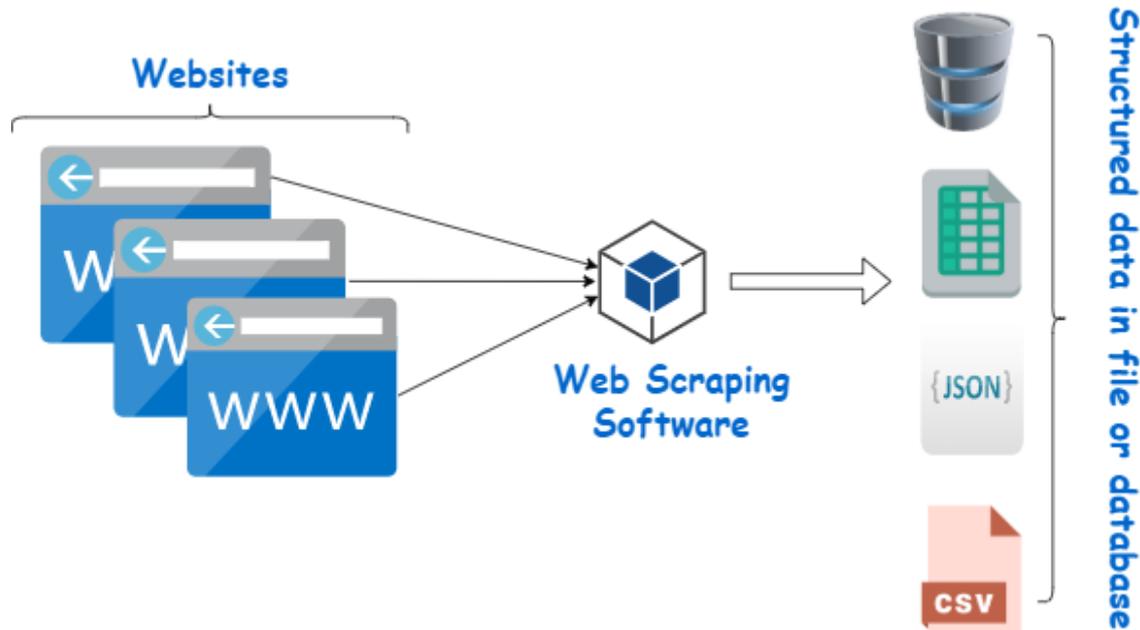
<https://fullfact.org>

Check4Decision Research Project

Acquisition des données par ML



- Web Crawling:
 - Action d'indexer toute l'information d'une page web pour y avoir accès plus facilement.
- Web Scraping :
 - Action d'extraire / récupérer des données à partir de sites web, grâce à un programme ou un logiciel.
 - Automatisation intelligente du processus dans lequel une personne réaliserait des copier/coller manuels

