



École Supérieure Multinationale des Télécommunications



BÉNIN BURKINA FASO GUINÉE MALI MAURITANIE NIGER SÉNÉGAL

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

SPECIALITÉ : Ingénierie des Données et Intelligence Artificielle

THÈME

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 Belvin SANOGO

Promotion 2020 - 2023

Décembre 2024

¤ DÉDIDACE ¤

Je dédie ce modeste travail :

- A mon défunt père, **SANOGO** Oumar ;
- A ma chère mère ;
- A toute ma famille, tantes, oncles, cousines et cousins et ami(e)s ;
- A mes professeurs ;
- A tous ceux qui ont contribué à ma vie et à mon éducation.

☞REMERCIEMENTS ☞

Le présent travail, qui marque la fin de notre parcours dans le second cycle à l'ESMT, est le résultat de trois (03) années d'études, marquées par de multiples sacrifices, privations et efforts. Il n'aurait pas été possible sans les contributions diverses et précieuses de nombreuses personnes, à qui nous exprimons notre profonde et sincère gratitude. Nous tenons particulièrement à remercier :

- M. Adamou Moussa SALEY, Directeur général de l'ESMT ;
- M. Ahmed KORA, Directeur de l'Enseignement, de la Formation et de la Recherche ;
- Pr. Boudal NIANG, notre responsable pédagogique ;
- Pr Mamadou BOUSSO, directeur de notre mémoire, enseignant chercheur en informatique à l'université IBA DER THIAM de Thiès pour ses orientations ;
- M. Jean-Marie PREIRA, notre codirecteur de mémoire, enseignant à l'ESMT, pour son assistance et ses orientations ;
- Dr. Alla LO, enseignant à l'ESMT ;
- M. Hervé OUEDRAOGO, assistant programme du cycle ingénieur ;
- Monsieur le président ainsi qu'aux autres membres du jury pour le temps qu'ils ont passé à évaluer notre travail ;
- Tout le corps professoral et administratif de l'ESMT.

SIGLES ET ABREVIATIONS

BERT	Bidirectional Encoder Representations from Transformers
BES	Bert Extractive Summarizer
CBOW	Continuous Bag of Words
CNN	Convolutional Neural Networks
cTF-IDF	Class-based TF-IDF
GPU	Graphics Processing Unit
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
HTTPS	Hypertext Transfer Protocol Secure
IA	Intelligence Artificielle
InDIA	Ingénierie des Données et Intelligence Artificielle
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
LDA	Latent Dirichlet Allocation
LLM	Large Language Models
LSA	Latent Semantic Analysis
MLM	Masked Language Modeling
MNR	Maximum Marginal Relevance
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
NSP	Next Sentence Prediction
OHE	One Hot Encoding
PV-DBOW	Paragraph Vector - Distributed Bag of Words
PV-DM	Paragraph Vector - Distributed Memory
RNN	Recurrent Neural Networks
TALN	Traitemen Automatique du Langage Naturel
TF - IDF	Term Frequency – Inverse Document Frequency
UMAP	Uniform Manifold Approximation and Projection
URL	Uniform Resource Locator
VIH	Virus de l'Immunodéficience Humaine

LISTE DES FIGURES

Figure 1 : Approches de détection des fausses nouvelles	8
Figure 2: Exemple de classification de données avec l'algorithme K-Moyennes	10
Figure 3: Exemple de dendrogramme de clustering hiérarchique.....	11
Figure 4 : Vue sur différents algorithmes d'apprentissage automatique.....	12
Figure 5 : Architecture générale d'un réseau de neurones artificiel.....	13
Figure 6 : Relation entre NLP et IA	14
Figure 7 : Exemple de Pipeline pour prétraitement de texte en NLP	15
Figure 10 : Architecture CBOW et Skip-Gram de Word2Vec	19
Figure 11 : Architecture PV-DBOW et PV-DM de Doc2Vec	20
Figure 12 : Vue sur l'architecture d'un Transformer	20
Figure 13 : Exemple d'incorporation de phrases avec BERT	22
Figure 14 : Etapes de modélisation de sujet avec BertTopic	28
Figure 15 : Représentation graphique du modèle LDA	29

LISTE DES CAPTURES

Capture 1 : Exemple de script pour la vectorisation OHE	16
Capture 2 : Exemple de script pour la vectorisation BoW	17
Capture 3 : Schéma de la base de données.....	34
Capture 4 : Proportion des articles selon la source et la thématique	35
Capture 5 : Proportion des articles selon la source et la thématique	35
Capture 6 : Top 4 des sujets obtenus avec BertTopic	37
Capture 7 : Détermination du nombre de sujet optimal	38
Capture 8 : Script pour la génération des résumés de sujet.....	39
Capture 9 : Script pour calcul de similarité avec Doc2Vec	40
Capture 10 : Variance de similarité dans chaque sujet.....	40
Capture 11 : Courbe de distribution de variance.....	41
Capture 12 : Courbe de distribution de similarité	41
Capture 13 : Répartition des Fake News	42
Capture 14 : Visualisation des clusters articles de la base de données	43
Capture 15 : Catégories de sujets selon la proportion potentielle de Fake News	44
Capture 16 : Représentation graphique du sujet 111.....	47
Capture 17 : Représentation graphique du sujet 1.....	48

LISTE DES TABLEAUX

Tableau 1: Top sept des sujets obtenus avec LDA.....	38
--	----

LISTE DES ÉQUATIONS

Équation 1 : Terme Frequency	17
Équation 2 : Inverse Document Frequency	18
Équation 3 : TD-IDF	18

SOMMAIRE

INTRODUCTION.....	1
Chapitre 1 : Présentation Du Sujet	2
1.1. Problématique	2
1.2. Objectifs du mémoire	2
Chapitre 2 : Introduction aux Fakes News	4
2.2. Les éléments constitutifs des Fake News	4
2.3. Les différents types de Fake News	5
Chapitre 3 : Introduction à l'Intelligence Artificielle et au Traitement du Langage Naturel	9
3.1. Apprentissage automatique.....	9
3.2. L'apprentissage profond	12
3.3. Les applications de l'Intelligence Artificielle	13
Chapitre 4 : Conception du modèle de détection des Fake News	24
4.1. Présentation de la solution.....	24
4.2. Collecte des données : par Web Scraping	24
4.3. Modélisation de sujets	26
Chapitre 5 : Implémentation de la solution	32
5.1. Présentation du jeu de données.....	32
5.2. Prétraitement des données	32
5.3. Analyse exploratoire.....	34
CONCLUSION	51

INTRODUCTION

La question de l'information a toujours été au centre de l'expérience humaine. Depuis la nuit des temps, les hommes ont ressenti ce besoin pressant d'être au fait, d'actualiser l'information dont ils disposent et de comprendre le monde qui les entoure.

Au cours des siècles, les modes d'accès à l'information ont beaucoup évolué. La révolution numérique fait évoluer notre mode de consommation de l'information. Dans chacun des secteurs d'activité qui composent l'espace médiatique, de la presse traditionnelle à la télévision, l'avènement d'Internet est le symbole d'un changement des pratiques informationnelles qui ouvre la voie à un nouvel âge marqué par une multitude de services de presse en ligne et un accès instantané et généralisé à l'information concernant les événements nationaux et internationaux et à la communication entre les hommes.

En revanche, cette surabondance d'informations à disposition sur Internet soulève la question de l'authenticité et de l'honnêteté des acteurs de la communication. Les fausses informations, communément appelées « Fake News », c'est-à-dire des informations mensongères ayant pour but de désinformer le public se sont vulgarisées en se propageant au sein du public, alimentant ainsi la confusion, la désinformation, et la méfiance.

La presse en ligne, qui joue un rôle crucial dans la diffusion rapide de l'information, est particulièrement vulnérable aux Fakes News en raison de la rapidité de publication et du partage sur les réseaux sociaux. Des cas célèbres, tels que l'affaire « Pizzagate »¹, au moment de la campagne présidentielle américaine de 2016 montrent l'impact dévastateur de la désinformation. De fausses allégations relatives à un réseau de trafic sexuel d'enfants présidé par les démocrates et Hillary Clinton elle-même, impliquant un restaurant de Washington D.C, ont été à l'origine d'une attaque violente contre un restaurant innocent : la pizzeria Comet Ping Pong.

Dans le cadre de notre mémoire de fin d'études, nous avons entrepris la mise en place d'un modèle de détection de Fake News qui s'appuie sur les techniques du traitement automatique du langage naturel.

Notre travail est organisé autour de plusieurs chapitres qui sont autant d'axes de recherche associés à la problématique. D'un côté, le premier chapitre évoquera le sujet sous un prisme plus large qui sera le cœur de la problématique. D'autre part, le deuxième chapitre servira d'actualité sur le sujet des Fake News. Le troisième chapitre introduira les concepts et notions de l'Intelligence Artificielle et du Traitement Automatique du Langage Naturel dont certaines techniques, serviront à l'implémentation de notre solution. Enfin, les chapitres quatre et cinq présenteront la mise en œuvre réelle de notre solution. Le chapitre quatre présentera notre solution et les outils utilisés pour son implémentation. Le chapitre cinq présentera les résultats et engagera la discussion sur des perspectives d'amélioration du modèle proposé.

