



**Formation initiale**

**MÉMOIRE FIN DE FORMATION POUR L'OBTENTION DU DIPLÔME D'INGÉNIEUR DE  
CONCEPTION DES TÉLÉCOMMUNICATIONS**

**OPTION : Ingénierie des Données et Intelligence Artificielle**

**Conception d'un modèle de Clustering pour la détection de  
Fausses Informations au sein de la presse sénégalaise en ligne**

**SOUS LA DIRECTION DE**

**Pr Mamadou BOUSSO,  
Enseignant Chercheur à l'UIDT**

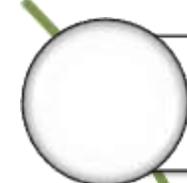
**M. Jean Marie PREIRA  
Enseignant à l' ESMT**

**PRÉSENTÉ ET SOUTENU PAR**

**M. Moussa Steve B. SANOGO**

**DÉCEMBRE 2024**





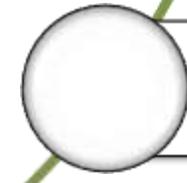
## Introduction

Généralités sur les Fake News

ML, NLP & Détection de Fake News

Conception du Modèle de Détection

Résultats & Perspectives



## Conclusion



# Introduction

**FAKE  
NEWS**



# Les Fake News

Les **Fake News** sont des **fausses histoires**, ressemblant aux récits d'information authentiques, diffusées sur Internet ou sur d'autres supports médiatiques, principalement à but politique ou comique.



*Caractéristiques*

## Eléments Constitutifs

Créateur/Diffuseur	Cible	Contenu	Contexte Social
 • <b>Humains</b> • <b>Robots</b>	 • <b>Électeurs</b> • <b>Entreprises</b> • <b>Personnalités Publiques</b>	 • <b>Textes</b> • <b>Images</b> • <b>Vidéos</b>	

## Types de Fake News



### Désinformation en Santé

- *Les vaccins*
- *Les remèdes miracles*
- *Les épidémies*



### Désinformation en Politique

- *Mensonges sur les candidats*
- *Résultats d'élections*
- *Propagande politique*



### Les Fake News de Divertissement

- *D'articles satiriques*
- *Buzz*

## Mécanismes de Propagation

Titres sensationnels  
et accrocheurs

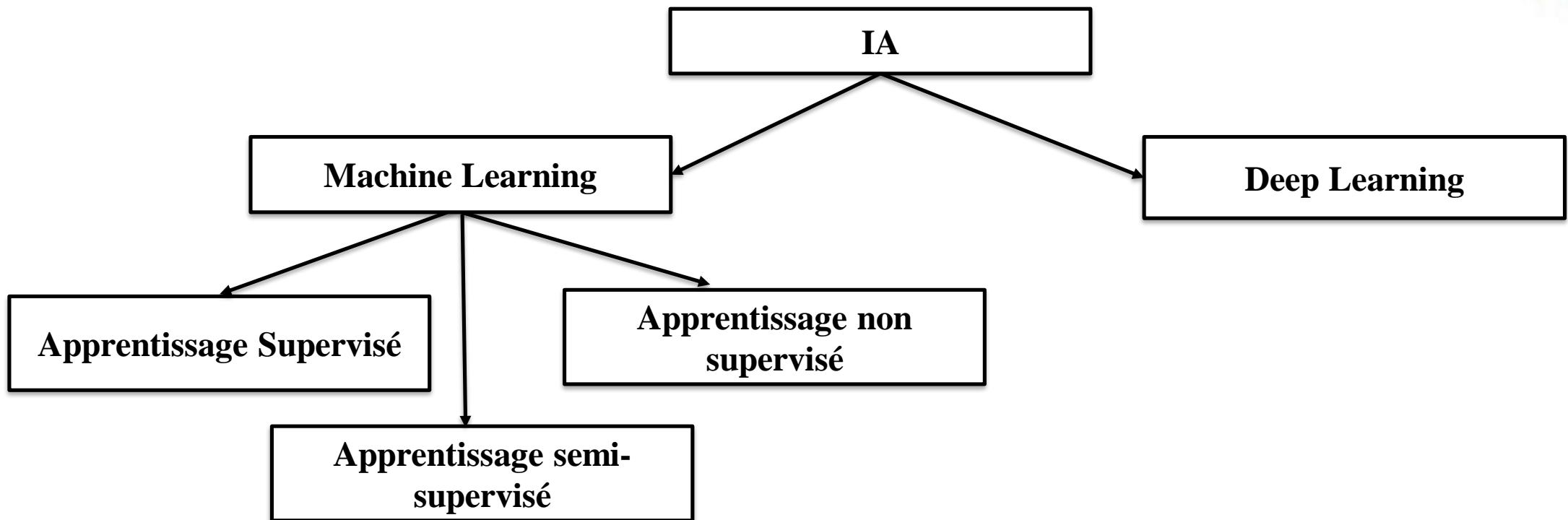
Les commentaires  
et interactions

Les réseaux  
sociaux

Techniques de  
référencement

L'IA est l'ensemble des **théories et techniques de simulation de l'intelligence humaine** sur des machines. C'est une discipline à part entière depuis les années 1950, qui a permis de réaliser d'importants progrès ces dernières années.

## Quelques domaines de l'IA



Les principaux algorithmes utilisés en Machine Learning, et Deep Learning sont les suivant :

## Apprentissage Supervisé



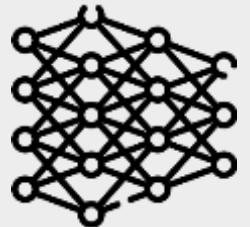
- Régression Linéaire
- Régression Logistique
- K plus proche voisin
- Machines à vecteurs de support (SVM)

## Apprentissage non supervisé



- K-moyennes
- Clustering hiérarchique
- Analyse en composantes principales (ACP)
- Décomposition en valeurs singulières (SVD)

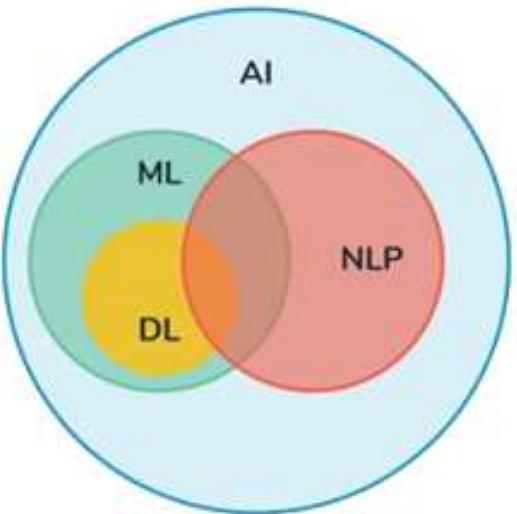
## Deep Learning



- Réseaux neuronaux convolutifs (CNN)
- Réseaux neuronaux récurrents (RNN)
- Réseaux de mémoire à long et court terme (LSTM)

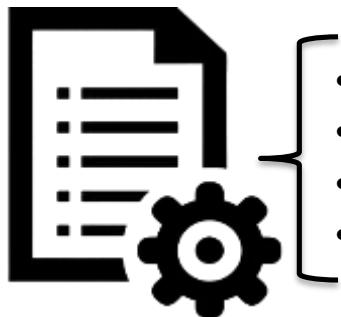
Le NLP est un sous-domaine de l'IA, qui se consacre à l'analyse et à la production du langage humain par des ordinateurs. Ces principales branches :

- ❖ Compréhension du Langage Naturel (NLU)
- ❖ Génération du Langage Naturel (NLG)
- ❖ L'Analyse Syntaxique

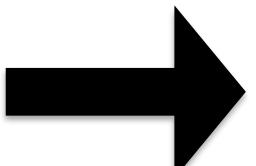


Globalement, nous pouvons distinguer **deux aspects** essentiels à tout problème de NLP :

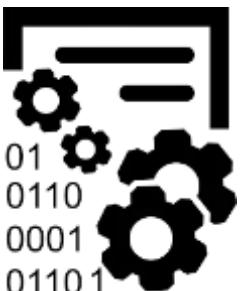
## Prétraitement



- Suppression ponctuation
- Suppression des Stopwords
- Tokenisation
- Etc.