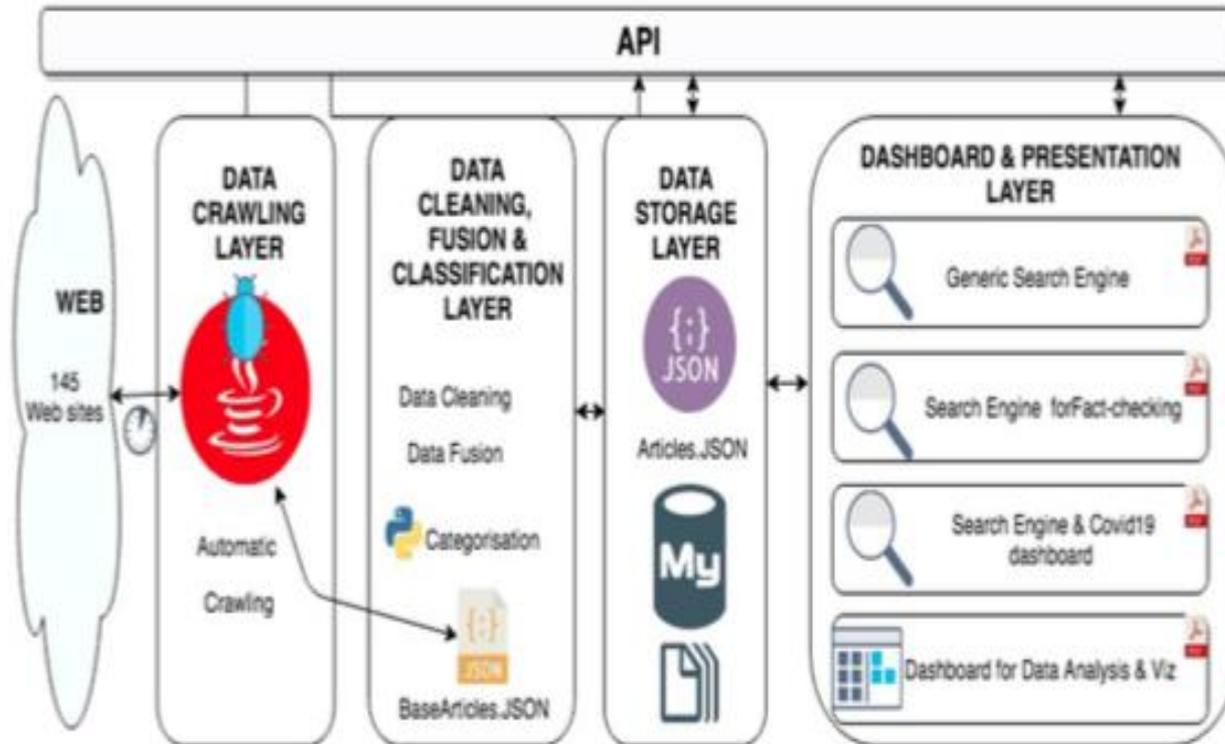


Utilisation:

- Moteurs de recherche
- Comparateurs de prix
- Veille concurrentielle

Check4Decision Research Project

Acquisition des données par ML



Activités

- Indexation des sources (Web Crawling)
- Extraction des articles & commentaires (web Scraping)
- Formatage et uniformisation (Hétérogénéité)
- Structuration et stockage (Format Leger : CSV, JSON, BDR)
- Catégorisation des données (Machine Learning)
- Analyse et Présentation (Moteur de recherche, Tableau de bord ..)