¹ Théorie conspirationniste du Pizzagate – [Wikipédia](#)

Chapitre 1 : Présentation Du Sujet

1.1. Problématique

Le bouleversement du paysage médiatique actuel qu'entraîne la révolution numérique, pose de nombreuses questions. S'il constitue sans aucun doute un progrès, il est à l'origine de la propagation d'informations erronées ou trompeuses, de discours haineux, de violences sur l'ensemble du continent.

Le cas du Nigeria illustre cette problématique, puisque les rumeurs sur le vaccin poliomyélitique ont induit des craintes face à ce dernier, le mettant en accusation de provoquer la stérilité, voire de transmettre le VIH. Ces craintes non fondées ont conduit à un rejet des campagnes de vaccination, nuisant à la santé publique : alors que la polio semblait absente, le pays a connu une flambée de cas entre 2000 et 2005, devenant même un centre de circulation virale en Afrique.

Au Sénégal, le cas qui a fait le plus de bruit est à chercher dans les élections présidentielles de 2019. La désinformation n'a pas épargné les figures politiques les plus en vue, dont l'opposant Ousmane Sonko. Le présumé financement de sa campagne par une entreprise pétrolière est le mensonge le plus médiatisé à ce jour, sans qu'aucune preuve n'ait démontré l'influence de ce mensonge sur les choix électoraux. Toujours au Sénégal, une étude parue en février 2021 évoque la désinformation comme cause de la réticence de certains Sénégalais à se faire vacciner contre la COVID-19. [B1]

Ces tentatives de manipulation de l'opinion publique soulignent la nécessité du caractère authentique de l'information dans un environnement politique et démocratique. Le Sénégal est confronté, comme un certain nombre de pays africains à un manque significatif de dispositifs nationaux efficaces pour la détection des Fake News et la lutte contre leur propagation. Il conviendrait alors de doter le pays de mécanismes de détection précoce et de prévention des Fake News, qui ne mettent pas en péril la crédibilité de l'information et surtout qui permettent aux citoyens d'accorder à l'information et à la prise de position des médias et des institutions, un minimum de confiance.

1.2. Objectifs du mémoire

L'objet de ce mémoire est de concevoir un modèle de classification des articles de la presse sénégalaise en ligne pour la détection des fausses informations. Ce travail s'articule autour des objectifs suivants :

- Identifier les outils et méthodes adaptés à la conception du modèle ;
- Développer un modèle de détection des Fake News basé sur des techniques de Machine Learning et de Traitement Automatique du Langage ;
- Proposer des recommandations pour l'amélioration du modèle conçu.

Ce mémoire s'inscrit également dans la validation de nos années d'études au cycle ingénieur de Conception Télécom, en Ingénierie des Données et Intelligence Artificielle (InDIA).

1.3. Démarche et méthodologie

Pour atteindre les objectifs définis, une démarche structurée a été adoptée. Les données ont été collectées par Web Scraping depuis des sites de presse en ligne, puis nettoyées et organisées dans un format standardisé (JSON). La modélisation des sujets a été effectuée en utilisant les bibliothèques open source BERTTopic et LDA, combinant des techniques de NLP et de clustering pour identifier les thèmes récurrents des articles. La qualité des résultats a été évaluée à l'aide de scores de cohérence et d'analyses qualitatives. Enfin, les résultats obtenus ont été interprétés pour mettre en évidence les caractéristiques distinctives des fake news.

1.4. Délimitations

Les données utilisées pour la conception du modèle sont issues de la presse en ligne sénégalaise pour l'année 2019 et ont été obtenues par Web Scrapping. Il ne s'agit pas de déployer le modèle proposé mais comme dit plus haut de concevoir un modèle de classification basé sur l'apprentissage non supervisé.

Chapitre 2 : Introduction aux Fakes News

L'avènement du numérique a entraîné le déferlement des Fake News, fausses informations diffusées en ligne ou via d'autres médias, dont la seule visée est d'influer sur les choix politiques de l'électorat, ou de divertir. Les réseaux sociaux et les services de presse en ligne sont entre autres des vecteurs de cette expansion rapide.

Nous examinerons dans ce chapitre divers aspects des Fake News, de leurs définitions à leurs modalités de diffusion, autant d'éléments nécessaires à l'analyse approfondie qui suivra.

2.1. Définition

D'après la définition donnée par le Cambridge Dictionary en 2019, 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.

Elles se diffusent de manière ininterrompue, perturbant et divertissant l'opinion publique. Les Fake News selon [B2] se définissent par leur :

- ⊕ **Volume** : elles sont produites à large échelle. Tout un chacun a bien la possibilité de publier une fausse nouvelle ;
- ⊕ **Variété** : elles ont d'innombrables sources (les rumeurs, les articles satiriques, les critiques fictives, les informations faussement orientées, la publicité trompeuse, le complotisme, les mensonges en politique, etc.) ;
- ⊕ **Rapidité** : elles se diffusent avec une telle vitesse qu'elles échappent aux dispositifs de surveillance.

2.2. Les éléments constitutifs des Fake News

Les Fake News reposent sur quatre éléments principaux [B2] : le créateur ou diffuseur, la cible de la désinformation, le contenu de l'information, le cadre social de la diffusion.

- ⊕ **Créateur/Diffuseur** : les personnes et/ou institutions qui sont à l'origine de la création et de la diffusion des Fake News peuvent être humaines ou non. Dans le cas non-humain, il s'agit de robots conçus et programmés pour générer automatiquement du contenu, interagir avec les utilisateurs à travers les réseaux sociaux, transmettre des rumeurs, des spams, des virus ou des contenus fallacieux.
- ⊕ **Cible de la désinformation** : les victimes des Fake News peuvent être des utilisateurs de médias sociaux ou d'autres plateformes en ligne. Les cibles des faux contenus peuvent changer en fonction des objectifs des faussaires : étudiants, électeurs, personnes âgées, etc.

- **Contenu de l'information** : le contenu de l'information englobe à la fois ses aspects physiques, tels que le titre, le corps du texte, les médias, ainsi que des aspects non physiques tels que l'objectif, les émotions, les sujets abordés, etc.
- **Contexte social** : le contexte social englobe la totalité du cadre d'activité et de l'environnement social où l'information se propage. En postant leurs expériences et leurs interactions au sein de groupes sociaux de pensée similaires, les utilisateurs en ligne contribuent à amplifier la portée des Fake News.

2.3. Les différents types de Fake News

Il existe différents types de désinformation. Dans ce point, il s'agira d'en présenter les principales tout en fournissant pour chacun d'elles des cas concrets.

2.3.1 Désinformation en politique

Les Fake News à caractère politique visent à influencer les opinions des citoyens sur des questions politiques et à perturber les processus démocratiques. Ils peuvent prendre différentes formes, telles que :

- **Mensonges sur les candidats** : la publication de fausses informations visant à nuire à l'image de candidats pour les élections ;
- **Résultats d'élections manipulés** : la diffusion de faux résultats d'élections, créant confusion et méfiance chez les électeurs ;
- **Propagande politique** : c'est la promotion d'opinions biaisées ou de thèses conspirationnistes afin d'aider à promouvoir un objectif politique.

2.3.2 Désinformation en santé

Les Fake News en matière de santé sont particulièrement problématiques, en raison de leurs effets potentiellement néfastes pour l'état de santé du public. Font notamment partie de ces types de fausses informations :

- **La désinformation sur les vaccins** : ce sont des campagnes de désinformation basées sur des éléments faux destinés à décourager la population d'utiliser un vaccin. Ce qui entraîne une baisse de la couverture vaccinale ;
- **Les remèdes miracles** : la promotion de traitement, de solutions incertaines ou de méthodes miraculeuses sans fondement scientifique contre des maladies graves ;
- **La désinformation sur les épidémies** : elle se caractérise notamment par la circulation de rumeurs et de théories du complot, ce qui nuit aux efforts de santé publique.

2.3.3 Les Fake News de divertissement

Les Fake News de loisir sont souvent créés dans un but humoristique ou d'attractivité, il s'agit plus particulièrement :

- **D'articles satiriques** : la rédaction d'articles satiriques destinés à divertir, mais parfois mal interprétés comme de véritables informations.
- **De buzz** : la promotion de fausses informations concernant des personnalités publiques est une stratégie couramment utilisée dans le but d'attirer une audience massive, discréder une figure influente, ou de générer des profits financiers.

2.4. Les mécanismes de propagation des Fake News

Nous aborderons les principaux mécanismes de propagation des Fake News plus particulièrement ceux concernant la presse en ligne :

2.4.1 Titres sensationnels et accrocheurs

Il s'agit de titres d'article conçus pour susciter un fort intérêt ou une forte émotion chez les lecteurs. Ce type de titre fait souvent appel à des mots ou des phrases provocateurs ou choquants. L'utilisation de tels titres est courant dans les médias en ligne pour inciter les utilisateurs à cliquer sur un article en particulier. Ils contribuent à la propagation des Fake News en créant une curiosité intense et en encourageant les lecteurs à partager l'information sans vérification.

2.4.2 Techniques de référencement

Les techniques de référencement permettent par exemple d'améliorer la visibilité d'un article dans les résultats de recherche des moteurs de recherche tel que Google. Les producteurs de Fake News peuvent utiliser des méthodes de référencement pour mettre leurs articles en haut des résultats de recherche lorsque les utilisateurs recherchent un thème précis. Cela sans doute augmente la probabilité que les Fake News soient découvertes et partagées.