## Data Science



- Vectorisation des données
- Application d'algorithmes de ML ou DL

## ❖ Vectorisation basée sur la syntaxe

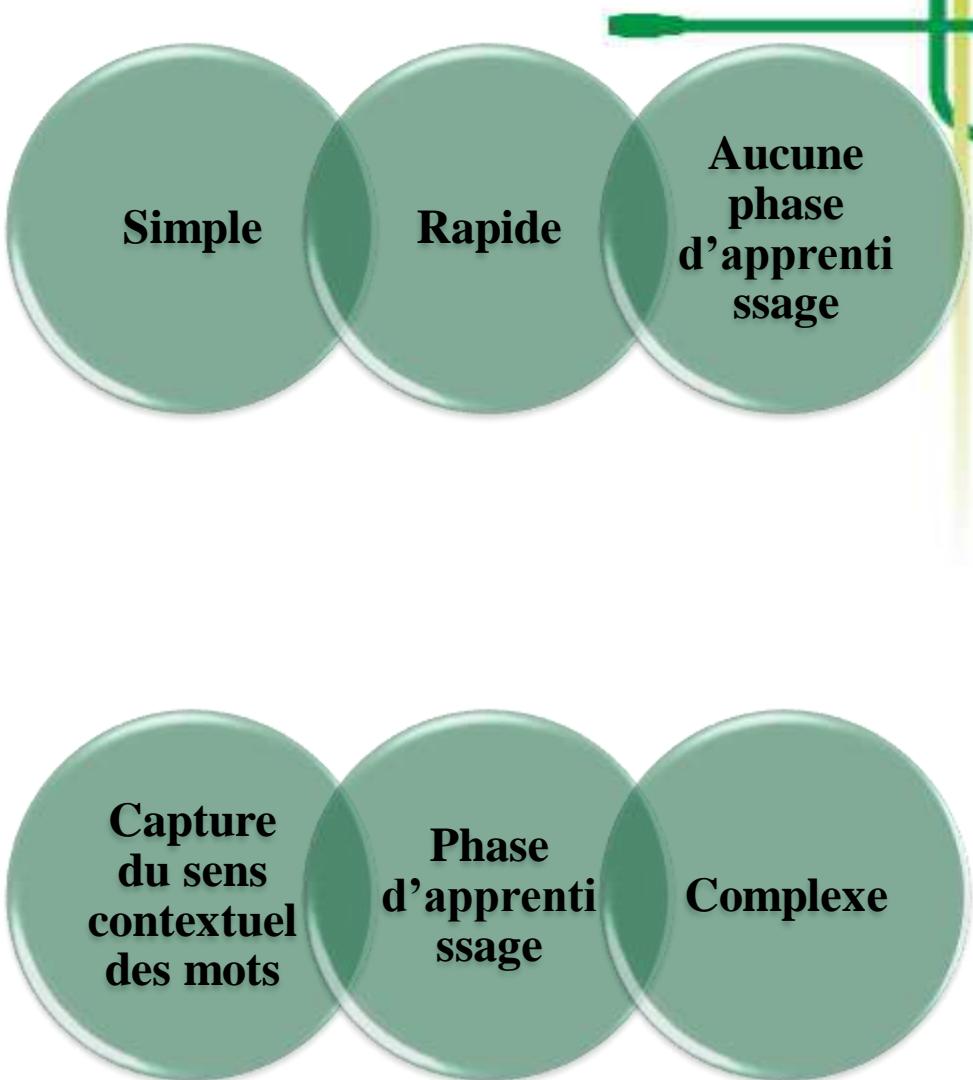
*L'ensemble des mots d'un document est considéré comme un sac de mots, sans tenir compte de l'ordre ni des relations entre eux.*

*Exemple : One Hot Encoding, Bag of Word, TD-IDF, etc.*

## ❖ Vectorisation basée sur la Sémantique

*Les mots sont transformés en vecteurs en tenant compte de leur sens et du contexte dans lequel ils apparaissent.*