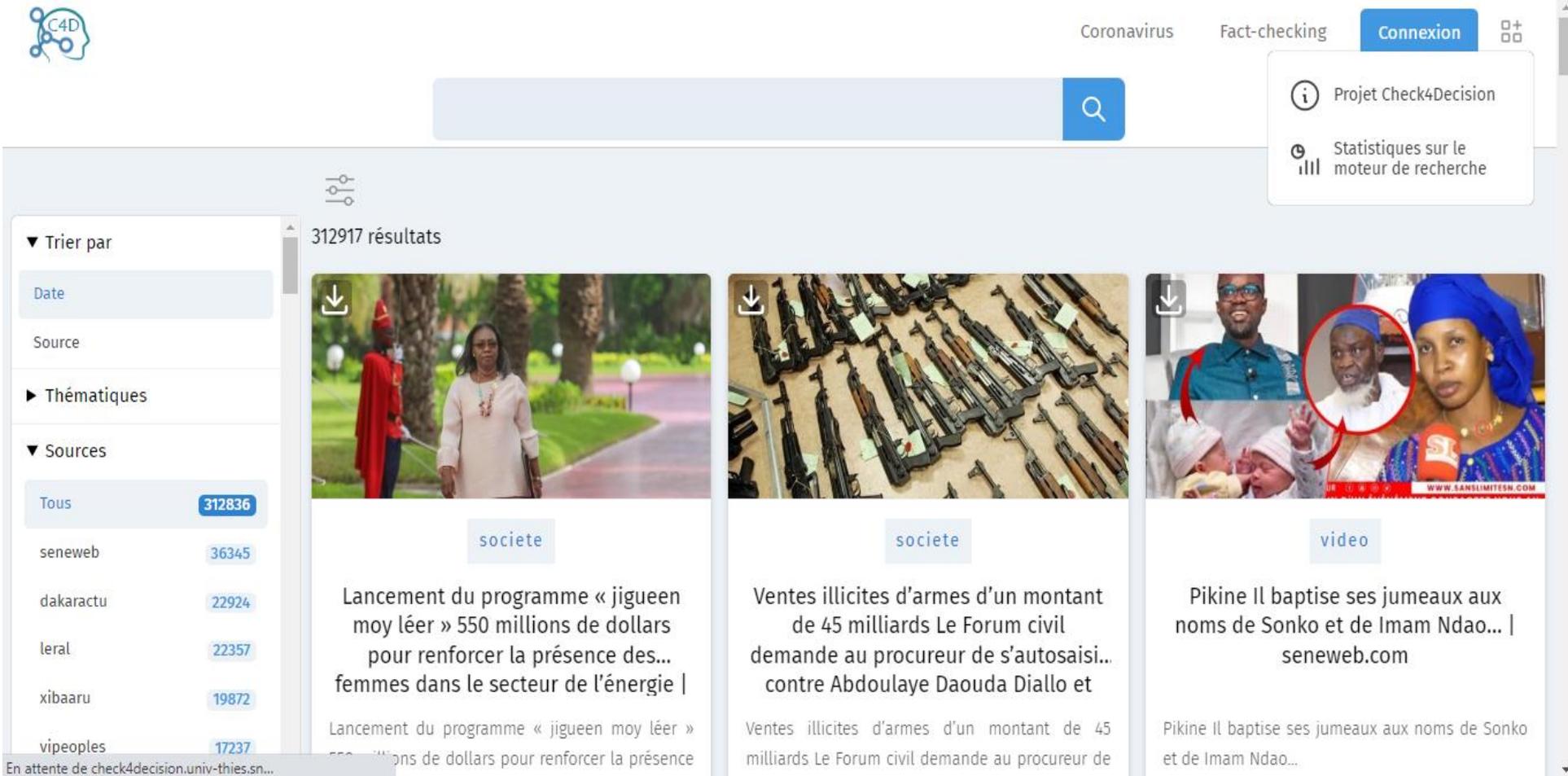
Check4Decision Research Project

Notre Corpus

- Lieu de Stockage : Serveur de UIDT (SSH, FTP & BD)
- Type de sources: Sites Web / Sénégalais
- Type de données : Articles et commentaires
- Format de données : Texte (CSV, JSON ou BDR) et Image (JPG & PNG)
- Nombre de sources : 111
- Fréquence de la collecte: 3 par jours
- Taille du corpus:
 - Articles: +500 000
 - Commentaires: + 10 000 000
- Mode d'accès: OPEN
- Exploitation des données : Plateforme & Tableau de bord
 - Moteur de recherche:
 - <https://check4decision.univ-thies.sn/search/>
 - Tableau de bord:
 - <https://check4decision.univ-thies.sn/search/dashboard/articles>