2.4.3 Les commentaires et interactions du lecteur

Les commentaires et interactions des lecteurs, tels que les likes et les partages exercent une force active qui contribue à la diffusion des Fake News. En réagissant positivement ou négativement à un article, le lecteur va participer à sa diffusion et accroître sa portée. Un commentaire peut renforcer la crédibilité des Fake News si plusieurs utilisateurs partagent le même avis et donnent l'impression d'un faux consensus. Les commentaires peuvent donner l'impression que l'opinion publique est favorable à une Fake News, ce qui encourage davantage de partages.

2.4.4 Les Réseaux sociaux

Ils sont au cœur du processus de partage des articles en ligne. En effet, les utilisateurs transmettent souvent les articles sur les réseaux sociaux comme Facebook, Twitter et WhatsApp. Les Fake News sont diffusées en exploitant ces réseaux, car les articles trompeurs peuvent s'y propager rapidement grâce aux partages en ligne.

2.5. Approches pour la détection des Fake News

La détection des Fake News représente un enjeu crucial qui mobilise une gamme variée d'approches et de méthodologies. De nombreuses techniques ont été développées pour identifier les informations fallacieuses, regroupant à la fois des méthodes classiques et des stratégies innovantes issues de l'apprentissage automatique. Voici un résumé des principales approches utilisées.

2.5.1 Approches basées sur les caractéristiques textuelles

Ces méthodes se concentrent principalement sur l'analyse des contenus textuels des articles. Elles utilisent des techniques de Traitement Automatique du Langage Naturel (TALN) pour extraire diverses caractéristiques du texte telles que les mots-clés, les entités nommées, la structure des phrases, ainsi que les émotions exprimées.

2.5.2 Approches Fondées sur les Métadonnées

En sus du texte, les métadonnées des articles, telles que la provenance, la date de publication, les hyperliens et les mentions afférentes au nom de l'auteur, peuvent parfois donner des indications pour évaluer un indice de confiance dans l'information. Les méthodes axées sur les métadonnées mettent en évidence des éléments de suspicion ou des sources récurrentes à l'origine des Fake News.

2.5.3 Approches Axées sur la Vérification des Faits

La vérification des faits ou « *Fact-checking* » repose sur le travail d'organisations spécialisées qui confrontent les informations trouvées dans les articles à des validations antérieures. Si une information contredit des données établies, elle est alors considérée comme suspecte. Conjointement, le *crowdsourcing* constitue une approche complémentaire pour faire appel au public afin de contribuer à détecter, évaluer ou corriger une information douteuse. Les principaux avantages de cette approche sont la rapidité de réaction et la démocratisation du processus de vérification des faits.

2.5.4 Approches basées sur l'apprentissage automatique

Au cœur de la détection des Fake News, on trouve l'apprentissage automatique **supervisé** ou **non supervisé**. En apprentissage supervisé, plusieurs algorithmes d'apprentissage automatique comme les arbres de décision, les réseaux de neurones, ou les machines à vecteurs de support (SVM), sont utilisés pour construire des modèles permettant de classifier des articles en articles vrais ou faux. L'apprentissage non supervisé, lui, exploite des caractéristiques intrinsèques des articles, fondées par exemple sur la similarité sémantique des articles, ou selon la fréquence de mots clés ou d'autres signaux. L'intérêt de l'approche non supervisée se fait particulièrement sentir lorsque l'on peine à collecter des données étiquetées. Sa logique étant que les Fake News présentent souvent suffisamment d'éléments susceptibles d'être exploités pour les distinguer du contenu d'information légitime, et ce, par de l'analyse de contenu, sans avoir à s'interroger sur l'étiquetage de la donnée.

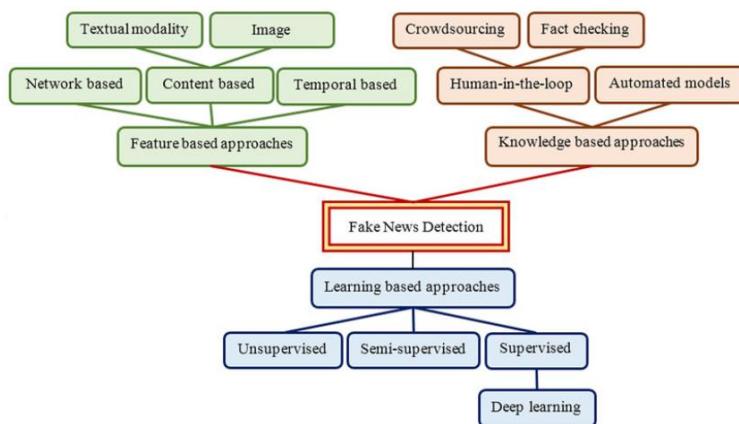


Figure 1 : Approches de détection des fausses nouvelles

Conclusion

Ce chapitre a tracé un panorama des Fake News, des mécanismes de leur propagation aux différentes approches de détection. Alors que les Fake News continuent de se propager à travers les médias en ligne et les plateformes de médias sociaux, il est devenu essentiel de mettre en place des systèmes de détection efficace.

Chapitre 3 : Introduction à l'Intelligence Artificielle et au Traitement du Langage Naturel

L'Intelligence Artificielle (IA) et le Traitement Automatique du Langage Naturel (TALN) jouent un rôle essentiel dans la détection des fausses nouvelles. Dans ce chapitre, nous allons nous plonger dans les bases de l'IA et du TALN, en soulignant leur importance dans l'analyse des informations en ligne et leur rôle essentiel dans la lutte contre les fausses nouvelles.

A. L'Intelligence Artificielle

L'IA est définie comme « un ensemble de théories et des techniques mises en œuvre pour construire des machines pouvant simuler l'intelligence humaine ». C'est une discipline à part entière depuis les années 1950. Les avancées en Machine Learning (ML) et Deep Learning ont permis de réaliser d'importants progrès ces dernières années.

3.1. Apprentissage automatique

L'apprentissage automatique ou Machine Learning est un ensemble de méthodes qui consiste à apprendre aux machines à tirer des enseignements des données et à s'améliorer avec l'expérience sans d'être explicitement programmées pour le faire². Un algorithme est entraîné sur de grandes quantités de données, afin d'en identifier des motifs et d'être en mesure de produire des prédictions sur de nouvelles données. On distingue trois grands types d'apprentissage automatique : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement.

3.1.1 L'apprentissage supervisé

En apprentissage supervisé, les données d'entraînement sont étiquetées avec les réponses attendues (catégories, valeurs), c'est une des méthodes fondamentales de l'apprentissage par induction, qui a pour but la création automatisée d'un classifieur qui apprend à partir d'exemples déjà classifiés ou étiquetés, d'où son nom « supervisé », car le modèle de classification s'entraîne sur les classes cibles et sur les caractéristiques qui les définissent. L'apprentissage supervisé concerne essentiellement deux types de problèmes : la classification (ex. détection de spams, classification d'image, prédiction de Churn...) et la régression (ex. prédiction de prix...). L'apprentissage supervisé se compose de deux phases bien distinctes : la phase d'apprentissage et la phase de test :

- **Phase d'apprentissage** : également appelée phase d'entraînement, il s'agit d'une phase au cours de laquelle un modèle est construit via la mise en œuvre d'un algorithme prenant en entrée un ensemble d'exemples de données contenant les informations nécessaires pour caractériser le problème considéré.

² D'après l'américain [Arthur Samuel](#), l'un des pionniers de l'IA.

- **Phase de test :** au cours de la phase de test, le modèle ainsi généré est évalué au regard de sa capacité à prédire l'étiquette d'un nouvel exemple en fonction des valeurs de ces caractéristiques d'entrée.

Les algorithmes de machine learning les plus connus en apprentissage supervisé sont : la régression linéaire, la régression logistique, les arbres de décision, les forêts aléatoires et les machines à vecteurs de support (SVM).

3.1.2 L'apprentissage non supervisé

L'apprentissage non supervisé, ou apprentissage sans enseignant, s'intéresse à la recherche de proximité entre les données d'un ensemble afin de les répartir en clusters ou classes. Pour ce faire, les algorithmes mis en œuvre rapprochent les données présentant des similitudes tout en éloignant celles qui sont les plus éloignées en raison de leurs caractéristiques. Contrairement à l'apprentissage supervisé, cet apprentissage ne nécessite pas d'expert pour guider le processus. L'apprentissage se fait de manière autonome. Le principal objectif consiste à classifier les données suivant leurs proximités ou similarités. Parmi les algorithmes d'apprentissage non supervisé, on note le k-moyenne (k-Means), le clustering hiérarchique, l'Analyse en Composantes Principales (ACP).

3.1.2.1 Algorithme des K-moyennes (K-Means)

L'algorithme des K-moyennes est l'une des méthodes de clustering les plus populaires. Son principe repose sur la partition d'un jeu de données en un nombre prédefini de clusters, chacun caractérisé par un centroïde, ou point central. L'algorithme commence par initialiser aléatoirement les centres de chaque cluster, puis itère pour minimiser la variance intra-cluster. Concrètement, il assigne chaque point de données au cluster dont le centroïde est le plus proche, puis met à jour la position des centroïdes en calculant la moyenne des points dans chaque cluster. Ce processus est répété jusqu'à ce que la convergence soit atteinte. Le principal avantage de l'algorithme des K-moyennes est sa simplicité et son efficacité sur de grandes bases de données.