*Exemple : Word2Vec, BERT, etc.*



# ML, NLP & Applications



*La Santé*



*L'éducation*



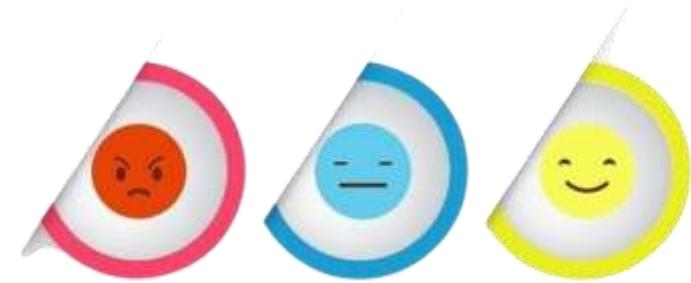
*L'agriculture*



*Les chatbots*

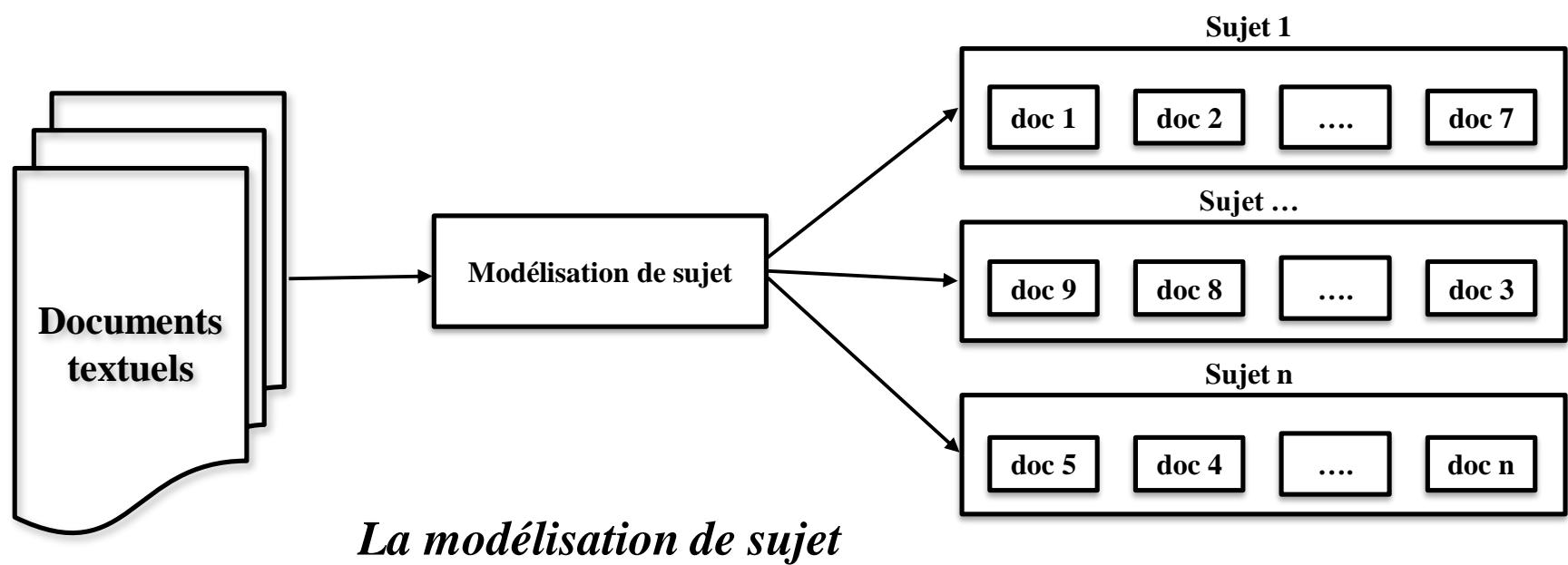
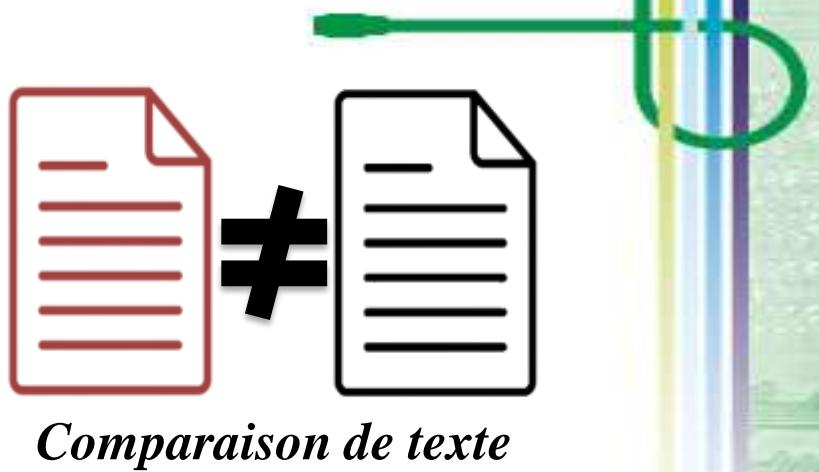
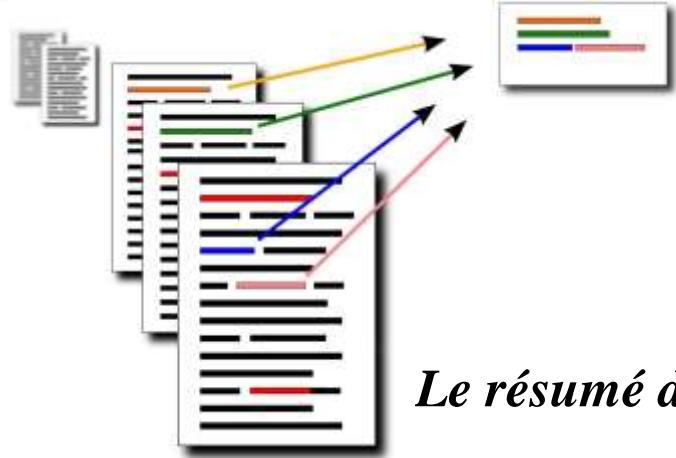