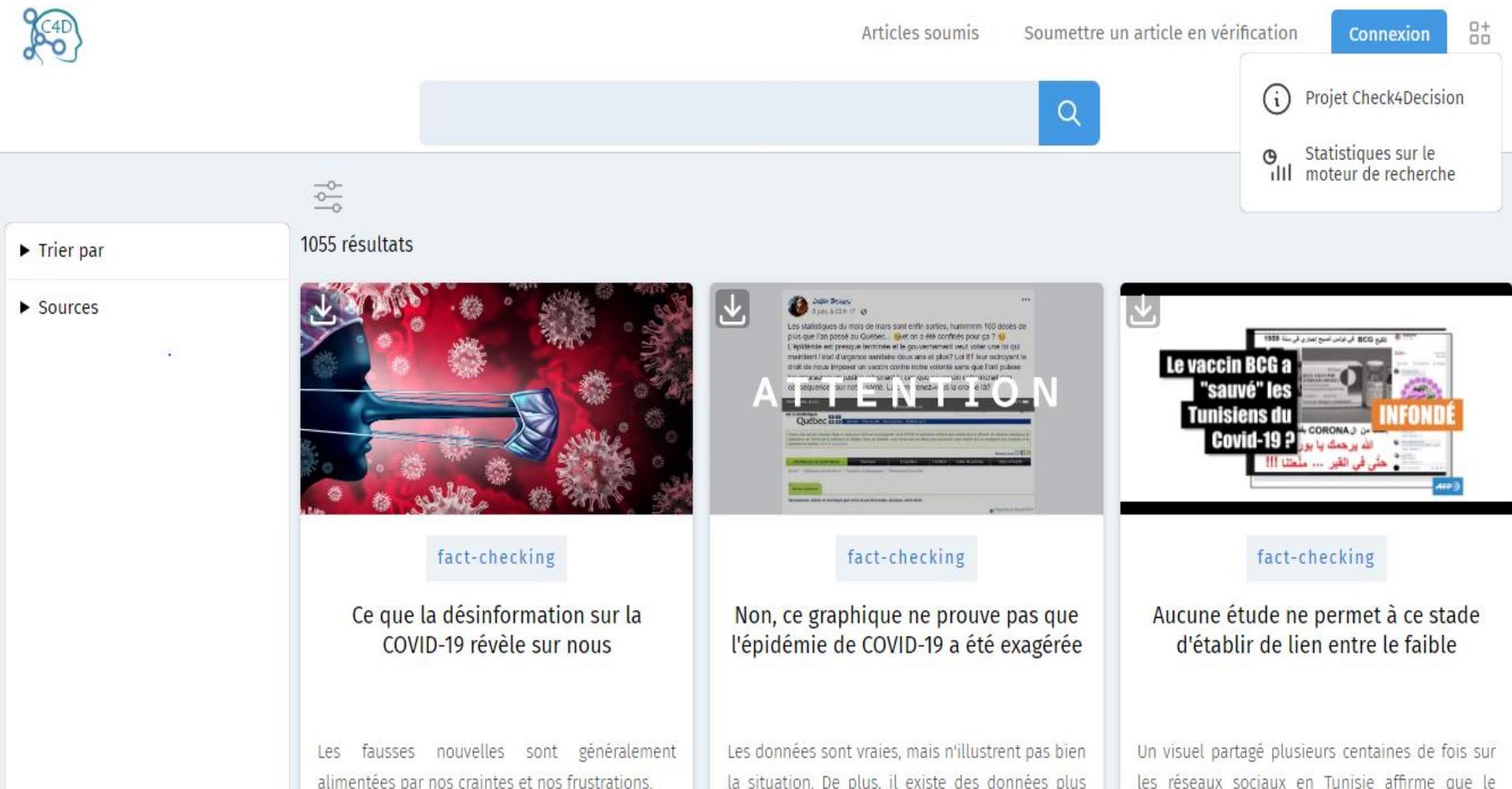
Projets **Check4Decision**

Résultats: Moteur de recherche



The screenshot displays the Check4Decision search engine interface. At the top, there is a search bar with a magnifying glass icon. To the right of the search bar, there are navigation links for 'Coronavirus', 'Fact-checking', and 'Connexion'. A dropdown menu is open, showing 'Projet Check4Decision' and 'Statistiques sur le moteur de recherche'. Below the search bar, the results are displayed in a grid. The first result is a video titled 'Lancement du programme « jiguen moy léer » 550 millions de dollars pour renforcer la présence des... femmes dans le secteur de l'énergie |', with a 'societe' tag. The second result is a video titled 'Ventes illicites d'armes d'un montant de 45 milliards Le Forum civil demande au procureur de s'autosaisi... contre Abdoulaye Daouda Diallo et', also with a 'societe' tag. The third result is a video titled 'Pikine Il baptise ses jumeaux aux noms de Sonko et de Imam Ndao... | seneweb.com', with a 'video' tag. On the left side, there is a sidebar with filters for 'Trier par' (Date, Source), 'Thématiques', and 'Sources' (Tous: 312836, seneweb: 36345, dakaractu: 22924, leral: 22357, xibaaru: 19872, vipeoples: 17237). At the bottom left, there is a status bar that says 'En attente de check4decision.univ-thies.sn...'. The total number of results is 312917.

<https://check4decision.univ-thies.sn/search>