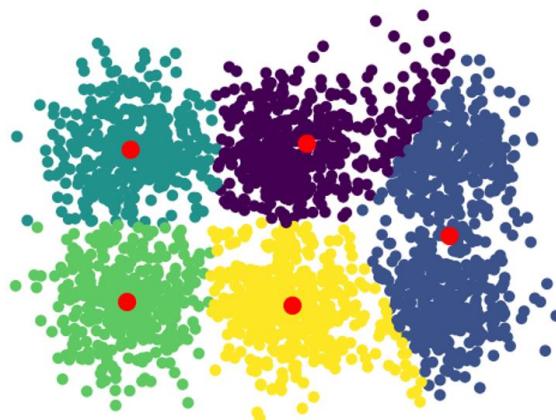


Figure 2: Exemple de classification de données avec l'algorithme K-Moyennes

3.1.2.2 Clustering hiérarchique

Le clustering hiérarchique est une méthode de classification non supervisée qui vise à créer une hiérarchie de clusters organisés sous forme d'arbre, ou dendrogramme. Contrairement à l'algorithme des K-moyennes, il ne nécessite pas de spécifier à l'avance le nombre de clusters. Ce type de clustering peut être divisé en **deux approches principales** : l'approche **agglomérative** et l'**approche divisive**. Dans l'approche agglomérative, chaque point de données commence comme un cluster individuel, et les clusters sont progressivement fusionnés en fonction de leur similarité jusqu'à ce qu'un seul cluster englobant tous les points soit formé. À l'inverse, l'approche divisive commence par un seul cluster englobant tous les points, qui est ensuite scindé en sous-clusters. Le clustering hiérarchique est particulièrement utile pour les ensembles de données où la structure en clusters est inconnue, mais il peut être moins efficace sur de très grands jeux de données, en raison de sa complexité computationnelle.

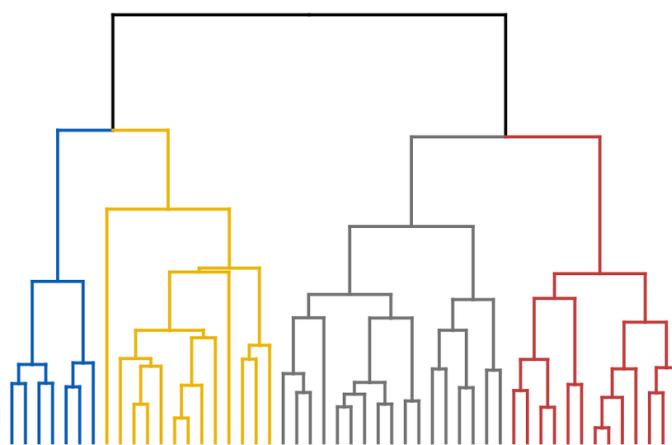


Figure 3: Exemple de dendrogramme de clustering hiérarchique

3.1.2.3 Algorithme HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise)

L'algorithme HDBSCAN est une extension de DBSCAN (Density-Based Spatial Clustering of Applications with Noise) qui introduit des notions de hiérarchie pour permettre un clustering plus flexible. Contrairement à DBSCAN, qui utilise un seuil fixe pour définir les densités, HDBSCAN utilise une approche hiérarchique qui permet de découvrir des clusters à des densités variables et qui peut s'adapter aux structures complexes des données. En regroupant les points dans des zones de haute densité, il identifie des clusters naturels dans les données, tout en marquant les points isolés comme du bruit. HDBSCAN est particulièrement utile pour les ensembles de données présentant des densités hétérogènes et permet d'éviter de spécifier un nombre de clusters à l'avance, contrairement à l'algorithme des K-moyennes. Son avantage réside également dans sa capacité à ignorer le bruit, rendant ainsi le clustering plus robuste face aux points aberrants.

3.1.3 L'apprentissage par renforcement

L'apprentissage par renforcement est une méthode conçue afin de résoudre des problèmes complexes où l'objectif à atteindre est indiqué, mais le système doit apprendre par itération et erreur pour y parvenir. Ce type d'apprentissage est axée sur l'acquisition de compétence au fil des expériences avec comme objectif l'optimisation d'une récompense numérique au fil du temps. L'apprentissage par renforcement regroupe trois éléments principaux : l'agent, l'environnement et les actions. L'agent est le décideur ou l'apprenant, l'environnement comprend tout ce avec quoi l'agent interagit, et actions ce que l'agent fait. L'apprentissage par renforcement est effectif lorsque l'agent choisit des actions maximisant la récompense attendue. Un algorithme classique d'apprentissage par renforcement est le Q-Learning.

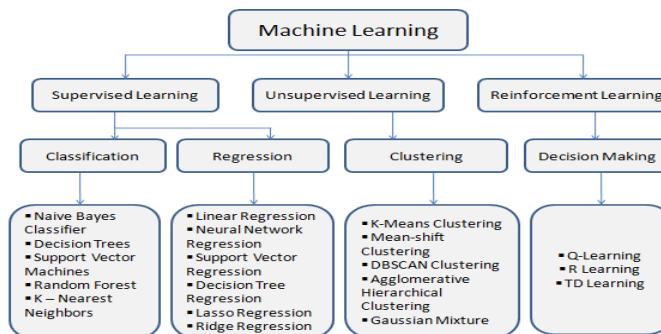


Figure 4 : Vue sur différents algorithmes d'apprentissage automatique

3.2. L'apprentissage profond

Le Deep Learning ou apprentissage profond, est l'un des volets de l'apprentissage automatique qui s'appuie sur des réseaux de neurones artificiels constitués de plusieurs couches. Ces réseaux de neurones artificiels se basent sur le fonctionnement des neurones biologiques et comptent plusieurs couches : une couche d'entrée, des couches cachées, et une couche de sortie. À cela s'ajoute un ensemble de poids reliant les neurones entre les différentes couches et d'une fonction d'activation au sein de chaque neurone. Un algorithme de rétropropagation y est utilisé pour effectuer les modulations nécessaires sur les poids des neurones, de manière à minimiser les erreurs existantes entre les prédictions du modèle et les vraies étiquettes du corpus.

Les réseaux de neurones profonds ont touché tous les domaines d'application de l'intelligence artificielle, affichant des résultats importants en reconnaissance vocale, en reconnaissance visuelle, en traduction automatique, etc. En exemple, on peut citer les modèles de réseaux de neurones suivants : Convolutional Neural Networks (CNN) pour le traitement d'images, Recurrent Neural Networks (RNN) pour la traduction automatique et plus largement pour la génération de texte, les Auto-encodeurs pour la réduction de dimensionnalité et les Transformers pour le traitement du langage humain.

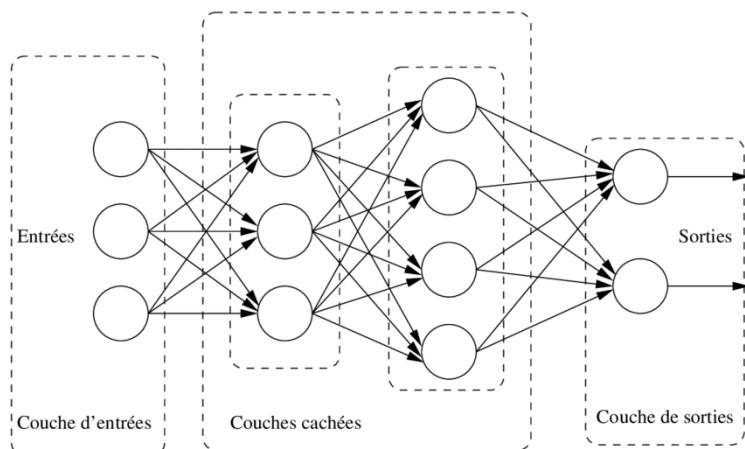


Figure 5 : Architecture générale d'un réseau de neurones artificiel

3.3. Les applications de l'Intelligence Artificielle

3.3.1 La médecine

Les nouvelles technologies basées sur l'IA transforment la médecine, car elles facilitent des diagnostics plus rapides et plus sûrs. Les applications d'IA traitent des images médicales, comme les scanners et radiographies, afin d'y déceler des maladies. De plus les robots de chirurgie dotés d'IA permettent aux chirurgiens d'effectuer des interventions délicates avec une précision sans commune mesure.

3.3.2 Les transports

L'IA permet de concevoir des véhicules autonomes, des drones de livraison, des systèmes de gestion du trafic. Tout cela permet de réduire les accidents sur la route et de diminuer les rejets de gaz à effet de serre. Elle offre aussi des solutions plus adaptées et plus écologique à la mobilité.

3.3.3 Le service client

Les chatbots ou agents virtuels d'IA permettent un service client 24h/7j. Ils sont capables de répondre aux questions des clients, de résoudre des problèmes habituels et d'orienter les utilisateurs vers les informations dont ils ont besoin. L'IA permet de rendre les centres d'appels plus efficaces et surtout de réduire les temps d'attente.

B. Traitement Automatique du Langage

Le Traitement Automatique du Langage Naturel (TALN), est un sous-domaine de l'IA, qui se consacre à l'analyse et à la production du langage humain par des ordinateurs. Il est adapté à toutes les langues et comprend des domaines tels que la traduction automatique, l'analyse des émotions et la conception de chatbots. Les principales branches du NLP sont la Compréhension du Langage Naturel (NLU), la Génération du Langage Naturel (NLG) et l'analyse syntaxique. Ces branches permettent respectivement d'interpréter le sens des mots, de générer du texte et d'explorer la structure grammaticale des phrases.