*Traduction automatique*

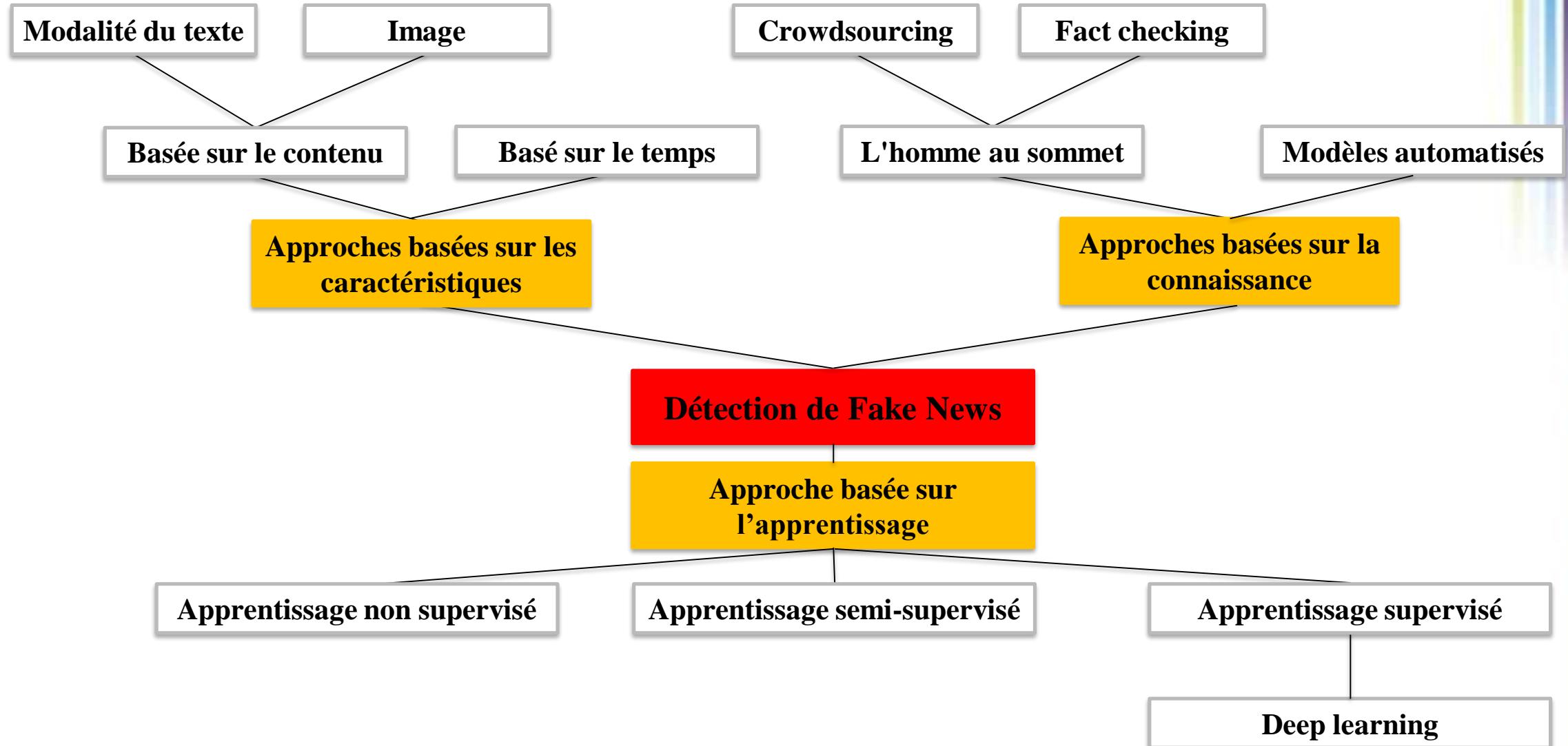


*L'Analyse de Sentiment*

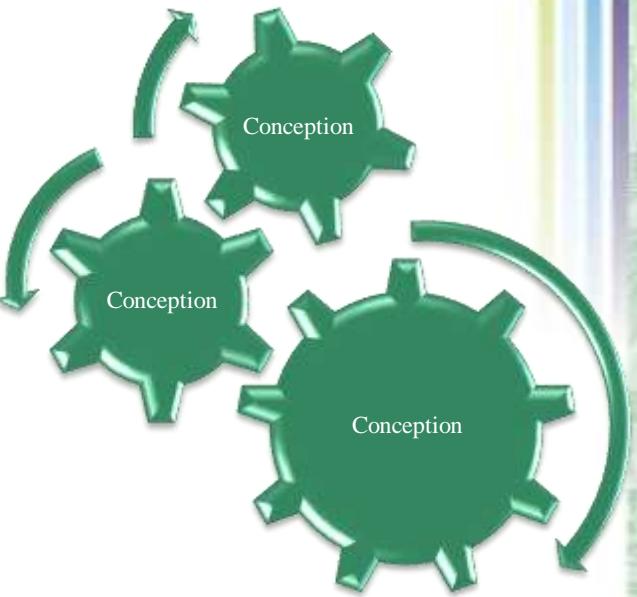
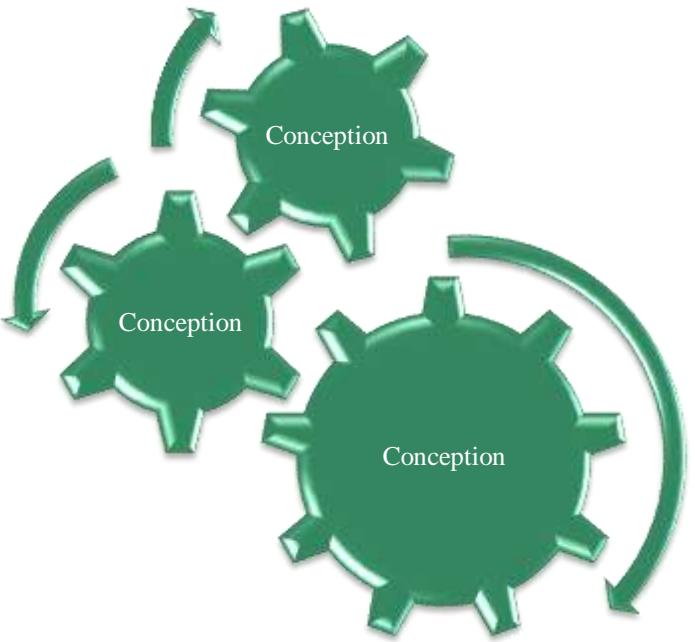
# ML, NLP & Applications



# ML, NLP et Détection de Fake News



# Conception du Modèle

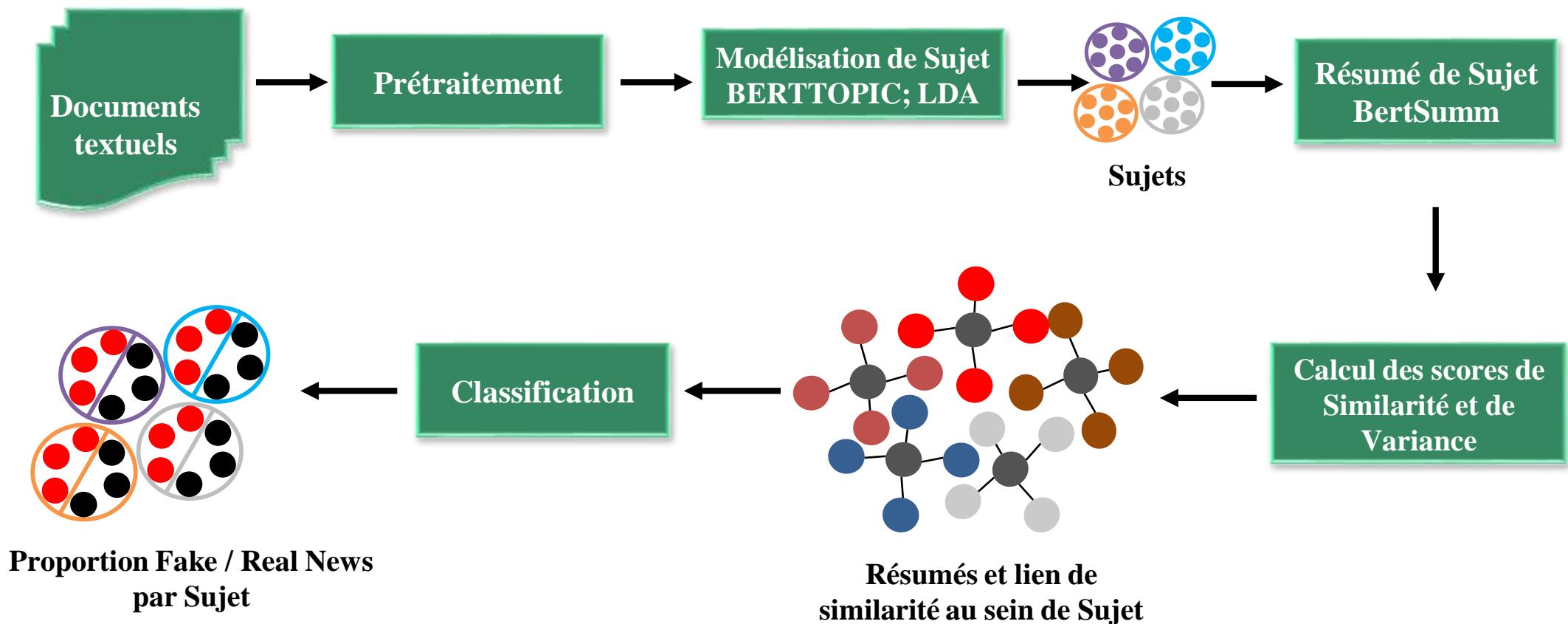


- ❖ *Présentation du modèle*
- ❖ *Classification des articles*

- ❖ *Bibliothèques & Env. d'exécution*
- ❖ *Perspectives d'amélioration*

# Modèle de détection

Le modèle de détection proposé repose sur des techniques d'apprentissage **non supervisé**. Il combine les techniques de **Topic Modeling**, de **résumés de texte**, et de **calcul de similarité** entre articles pour identifier les contenus potentiellement douteux.



# Source de Données

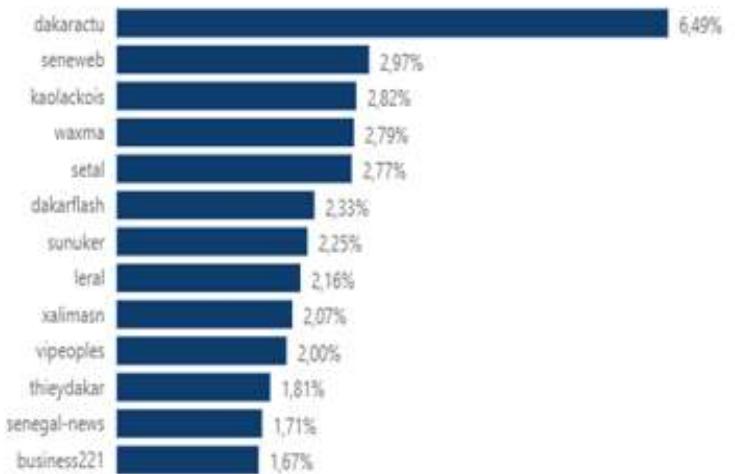
Notre base de données est au format JSON, pèse environ 31 Mo et a été obtenue par **Web Scraping**. La structure du fichier est basée sur une liste d'articles de presse en ligne, où chaque article est représenté par un objet JSON.