The screenshot shows the Check4Decision search results page. At the top, there is a search bar with a magnifying glass icon. To the right of the search bar, there are navigation links: "Articles soumis", "Soumettre un article en vérification", and "Connexion". Below the search bar, there are two filters: "Projet Check4Decision" and "Statistiques sur le moteur de recherche". On the left side, there is a sidebar with "Trier par" and "Sources" options. The main content area displays 1055 results. Three results are visible, each with a thumbnail image, a "fact-checking" label, and a title. The first result has a thumbnail of a person wearing a face shield and a face mask, with the title "Ce que la désinformation sur la COVID-19 révèle sur nous". The second result has a thumbnail of a social media post with the word "ATTENTION" overlaid, and the title "Non, ce graphique ne prouve pas que l'épidémie de COVID-19 a été exagérée". The third result has a thumbnail of a social media post with the text "Le vaccin BCG a 'sauvé' les Tunisiens du Covid-19?" and "INFONDÉ", and the title "Aucune étude ne permet à ce stade d'établir de lien entre le faible".

Articles soumis Soumettre un article en vérification Connexion

Projet Check4Decision
Statistiques sur le moteur de recherche

1055 résultats

Trier par
Sources

fact-checking
Ce que la désinformation sur la COVID-19 révèle sur nous

fact-checking
Non, ce graphique ne prouve pas que l'épidémie de COVID-19 a été exagérée

fact-checking
Aucune étude ne permet à ce stade d'établir de lien entre le faible