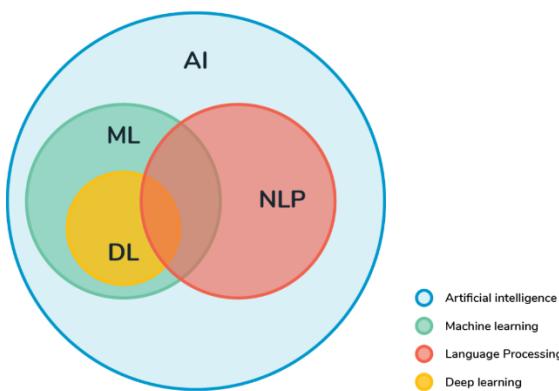


Figure 6 : Relation entre NLP et IA

3.4. Fondement du NLP

Globalement, nous pouvons distinguer deux aspects essentiels à tout problème de NLP : une partie « linguistique », qui consiste à prétraiter et transformer les informations en entrée en un jeu de données exploitable. Et la partie « apprentissage automatique » ou « Data Science », qui porte sur l'application de modèles de Machine Learning ou Deep Learning à ce jeu de données. Nous découvrons plus en détails l'aspect linguistique dans les points qui suivent.

3.5. Le prétraitement des données

L'objet du prétraitement est de préparer et de nettoyer les données textuelles. Il permet surtout de représenter les données textuelles sous une forme exploitable par les algorithmes de l'IA. Cela implique plusieurs opérations :

- ⊕ **La tokenisation :** Il s'agit de la décomposition du texte en unités de travail plus petites appelées tokens. On confond souvent le token avec un mot du texte, mais selon le problème à résoudre, il peut s'agir d'un caractère isolé (une lettre), voire d'une sous-partie du mot.

- **La suppression des stopwords :** Il s'agit de mots couramment utilisés qu'il convient de supprimer afin de garder les mots les plus informatifs et réduire la taille du vocabulaire³.
- **La lemmatisation et le stemming :** La lemmatisation et le stemming sont deux techniques qui permettent également de réduire la taille du vocabulaire. Plus rapide, le stemming consiste à supprimer les terminaisons des mots sans garantir la qualité sémantique des résultats. La lemmatisation quant à elle prend en compte le contexte et fait correspondre les mots avec leur forme de base. Elle garantit donc un meilleur résultat du point de vue sémantique. Cependant, elle est plus coûteuse car reposant sur des tables de recherche permettant d'identifier une forme de base des mots. La lemmatisation est donc à préconiser dans un cas de figure où on priviliege la qualité d'analyse à la rapidité.
- **L'étiquetage morpho-syntaxique :** L'étiquetage morpho-syntaxique, également appelé Part-of-Speech Tagging, consiste à attribuer à chaque mot une étiquette pour définir sa fonction grammaticale (nom propre, adjectif, déterminant, etc.).
- **La reconnaissance d'entités nommées (NER) :** elle permet d'identifier les entités telles que des personnes, des entreprises ou des lieux dans un texte.

Selon la nature du problème, il est possible d'effectuer d'autres opérations de prétraitement. Cela inclut la suppression des chiffres, de la ponctuation, des symboles, etc.



Figure 7 : Exemple de Pipeline pour prétraitement de texte en NLP

³ Le vocabulaire est l'ensemble des mots ou tokens présents dans un corpus de données textuelles.

3.6. La vectorisation des données

La vectorisation est le processus de codage du texte sous forme d'entiers, c'est-à-dire sous forme numérique pour créer des vecteurs de caractéristiques afin que les algorithmes d'apprentissage automatique puissent comprendre nos données [W2]. Il existe deux principaux types de vectorisation des données en NLP : la vectorisation basée sur la sémantique et la vectorisation basée sur la syntaxe.

3.6.1 Vectorisation basée sur la syntaxe

3.6.1.1 One Hot Encoding

Chaque mot du vocabulaire V est associé à un indice entier, i (de 0 à V-1), et la représentation vectorielle de chaque mot a une longueur V avec des 0 partout, sauf un 1 à l'indice i correspondant au mot. Il s'agit d'une représentation du niveau mot. Cependant, cette méthode présente plusieurs inconvénients :

- Elle ne tient pas compte de la sémantique des mots. Par exemple, nous savons que les représentations vectorielles de "garçon" et "garçons" devraient être proches (petite distance euclidienne ou score de cosinus proche de 1). Mais dans le cas du OHE, chaque token est à la même distance de tout autre token.
- La représentation vectorielle est éparse (la plupart des valeurs sont 0), ce qui peut entraîner des inefficacités computationnelles. Comme la plupart des corpus ont des vocabulaires immenses, cela peut poser un gros problème.
- Un problème de représentation de longueur fixe par phrase se pose. Si une phrase comporte 10 tokens et une autre en comporte 9, leur représentation vectorielle n'a pas la même taille, car chaque token a une longueur V. Cela peut poser problème lors de la formation de modèles avec ces données.
- Enfin, la dimension est très élevée, car chaque mot a une longueur de vecteur égale à la taille totale du vocabulaire.

```
def OHE(text):
    tokens = set(text.lower().split())
    length = len(tokens)
    index_map = {x:index for x,index in zip(tokens,range(length))}
    ohe_matrix = []
    for token in tokens:
        ohe = np.zeros(length)
        ohe[index_map[token]] = 1
        print(token, ohe)
        ohe_matrix.append(ohe)

OHE('he is a good boy but he is naughty')

is [1. 0. 0. 0. 0. 0. 0.]
good [0. 1. 0. 0. 0. 0. 0.]
naughty [0. 0. 1. 0. 0. 0. 0.]
he [0. 0. 0. 1. 0. 0. 0.]
a [0. 0. 0. 0. 1. 0. 0.]
but [0. 0. 0. 0. 0. 1. 0.]
boy [0. 0. 0. 0. 0. 0. 1.]
```

Capture 1 : Exemple de script pour la vectorisation OHE

3.6.1.2 Sac-de-mots (bag of words)

C'est l'une des techniques les plus simples. Elle comprend trois étapes fondamentales : la tokenisation du texte d'entrée, la création d'un vocabulaire (V) et enfin, la construction de vecteurs. Le processus de création du vocabulaire commence par l'extraction des tokens uniques du texte, triés par ordre alphabétique. Chaque mot du vocabulaire V se voit attribuer un indice entier, i (de 0 à V-1). Ensuite, une matrice creuse est générée pour l'entrée en utilisant la fréquence des mots dans le vocabulaire. Chaque ligne de cette matrice creuse représente un vecteur de phrase, dont la longueur correspond à la taille du vocabulaire.

Cette technique ignore l'ordre des mots dans la phrase et ne tient pas compte de leurs similarités. Par exemple elle confond une phrase telle que « le chien mord l'homme » avec « l'homme mord le chien ». Elle est souvent utilisée pour des raisons d'efficacité dans le cadre de tâches de recherche d'informations de grande envergure, telles que les moteurs de recherche.

```
from sklearn.feature_extraction.text import CountVectorizer
import pprint
pp = pprint.PrettyPrinter(indent=4)
cv = CountVectorizer()
sentence = ['he is a good boy but he is naughty', 'that girl is a good basketball player']
bow_rep = cv.fit_transform(sentence)
pp.pprint(cv.vocabulary_)

print("BoW representation for {}: {}".format(sentence[0]), bow_rep[0].toarray())
print("BoW representation for {}: {}".format(sentence[1]), bow_rep[1].toarray())

{'basketball': 0,
 'boy': 1,
 'but': 2,
 'girl': 3,
 'good': 4,
 'he': 5,
 'is': 6,
 'naughty': 7,
 'player': 8,
 'that': 9}
BoW representation for he is a good boy but he is naughty: [[0 1 1 0 1 2 2 1 0 0]]
BoW representation for that girl is a good basketball player: [[1 0 0 1 1 0 1 0 1 1]]
```

Capture 2 : Exemple de script pour la vectorisation BoW

3.6.1.3 TF-IDF (Term Frequency–Inverse Document Frequency)

Cette technique attribue un score numérique à chaque mot pour refléter son importance dans un corpus [W1]. Le score est basé sur deux éléments : la fréquence du mot dans le document (TF) et sa fréquence inverse dans l'ensemble des documents (IDF) :

- **TF (Term Frequency)**

Il peut être compris comme un score de fréquence normalisé. Il est calculé via la formule suivante :

$$TF_{(x,y)} = \frac{\text{La fréquence du mot}(x)\text{dans le document } (y)}{\text{Nombre total de mot dans ce document } (y)}$$

Équation 1 : Terme Frequency

- **IDF (Inverse Document Frequency)**

Il est donné par la formule suivante :

$$IDF(x) = \log \left(\frac{\text{Nombre total de document}}{\text{Nombre de document contenant le mot } (x)} \right)$$

Équation 2 : Inverse Document Frequency

Comme nous l'avons vu ci-dessus, l'intuition sous-jacente est que plus un mot est commun à tous les documents, moins il a d'importance pour le document actuel. Un logarithme est utilisé pour atténuer l'effet de l'IDF dans le calcul final.