Champ	Format
urlArticle	C.C
sourceArticle	C.C
datePublicationArticle	C.C
contenuArticle	C.C
nombreLikesArticle	Entier
nombreLecturesArticle	Entier
nombreCommentairesArticle	Entier
nombrePartagesArticle	Entier

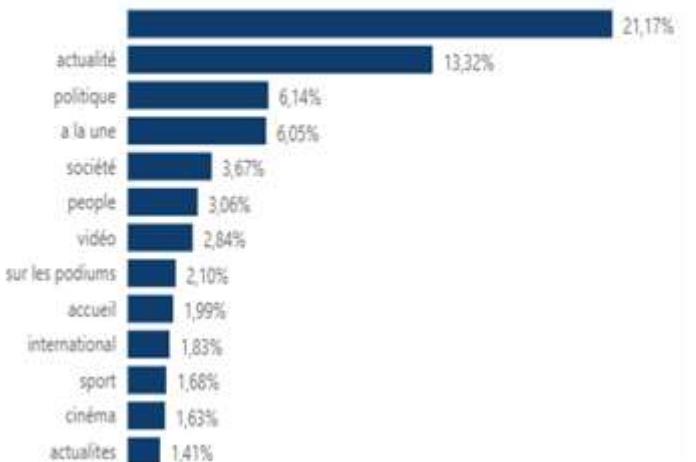


# Analyse Exploratoire

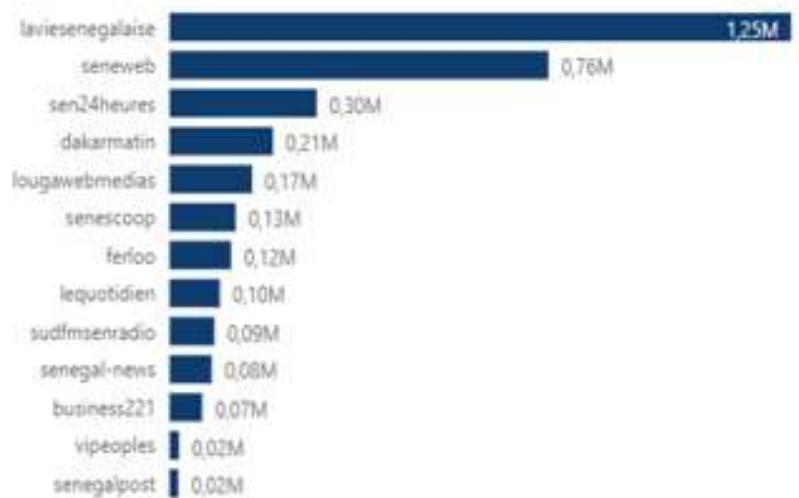
% article selon la source



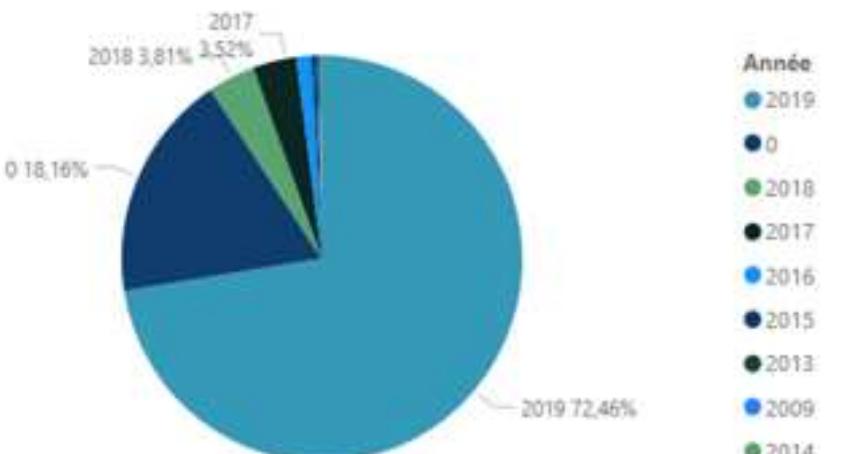
% article selon la Thématique



Nombre de lectures par source d'article



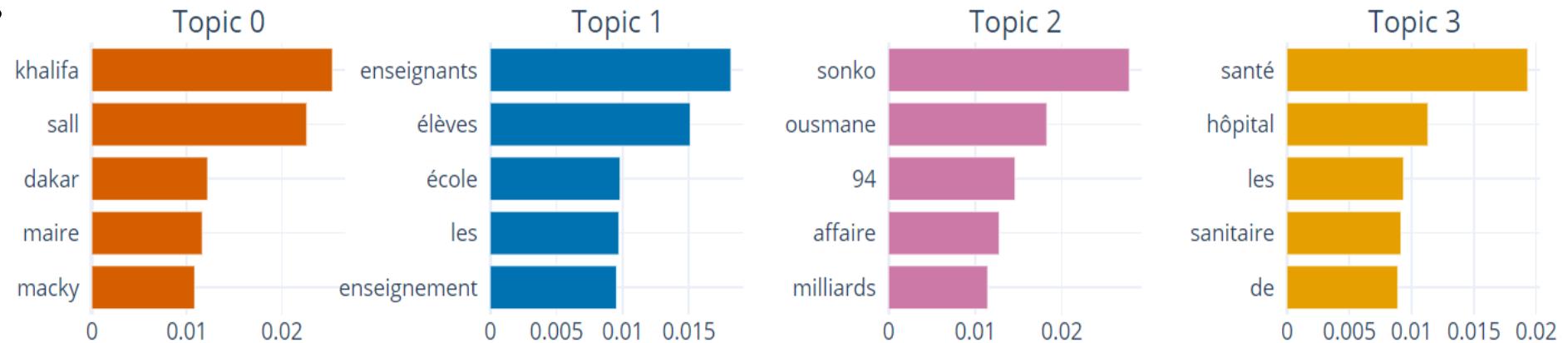
% article selon l'année de publication



# Modélisation de Sujet : BertTopic

Après la phase d'entraînement du modèle sur **5530 articles**, nous avons obtenu un ensemble de **115 sujets très variés**, ainsi que **1943 articles** classés comme n'appartenant à **aucun sujet**. Les sujets obtenus sont très variés :

- ❖ Ousmane Sonko et l'affaire des 94 milliards
- ❖ Réchauffement climatique
- ❖ Sécurité aéroportuaire
- ❖ Etc;