Projets **Check4Decision** Résultats: **Dashboard**



Tableau de bord

Articles collectés

Recherche



307940

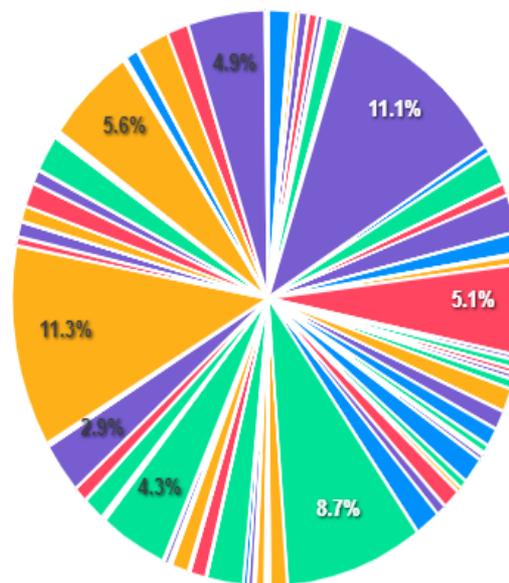
Articles collectés



111

Sources

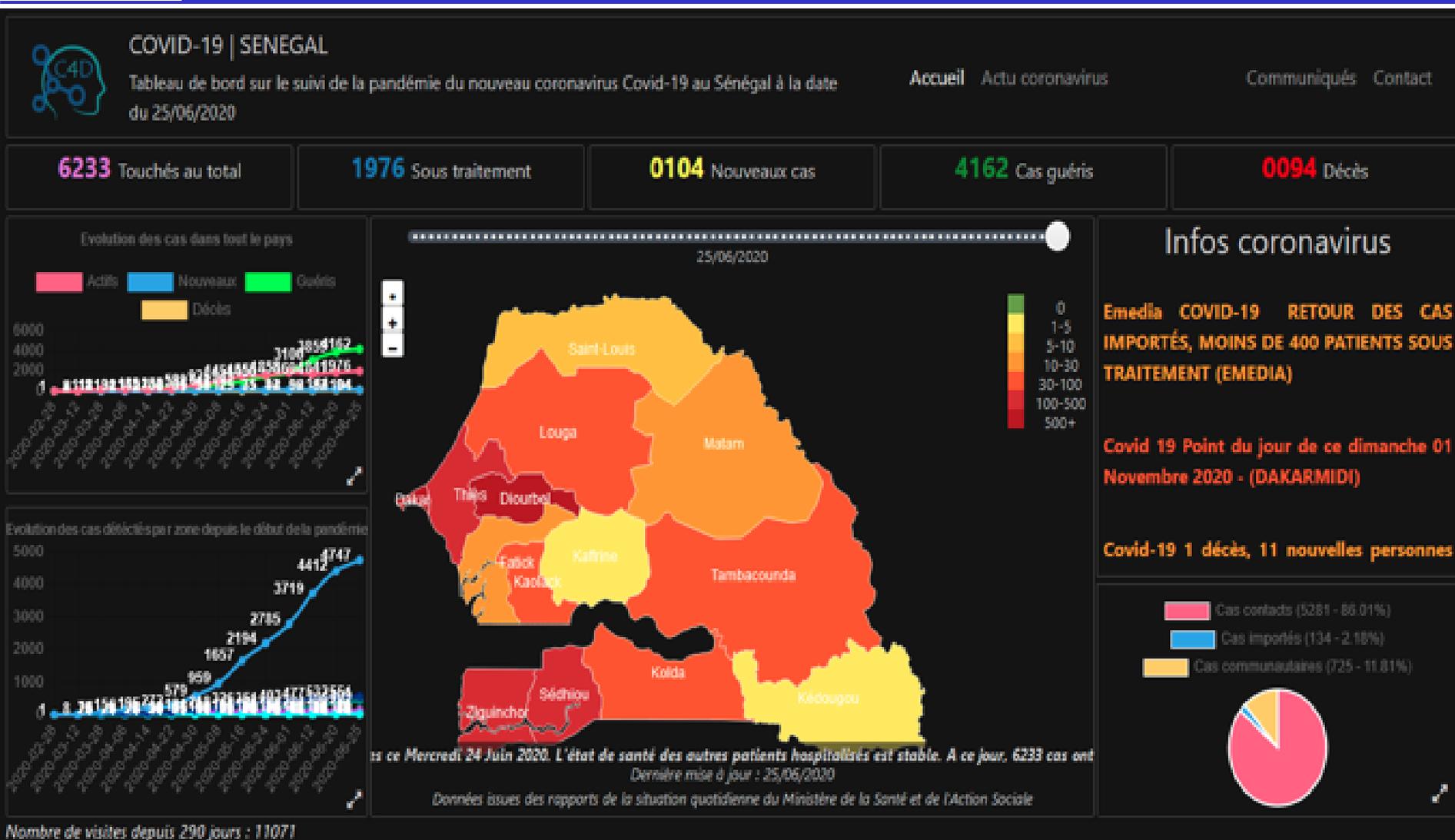
Nombre d'articles par source



- ACTU24 (5594)
- ACTUSEN (808)
- ACTUSENEGAL (1216)
- AFRICACHECK (64)
- ALLODAKAR (2334)
- APS (18)
- ASSIROU (51)
- AZACTU (13)
- BAOLINFO (2263)
- BESTINFOS (1395)
- BUSINESS221 (647)
- CHRONIQUES (4636)
- DAKAR24SN (911)
- DAKAR92 (133)
- DAKARACTU (43992)
- DAKARBUZZ (1847)
- DAKARFLASH (7392)
- DAKARHEBDO (33)
- DAKARMATIN (2815)
- DAKARMIDI (8362)