Le score final TF-IDF est le suivant :

$$TD - IDF_{(x,y)} = TF_{(x,y)} * IDF(x)$$

Équation 3 : TD-IDF

C'est ainsi que la technique TF-IDF parvient à incorporer l'importance d'un mot. Plus le score est élevé, plus ce mot est important.

3.6.2 Vectorisation basée sur la sémantique

La vectorisation basée sur le sens est une approche qui vise à représenter le texte en tenant compte du sens sémantique des mots et de leurs relations. Contrairement à la vectorisation syntaxique, qui repose principalement sur la fréquence d'apparition des mots, cette méthode s'efforce de capturer le sens et la signification des mots dans un contexte donné. Il existe plusieurs techniques de vectorisation basée sur la sémantique, les plus connues sont les suivantes :

3.6.2.1 Word2Vec

Word2Vec est un algorithme de prolongement de mot (Word Embedding) développé par Google [B9]. Il repose sur des réseaux de neurones à deux couches et cherche à apprendre les représentations vectorielles des mots composant un texte, de telle sorte que les mots qui partagent des contextes similaires soient représentés par des vecteurs numériques proches. Il existe deux architectures principales : Skip-Gram, qui prédit les mots voisins à partir du mot d'origine, et Continuous Bag of Words (CBOW), qui prédit le mot d'origine à partir des mots voisins.

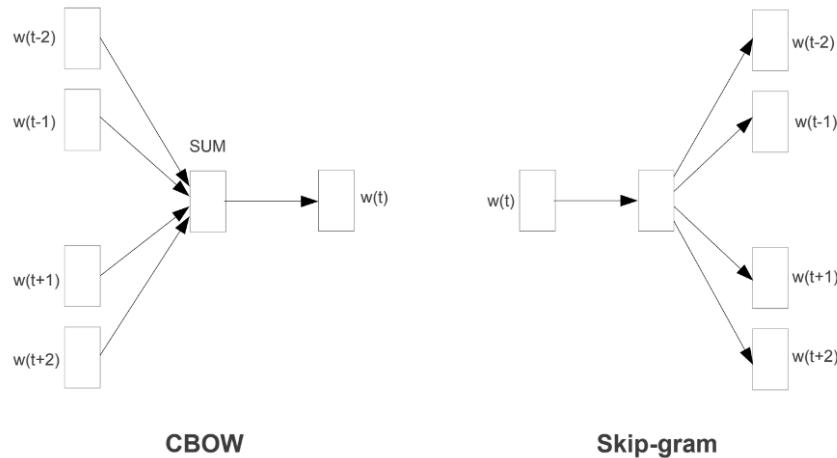


Figure 8 : Architecture CBOW et Skip-Gram de Word2Vec

3.6.2.2 FastText

Il s'agit d'une technique de Word Embedding qui étend Word2Vec en prenant en compte la morphologie des mots [B10]. Contrairement à Word2Vec, qui traite les mots comme des **unités indivisibles**, il crée des représentations de mots en fonction de leurs n-grammes lorsque le mot lui est inconnu. FastText divise alors chaque mot inconnu en sous-unités plus petites appelées « n-grammes »⁴. Par exemple, le mot "chat" peut être décomposé en "cha", "hat", "ch", "ha", "at", "c", "h", "a", "t". Puis, il crée des vecteurs pour ces sous-unités et les combine pour obtenir un vecteur représentatif du mot complet.

3.6.2.3 Doc2Vec

Également connu sous le nom de Paragraph Vector, Doc2Vec [W9] étend les fonctionnalités de Word2Vec en permettant la représentation vectorielle de l'ensemble des documents. Son concept ingénieux s'appuie sur le modèle word2vec en introduisant un autre vecteur : l'ID de paragraphe. Ainsi, lors de l'entraînement des vecteurs de mots W, le vecteur de document D est également entraîné et, à la fin de l'entraînement, il contient une représentation numérique du document. Ce modèle utilise deux architectures principales : PV-DM (Distributed Memory) et PV-DBOW (Distributed Bag of Words). PV-DM maintient un vecteur de contexte partagé pour le document et des vecteurs individuels pour chaque mot, tandis que PV-DBOW se concentre exclusivement sur la prédiction des mots du document en utilisant son vecteur. Doc2Vec offre ainsi une représentation sémantique des documents, ouvrant la voie à des applications telles que la recherche sémantique et la classification de documents.

⁴ Un n-gramme désigne une séquence de n mots ou caractères consécutifs dans un texte donné.

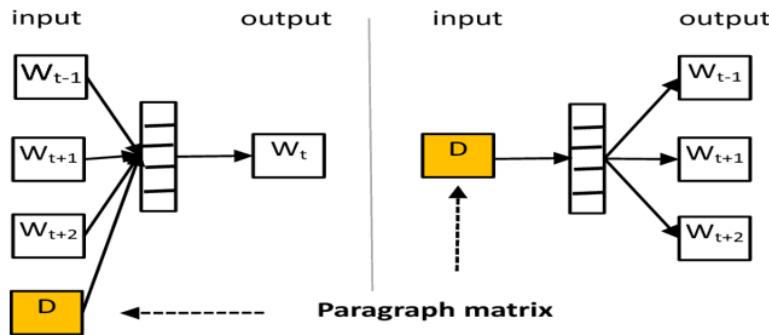


Figure 9 : Architecture PV-DBOW et PV-DM de Doc2Vec

3.7. Les Transformers

Les Transformers [W10] ont révolutionné le domaine du NLP en introduisant une architecture capable de surmonter les limitations des réseaux de neurones récurrents (RNN). Avant leur apparition, les RNN traitaient les mots de manière séquentielle, ce qui posait des problèmes pour les longues séquences, notamment en raison du gradient qui disparaît⁵ et de l'impossibilité de paralléliser les calculs efficacement. Les Transformers, introduits avec l'architecture self-attention, permettent de capturer les dépendances entre les mots d'une séquence indépendamment de leur distance, tout en offrant une parallélisation plus efficace des calculs.

3.7.1 Architecture des Transformers

Les Transformers sont composés de pile d'encodeur et de décodeur. Chacun d'eux étant constitué de plusieurs couches. L'encodeur utilise des couches d'auto-attention et des réseaux entièrement connectés (feed-forward) pour traiter l'entrée, tandis que le décodeur possède une architecture similaire avec une couche supplémentaire appelée d'attention encodeur-décodeur qui capture l'importance des mots de l'encodeur pour générer une sortie contextuelle.

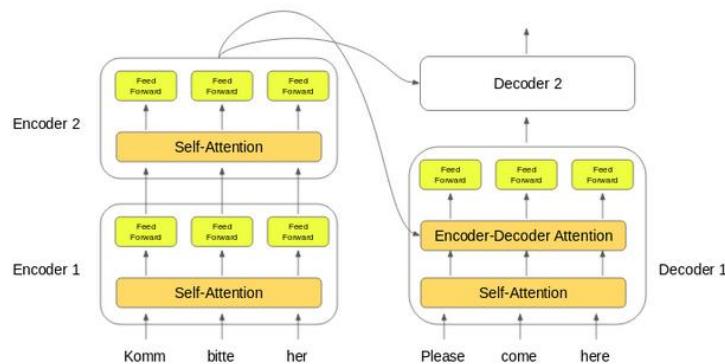


Figure 10 : Vue sur l'architecture d'un Transformer

⁵ Vanishing Gradient : Problème rencontré lors de l'entraînement par rétropropagation dans les réseaux de neurones. Pour plus de détails cliquez [ici](#).

3.7.2 Concepts clés

3.7.2.1 Codage positionnel

Une fois les mots convertis en vecteurs, le codage positionnel introduit des informations sur la position des mots dans la séquence d'entrée à l'aide de fonctions sinusoïdales, permettant au modèle de comprendre l'ordre des mots dans une phrase.

3.7.2.2 Mécanisme d'auto-attention

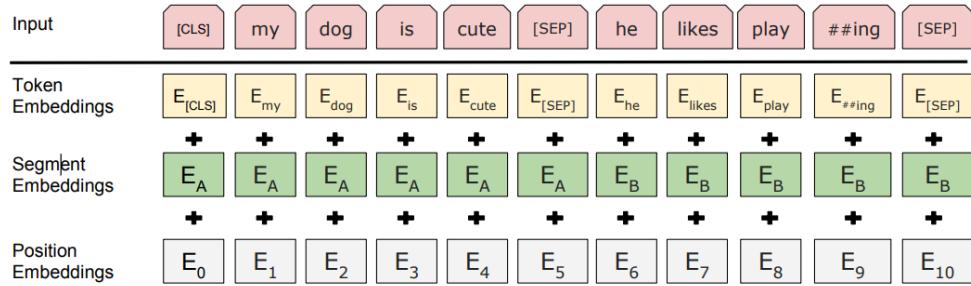
L'auto-attention permet à chaque mot d'une séquence d'interagir avec tous les autres mots pour évaluer leur importance relative. Cela capture les relations complexes au sein d'une séquence. Le mécanisme d'auto-attention permet aux Transformers de se concentrer sur les parties pertinentes d'une séquence, améliorant ainsi la compréhension des relations à longue distance.

Cette approche rend les Transformers extrêmement puissants pour des tâches complexes comme la traduction automatique, le résumé de texte et la classification.