*Top 4 des sujets détectés*

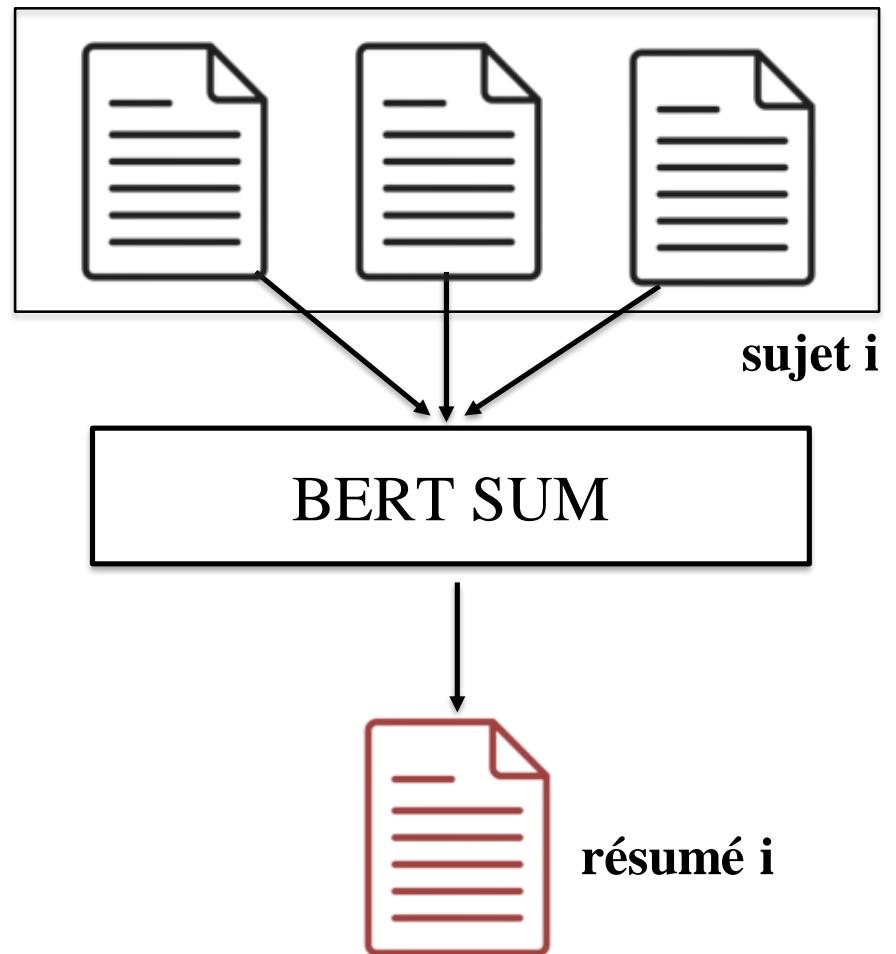
# Modélisation de Sujet : LDA

LDA est utilisé en complémentarité avec BertTopic. Grâce à son approche statistique basée sur les fréquences de mots, permettra de regrouper les **1943** outliers en nouveaux sujet cohérents.

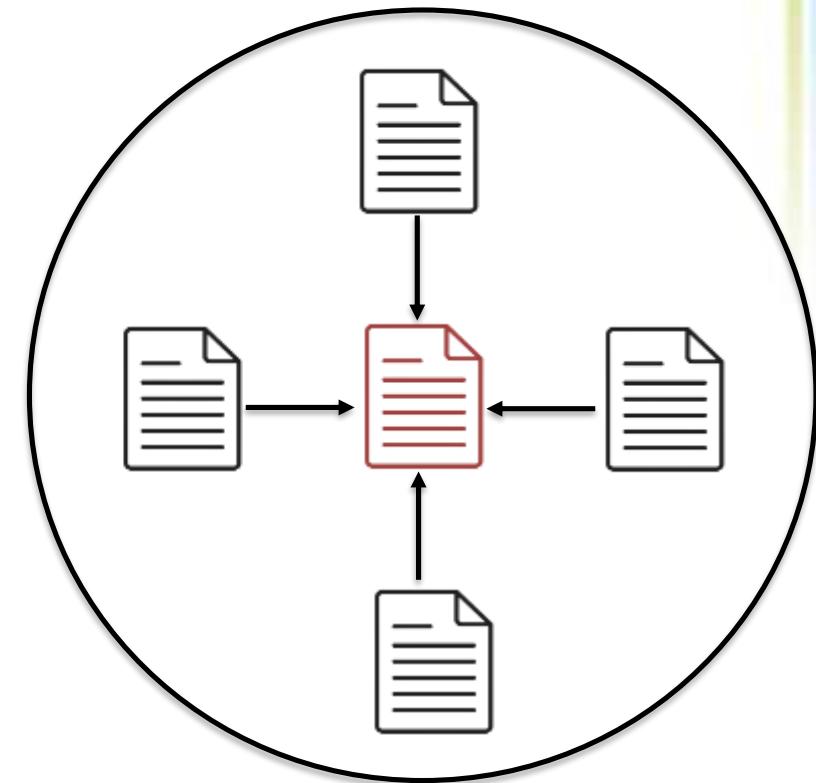
Sujet 1	Sujet 2	Sujet 3	Sujet 4	Sujet 5	Sujet 6	Sujet 7
karim	priver	sall	prendre	cheikh	afrique	2019
latifah	secteur	macky	malick	serigne	développement	tweet
bien	secteur_priver	politique	voir	général	milliard	match
mohamed	etat	pouvoir	aicha	touba	economique	aliou
aller	sall	wade	savoir	religieux	projet	joueur
venir	contrat	bien	fois	homme	africain	brésil