Projets **Check4Decision**

Résultats: Moteur Fact-Checking



<https://check4decision.univ-thies.sn/search/corona/>

Réutilisation de la Data

Travaux en cours

- Crawling (Indexation) Intelligent par apprentissage profond des données (UADB)
- Extraction large spectre de données Web (Utt)
- La classification automatique des articles de presse par machine Learning (UCAD)
- Modération intelligente des contenus journalistique (UASZ)
- Modélisation de la propagation des fakenews dans la presse en ligne (Dr Mansal-UCAO)
- Identification automatique des faits dans les affirmations journalistique (UIDT)
- Impacts des fakenews dans le choix des électeurs : le cas du Sénégal (Jean d'arc Post BAC-UCAO)
- Opinion Mining sur les commentaires journalistique (Dr Faty)
- Aspects légaux du web Scraping et du Crawling (Dr Bassene-UASZ)

Perspectives

Une base de connaissance

- **Une base de connaissance:**
 - Regroupe des connaissances spécifiques à un domaine spécialisé donné, sous une forme exploitable.
 - Va centraliser toutes les informations disponibles par rapport à un sujet ou un domaine bien donné.
- A court terme : CEDEAO
- A moyen terme: Presse Francophone Africaine
- A long terme : Afrique

Pour plus d'informations

From Intelligent Crawling to Inclusive Fact-Checking: An End-to-End System

Edouard Ngor SARR

UFR SES

Université Assane SECK de Ziguinchor
Ziguinchor-SENEGAL
edouard-ngor.sarr@univ-zig.sn

Lamine FATY

Check4Decision Research Project

Université Assane SECK de Ziguinchor
Ziguinchor-SENEGAL
lamine.faty@univ-zig.sn

Moussa Déthié SARR

UFR-Sciences et Technologies
Université de Thiès
Thiès-SENEGAL
mdsarr@univ-thies.sn

Mouhamadou Moustapha SISSOKHO

Check4Decision Research Project

Université de Thiès
Thiès-SENEGAL
siskomouhamed@gmail.com

Fatima TOURE

Check4Decision Research Project

Université de Thiès
Thiès-SENEGAL
toure.fatima947@gmail.com

Ousmane SALL

UFR-Sciences et Technologies
Université de Thiès
Thiès-SENEGAL
osall@univ-thies.sn

Abstract: In this article, we present an aggregation platform of journalistic contents based on an intelligent crawler of articles from Online Press, of an inclusive fact-checking approach and an Opinion Mining method. More than 300 000 press articles from more than 145 Senegalese online information websites have been collected, processed, analyzed, aggregated, stored, and classified by category and theme, thanks to Machine Learning. The primary objective in these researches is to provide Web surfers, mainly journalists and fact-checkers, with a unique platform aggregating all the data related to Senegalese current affairs with analysis options, display and fact-checking.

Keywords: Fact-checking, Crawling, Scraping, Machine Learning, Classification, Search engine.

I. INTRODUCTION & MOTIVATIONS

- The first search⁵ engine is generic and has more than 300 000 articles from the Senegalese online press.
- The second search engine is the result of a filter related to Covid-19.
- The third and last one (fact-checking search engine) only shows press articles which have been already checked by certified fact-checkers journalists. That last search engine is directly interfaced with many famous sites such as AfricaCheck⁶, ICI Radio Canada⁷, Factual AFP⁸, etc. Its role is to minimize the negative impacts of fake news on public opinion.

An inclusive fact-checking opinion is also inserted into each of these search engines as well as two dashboards; one about Covid-19 in Senegal and another one for the analysis

Merci de votre attention