3.7.3 Cas d'un modèle de langage utilisant les Transformers : BERT

BERT, ou *Bidirectional Encoder Representations from Transformers*, est un modèle de langage basé sur l'architecture Transformer n'utilisant que sa partie encodeur [B8]. Il s'agit d'un modèle développé par Google et publié en open source en 2018. L'architecture de BERT se compose de couches d'encodeur bidirectionnelles, permettant une compréhension contextuelle complète des mots dans une phrase, en tenant compte des mots qui précèdent et suivent un mot donné. BERT a été pré-entraîné sur de grands ensembles de données comme Wikipédia et BookCorpus pour accomplir deux tâches principales : la modélisation du langage masqué (*Masked Language Modeling – MLM*), et la prédiction de la phrase suivante (*Next Sentence Prediction – NSP*). Grâce au *fine-tuning*, BERT peut être utilisé pour des tâches spécifiques de NLP comme le résumé automatique de documents, la génération de texte, la réponse à des questions, ou la classification de texte.

Pour faciliter ces tâches, il utilise des jetons spéciaux comme [CLS], qui marque le début d'une séquence et est utilisé pour la classification, et [SEP], utilisé pour séparer des paires de phrases. Ces jetons sont intégrés pendant la phase de pré-entraînement et lors des ajustements spécifiques aux tâches.

**Figure 11 : Exemple d'incorporation de phrases avec BERT**

3.8. Applications du NLP

Le NLP est omniprésent dans notre vie quotidienne, bien que nous ne le remarquions pas toujours :

- **Assistants Virtuels** : Des assistants comme Siri, Cortana et Alexa utilisent le NLP pour comprendre et répondre aux commandes vocales.
- **Traduction Automatique** : Des outils tels que Google Translate se basent sur le NLP pour traduire instantanément des textes entre différentes langues.
- **Analyse des Sentiments** : Les entreprises utilisent le NLP pour analyser les opinions et les réactions des clients sur les réseaux sociaux et les forums.
- **Modération de Contenu** : Le NLP est employé pour filtrer et modérer le contenu en ligne, assurant ainsi un environnement plus sûr sur Internet.

3.9. Machine Learning, NLP et détection de Fake News

Le Machine Learning et le NLP jouent un rôle essentiel dans la lutte contre la propagation des fausses nouvelles. Les algorithmes d'apprentissage automatique, formés sur de vastes ensembles de données, se révèlent efficaces pour repérer divers indicateurs de désinformation présents dans un texte. Par exemple, l'analyse sémantique permet de détecter les incohérences entre le titre trompeur d'un article et son contenu réel. L'extraction d'entités nommées contribue à identifier les personnalités politiques mentionnées ainsi que le contexte associé.

L'apprentissage supervisé et l'apprentissage non supervisé sont deux approches couramment utilisées pour la détection des fausses nouvelles. Néanmoins, l'apprentissage automatique supervisé traditionnel se heurte à divers défis. Les limites consistent en la disponibilité restreinte de données annotées et la nécessité de mises à jour constantes des modèles. L'étiquetage des ensembles de données s'avère souvent chronophage et coûteux et les expériences passées peuvent s'avérer peu prédictives pour les événements futurs.

3.9.1 Détection non supervisée des Fake News

Dans le contexte de la détection non supervisée des fausses nouvelles, les contraintes des méthodes supervisées et semi-supervisées découlent principalement de la nécessité d'avoir des données correctement annotées. Or, dans des situations en temps réel, telles qu'une couverture médiatique soudaine d'un événement critique, l'obtention rapide de données étiquetées peut s'avérer impossible. Pour relever ces défis, des recherches récentes ont exploré des approches d'apprentissage automatique non supervisées, fondées sur l'utilisation de sources fiables, de données utilisateur, de modèles de propagation de l'information, et de comportement entre utilisateurs.

Par exemple, Gaglani et ses collègues [B3] ont mis en œuvre une approche basée sur la similarité sémantique pour détecter les fausses nouvelles sur les réseaux sociaux. Cette méthode non supervisée se concentre sur le contenu, mais ne prend pas en compte le contexte social des médias sociaux. De même, Gangireddy et Coll. [B4] ont proposé une méthode non supervisée basée sur le comportement des utilisateurs, négligeant l'inclusion du contenu dans la modélisation. Li [B5], quant à lui, a utilisé un auto-encodeur dans une approche d'apprentissage automatique non supervisé, en extrayant quatre types de caractéristiques, à savoir le contenu, la propagation, l'utilisateur et les éléments visuels à partir des médias sociaux. Enfin, Jwa et ses collègues [B6] ont utilisé BERT pour détecter automatiquement les fausses nouvelles en analysant le titre et le contenu des articles de presse.

Malheureusement, les recherches sur la détection des fausses nouvelles sur les articles de presse sont limitées. Cette limitation peut s'expliquer par plusieurs facteurs. Tout d'abord, contrairement aux médias sociaux où la désinformation peut être largement répandue par des utilisateurs individuels, les articles de presse sont généralement rédigés par des professionnels du journalisme et suivent des normes de rédaction strictes. Cette nature professionnelle rend souvent difficile l'identification de caractéristiques distinctives pour distinguer les vrais articles des fausses nouvelles. De plus, les articles de presse sont plus longs et détaillés, ce qui nécessite une analyse plus approfondie pour détecter toute forme de manipulation ou de fausse information.

En outre, il existe un manque d'ensembles de données étiquetés spécifiquement pour les articles de presse, ce qui entrave la capacité à former des modèles de détection des fausses nouvelles spécifiques à ce type de contenu. Ces défis rendent essentiel le développement de nouvelles approches et méthodologies pour détecter la désinformation dans ce type de média, garantissant ainsi la diffusion d'informations vérifiées et précises.

Conclusion

Ce chapitre nous a introduit aux concepts fondamentaux de l'IA, du NLP et a souligné leur rôle essentiel dans la détection des Fake News. Nous avons exploré les principes de l'IA, tels que l'apprentissage automatique et l'apprentissage profond, ainsi que ses applications dans divers domaines. En ce qui concerne le NLP, nous avons examiné comment il permet de comprendre le langage humain, en mettant l'accent sur la préparation des données et la vectorisation. Enfin, nous avons abordé le rôle crucial du Machine Learning et du NLP dans la lutte contre la désinformation, en mettant en lumière les défis spécifiques liés à la détection des Fake News dans les articles de presse.

Chapitre 4 : Conception du modèle de détection des Fake News

4.1. Présentation de la solution

La solution proposée pour détecter les fausses nouvelles utilise à la fois des méthodes de classement de similarité et de variance en conjonction avec du Topic Modeling, une technique NLP. La solution comprend deux étapes principales. La première étape consiste à regrouper les documents à l'aide d'une modélisation contextuelle de sujets pour extraire un résumé global de chaque sujet. La deuxième étape consiste à mesurer la similarité entre les articles de presse et le résumé principal au sein d'un même sujet. Grâce au résumé et à la modélisation contextuelle de sujets, les articles pertinents contenant des informations contextuelles similaires sont regroupés. Au sein des articles partageant des informations similaires, leur similarité et leur variance sont mesurées pour identifier les sujets potentiels présentant une plus forte probabilité de contenir de fausses nouvelles.

Sur la base de l'hypothèse selon laquelle les articles de presse sur chaque sujet devraient partager des contenus similaires et que le résumé extrait devrait contenir des informations contextuelles de base, une variance de similarité élevée peut être une indication de différences contextuelles au sein de ce sujet. Enfin, la variance de similarité est utilisée pour classer les sujets qui ont plus de chance de contenir de fausses nouvelles.

4.2. Collecte des données : par Web Scraping

Les données de presse en ligne utilisées seront obtenues par Web Scraping. Le Web Scraping⁶, également appelé extraction de données web, est une technique qui consiste à convertir des données semi-structurées au format HTML en données structurées adaptées à des outils d'analyse ou des bases de données.

Il existe trois modalités principales pour réaliser le web scraping :

- **Extraction manuelle** : l'utilisateur récupère les données à la main en naviguant et copiant le contenu des pages web.
- **Extraction semi-automatique** : un logiciel ou une application web est utilisé pour extraire et nettoyer des éléments spécifiques d'une ou plusieurs pages web.
- **Extraction automatique** : un programme informatique (script) simule un navigateur web pour parcourir les pages, suivre les liens et extraire les données de manière autonome.

4.2.1 Fonctionnement

Dans le cadre d'un processus d'extraction automatique ou semi-automatique, il est nécessaire d'identifier les éléments d'intérêt dans les documents HTML afin de les extraire et de les structurer. Ces principales étapes sont :

⁶ Qu'est-ce que le Web Scraping - [Jedha Bootcamp](#)

Exploration du site

Un robot d'indexation web (web crawler) explore le site cible en effectuant des requêtes HTTP/HTTPS vers des URL spécifiques. Il peut suivre des modèles logiques comme la pagination ou les liens internes. Le contenu HTML obtenu est ensuite transmis à un module d'analyse. Exemple : un crawler démarre à l'adresse <https://www.seneweb.com> et suit les liens de la page d'accueil pour explorer l'ensemble des pages liées.

Extraction des données

Le contenu HTML collecté est analysé à l'aide d'un parseur qui extrait les informations pertinentes sous une forme semi-structurée.

Nettoyage et transformation des données

Les données brutes extraites ne sont souvent pas exploitables immédiatement. Elles nécessitent une transformation ou un nettoyage (normalisation, suppression des doublons, etc.). Des outils comme les expressions régulières peuvent être utilisés pour standardiser les données.