# Résumé de Sujet et Calcul de similarité

## Phase de résumé de sujet



## Calcul de similarité



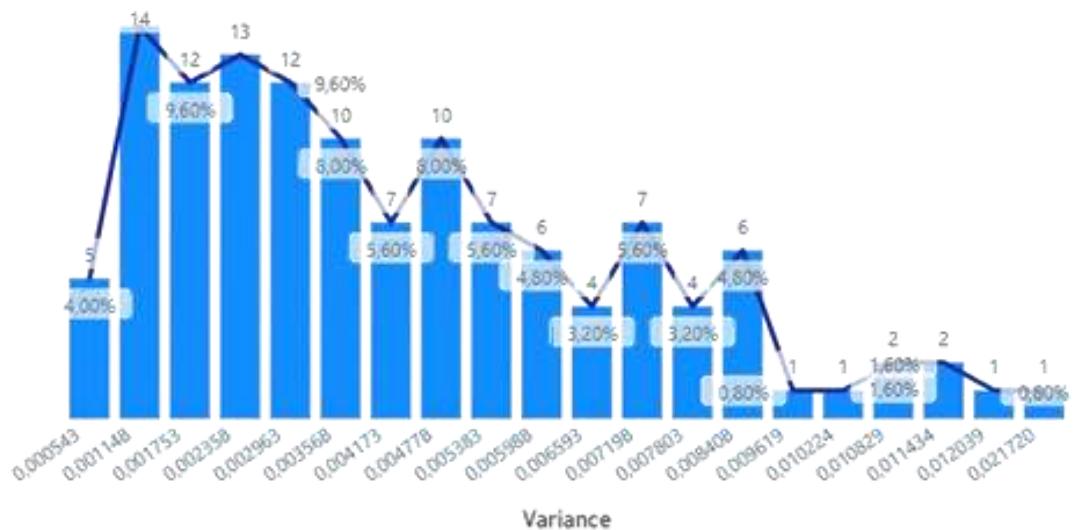
article



résumé

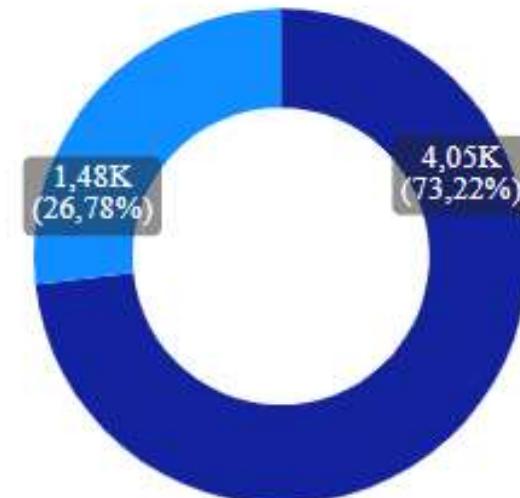
Score de similarité

# Classification des articles



*Seuil de Similarité  
déterminé à 0.11*

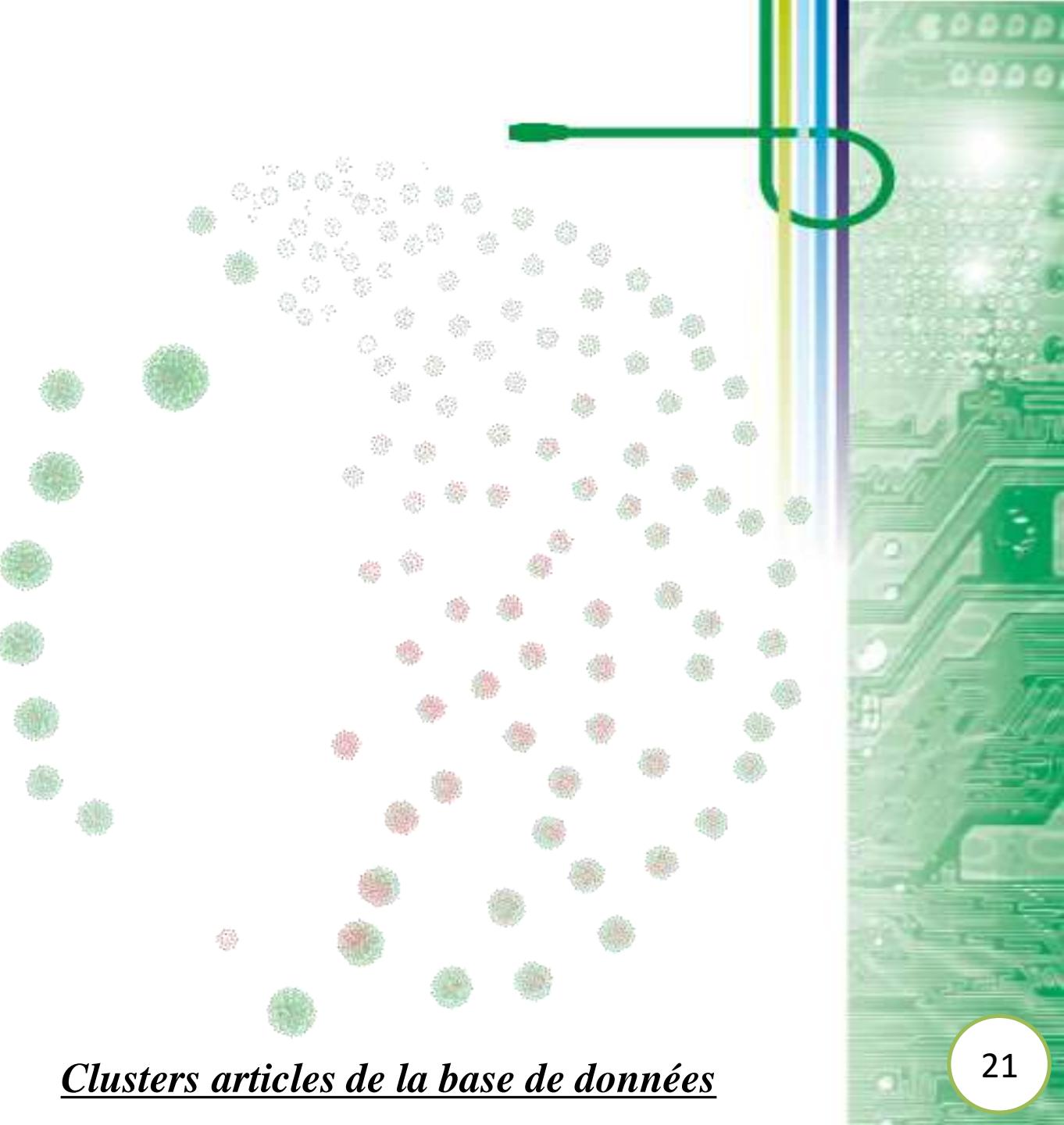
- Classe
- NoFake
  - Fake



# Classification des articles

L'image ci-contre représente notre base de données après application du modèle. On peut y observer :

- ❖ 125 clusters d'articles
- ❖ Chaque cluster représente un sujet
- ❖ En orange les résumés de sujet
- ❖ En rouge les articles potentiellement fallacieux
- ❖ En vert les articles factuels



Clusters articles de la base de données

# Classification des articles



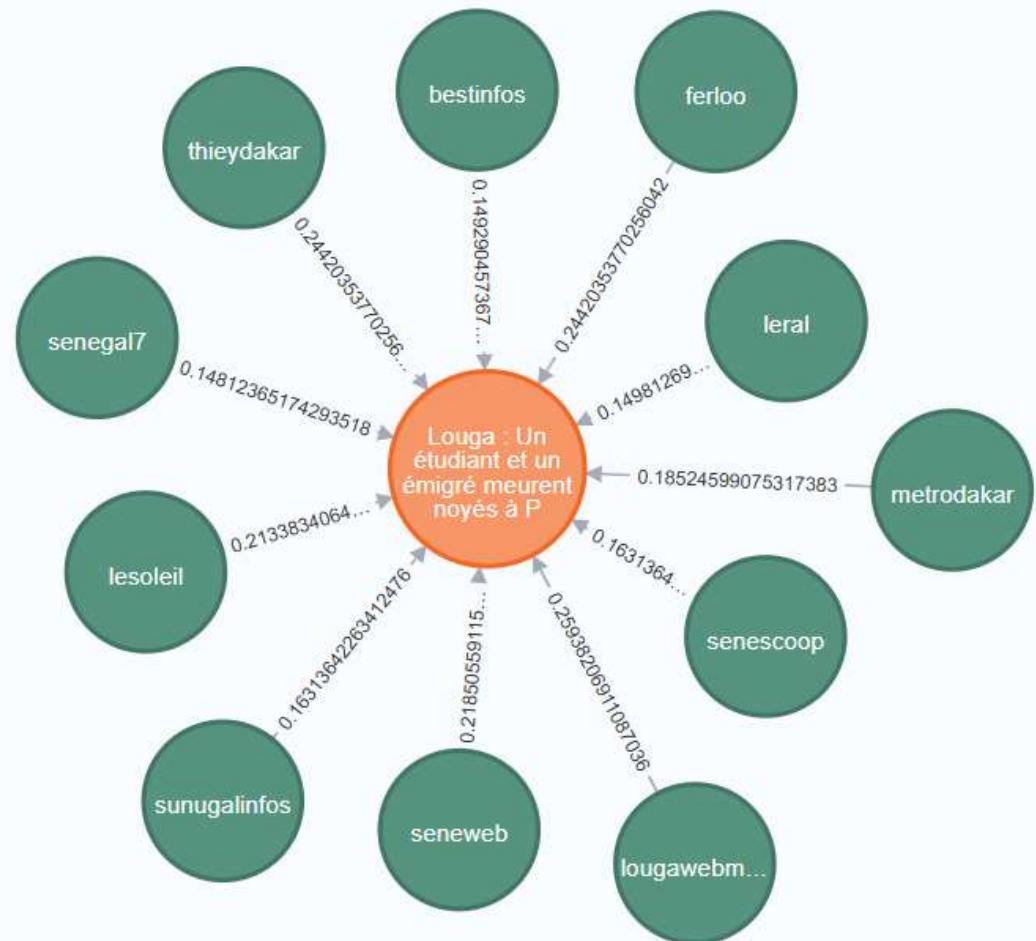
TYPES DE SUJET				
Sans Fake News	Avec peu de Fake News	Avec une proportion modérée de Fake News	Avec une forte proportion de Fake News	
TOTAL				
19	23	27	56	
CARACTÉRISTIQUES				
<ul style="list-style-type: none"><li>• Domaines juridiques et administratifs</li><li>• Sécurité et diplomatie (relations internationales, Procédures de contrôle)</li></ul>	<ul style="list-style-type: none"><li>• Domaines institutionnels (Justice, Diplomatie)</li><li>• Événements culturels (Prix, événements religieux)</li></ul>	<ul style="list-style-type: none"><li>• Politique et institutions (processus électoraux, personnalités politiques)</li><li>• Santé, société et économie et dévelop.</li></ul>	<ul style="list-style-type: none"><li>• Sport et événements sportifs</li><li>• Politique et événements territoriaux (Collectivités</li></ul>	

# Vue sur le sujet 111

Le sujet 111 aborde un événement tragique : la noyade de plusieurs jeunes hommes à Louga, au Sénégal.

Parmi les médias ayant couvert cette tragédie figurent notamment **Seneweb**, **Senescoop**, **Metro Dakar**, **Leral**, **Ferloo**, **Bestinfos**, **Thiey Dakar**, **Senegal7**, **Le Soleil**, **Sunugalinfos** et **Lougawebmedias**.

Ces articles se distinguent par leur forte cohérence narrative. La plupart citent **Radio Sénégal** comme source principale.



Représentation graphique du sujet 111

# Vue sur le sujet 1

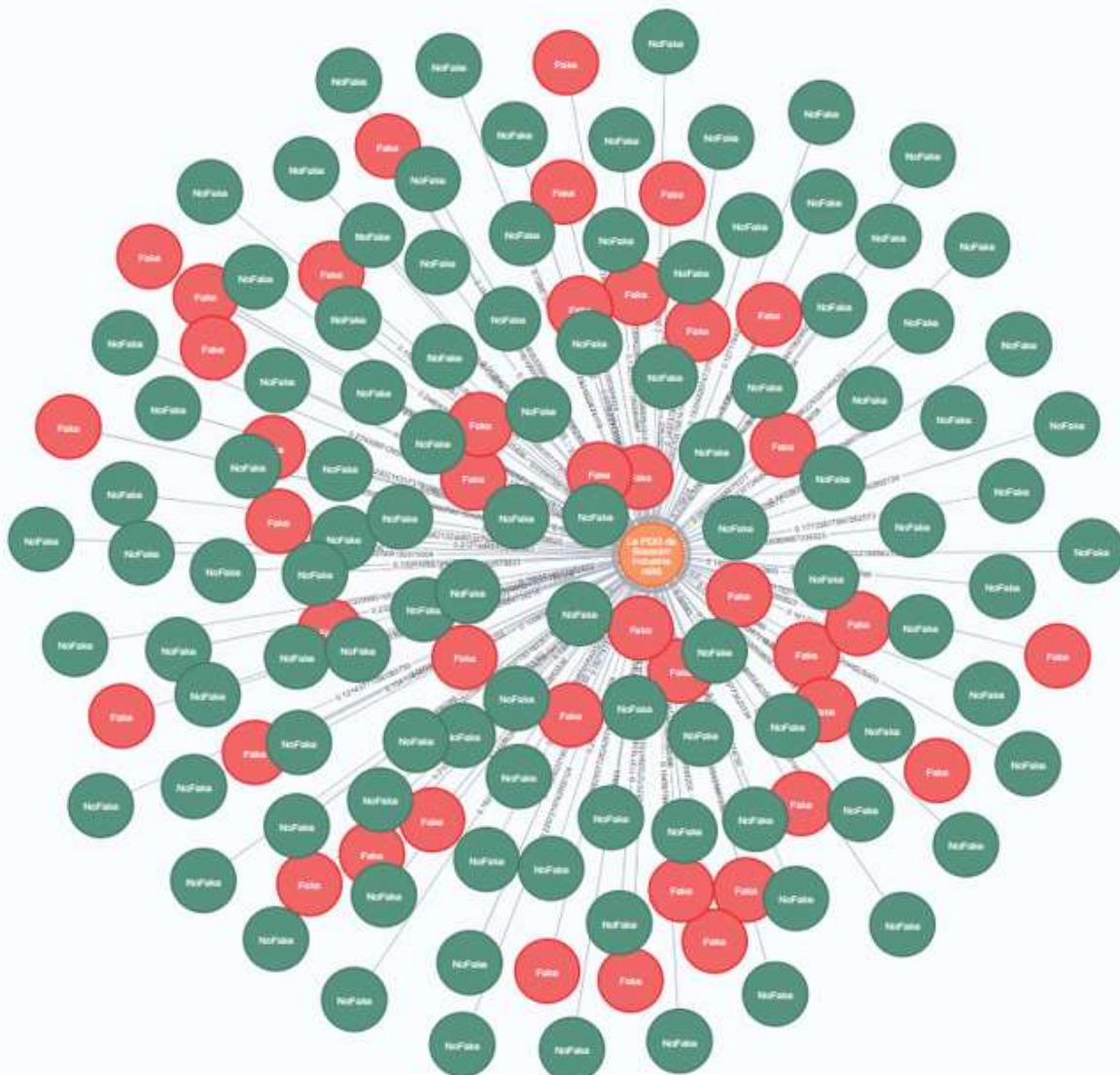
Le sujet 1 est indéniablement centré sur l'éducation au Sénégal.

Les articles mettent en lumière plusieurs aspects clés du système éducatif sénégalais :

- ❖ l'accès à l'éducation,
- ❖ les conditions d'enseignement
- ❖ les politiques
- ❖ la vie universitaire

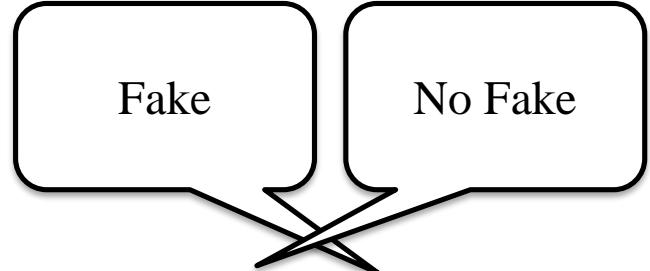
Parmi ses 144 articles, **29,17 %** ont été identifiés comme **potentiellement fallacieux**.

Les articles proviennent de sources variées, telles que **Seneweb**, **Senescoop**, **Seneplus**, **SenegalDirect**, **SenegalActus**, **Sen360**, **Leral**, et **Le Quotidien**.



Représentation graphique du sujet 1

# Perspectives d'amélioration



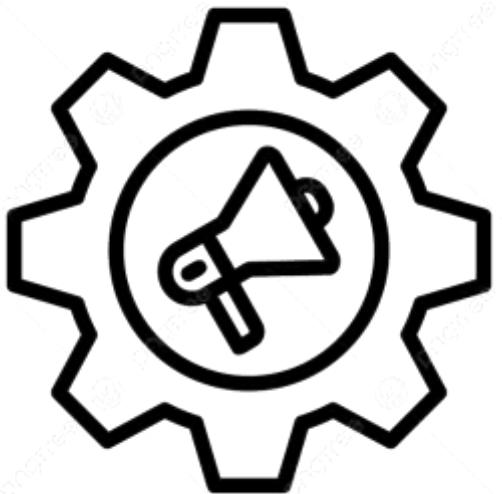
**Etiquetage Manuelle**



**Création d'organismes et de plateformes de détection**



**Semi-Supervised Learning**



**Sensibilisation et Formation des acteurs**

## Environnement de Travail & Bibliothèque



spaCy



LDA

**MERCI DE  
VOTRE  
ATTENTION**



**Formation initiale**

**MÉMOIRE FIN DE FORMATION POUR L'OBTENTION DU DIPLÔME D'INGÉNIEUR DE  
CONCEPTION DES TÉLÉCOMMUNICATIONS**

**OPTION : Ingénierie des Données et Intelligence Artificielle**

**Conception d'un modèle de Clustering pour la détection de  
Fausses Informations au sein de la presse sénégalaise en ligne**

**SOUS LA DIRECTION DE**

**Pr Mamadou BOUSSO,  
Enseignant Chercheur à l'UIDT**

**M. Jean Marie PREIRA  
Enseignant à l' ESMT**

**PRÉSENTÉ ET SOUTENU PAR**

**M. Moussa Steve B. SANOGO**

**DÉCEMBRE 2024**