Stockage et sérialisation

Une fois nettoyées, les données sont formatées et stockées dans des bases de données relationnelles (Oracle, MySQL, etc.), des fichiers JSON ou CSV, ou encore transmises à des entrepôts de données pour une utilisation ultérieure.

Dans notre cas figure, le crawler explorera les pages internet de différents médias de presse en ligne sénégalaise que nous lui auront fourni. Par la suite le contenu HTML collecté sera traité en extrayant que les champs que nous avons besoin pour chaque article : les champs tel que le titre de l'article, son contenu, le nombre de partages et de commentaires, la thématique, la date de publication, etc... Enfin l'ensemble des articles collectés seront stockés dans une base de données au format JSON.

4.3. Modélisation de sujets

La modélisation de sujets, ou Topic Modeling [W8], est une technique d'analyse de texte visant à extraire des thèmes à partir d'un ensemble de documents. Son objectif principal est d'identifier les structures sous-jacentes et les sujets qui se dégagent naturellement des documents, sans supervision explicite.

En fonction du modèle utilisé, cette méthode implique plusieurs étapes, notamment la préparation des données par le nettoyage et la transformation du texte, puis la création d'une matrice document-terme pour quantifier la fréquence des mots dans les documents. Enfin, le modèle est appliqué pour identifier les thèmes et attribuer les documents à ces thèmes.

Les méthodes de Topic Modeling peuvent être classées en trois grands groupes en fonction de leurs approches et caractéristiques. Le premier groupe comprend les modèles probabilistes, tels que le Latent Dirichlet Allocation (LDA). Le deuxième groupe utilise des Large Language Models (LLM) tels que BertTopic, GPT, et CamemBERT. Enfin, le troisième groupe se base sur une approche sémantique, à l'exemple du Latent Semantic Analysis (LSA).

Pour notre solution nous utilisons BertTopic et LDA en complémentarité pour le topic modeling. BertTopic exploite la puissance sémantique de BERT pré-entraîné, identifiant finement les topics. Cependant, il peut générer des hors topics (outliers) car BERT peine sur certains textes atypiques. Pour consolider, nous appliquons donc LDA qui grâce à son approche statistique basée sur les fréquences de mots, permettra de regrouper ces outliers en nouveaux topics cohérents. Ainsi, LDA compense les lacunes de BertTopic pour renforcer la qualité du topic modeling final.

4.4.1 BertTopic

BertTopic [W7] est une variante de BERT, développée en 2020 par Maarten Grootendorst pour s'attaquer aux tâches de modélisation thématique. Il s'agit d'une combinaison de techniques qui utilisent des Transformers et la classe TF-IDF pour produire des clusters denses de documents qui sont faciles à comprendre tout en conservant des mots significatifs dans la description du sujet. Cette approche de Deep Learning prend en charge le modèle des transformateurs de phrases pour plus de 50 langues pour l'extraction de l'intégration de documents, elle permet également d'interpréter et de visualiser facilement les sujets générés.

4.4.1.1 Fonctionnement de BertTopic

BertTopic peut être considéré comme une séquence d'étapes pour créer ses représentations de sujets. Ce processus peut être regroupé en trois grandes étapes : l'incorporation de documents, le regroupement de documents, et la représentation de sujet [W7] :



Incorporation les données textuelles

La technique d'incorporation de documents permet de représenter numériquement un document textuel, tel qu'un article, un paragraphe, un commentaire etc. Tout en capturant la signification sémantique et contextuelle du document, elle le transforme en un vecteur numérique.

Dans cette étape, BertTopic extrait des incorporations de documents grâce à S-BERT encore appelé Sentence-BERT, où il peut utiliser n'importe quelle autre technique d'incorporation. Cependant il est essentiel de noter que le choix de la technique d'incorporation peut avoir un impact significatif sur la qualité des sujets générés. Par défaut, il utilise les sentence Transformers suivants :

- ❖ "paraphrase-MiniLM-L6-v2" : Un modèle basé sur BERT en anglais spécialement conçu pour les tâches de similarité sémantique.
- ❖ "paraphrase-multilingual-MiniLM-L12-v2" : Il est similaire au premier, à la différence majeure qu'il fonctionne pour plus de 50 langues.



Regroupement des documents

Avant de regrouper les documents incorporés à la première étape, une réduction de la dimensionnalité est mise en œuvre. Cette réduction de dimension est réalisée à l'aide d'UMAP (Uniform Manifold Approximation and Projection), une technique de réduction de dimension non linéaire qui tient compte des relations complexes entre les données en les projetant dans un espace de dimension inférieure. La réduction de dimension présente plusieurs avantages, notamment l'élimination des caractéristiques redondantes, la réduction du temps de calcul et la facilitation de la visualisation des données.

Ensuite, le regroupement des données s'effectue grâce à l'algorithme de clustering HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). Cet algorithme permet de regrouper les incorporations réduites et de créer des groupes de documents sémantiquement similaires. Il est spécifiquement conçu pour identifier des clusters de manière robuste et efficace dans des données de haute dimension. Contrairement à certains autres algorithmes de clustering qui nécessitent de spécifier à priori le nombre de clusters, HDBSCAN est capable de détecter automatiquement le nombre de clusters présents dans les données.



Création d'une représentation de sujet

La dernière étape consiste à extraire et réduire les sujets en utilisant TF-IDF basé sur des classes (cTF-IDF). Cette méthode permet d'associer des mots clés pertinents à chaque classe c'est-à-dire le sujet auquel chaque document est attribué. Cela garantit une pondération précise des termes en fonction de leur importance pour chaque sujet. Ensuite, pour améliorer la cohérence des mots associés à chaque sujet, on fait appel au Maximal Marginal Relevance (MMR). Cette technique sélectionne les termes les plus pertinents pour chaque sujet tout en évitant la redondance, ce qui améliore la qualité de la représentation des sujets en veillant à ce que les mots choisis soient distincts et informatifs.

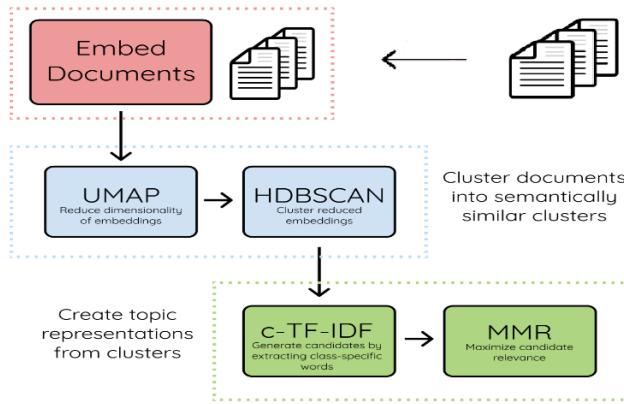


Figure 12 : Etapes de modélisation de sujet avec BertTopic

Bien que ces étapes soient par défaut, il y a une certaine modularité dans BertTopic. Chaque étape de ce processus a été soigneusement sélectionnée de manière à ce qu'elles soient toutes quelque peu indépendantes les unes des autres. En conséquence, BertTopic est assez modulaire et peut maintenir sa qualité de génération de sujets à travers une variété de sous-modèles. En d'autres termes, il permet de construire un modèle propre de Topic Modeling en choisissant à chaque étape le modèle le plus approprié au problème traité.

4.4.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) est une méthode de modélisation de sujets basée sur la probabilité. Elle est conçue pour extraire des thèmes à partir d'un corpus de documents en attribuant des mots à des sujets spécifiques. L'approche LDA considère chaque document comme une distribution de sujets, où chaque sujet est une distribution de mots. Les paramètres de LDA sont estimés à partir des données, ce qui permet d'obtenir une compréhension des sujets qui émergent naturellement dans le corpus. C'est une méthode non-supervisée générative qui se base sur les hypothèses suivantes [W8] :

- Chaque document du corpus est un ensemble de mots sans ordre (bag-of-words) ;
- Chaque document m aborde un certain nombre de thèmes dans différentes proportions qui lui sont propres $p(\theta_m)$;
- Chaque mot possède une distribution associée à chaque thème $p(\phi_k)$. On peut ainsi représenter chaque thème par une probabilité sur chaque mot.
- z_n représente le thème du mot w_n

La modélisation thématique avec l'algorithme de Latent Dirichlet Allocation (LDA) implique trois étapes fondamentales :



Préparation des données

Dans cette première étape, les documents textuels sont soumis à un processus de nettoyage visant à éliminer la ponctuation, les caractères spéciaux et les stopwords. Un dictionnaire de mots uniques est créé pour permettre la vectorisation des données.



Vectorisation des données (Modèle Bag of Words)

Au cours de cette phase, une matrice document-terme est générée pour quantifier la fréquence des mots dans chaque document. Cette matrice est cruciale pour appliquer l'algorithme LDA. Chaque document est transformé en un vecteur multidimensionnel, où chaque dimension représente un mot du dictionnaire. La valeur de chaque dimension est calculée en fonction de la fréquence d'occurrence des mots dans le document. Elle peut être mesurée en termes de fréquence brute ou en utilisant des métriques normalisées telles que le TF-IDF. La dimensionnalité de la matrice peut être réduite si nécessaire pour accélérer le traitement tout en conservant l'information essentielle.



Exécution du modèle

L'étape finale consiste à configurer le modèle LDA en définissant le nombre de sujets que l'on souhaite découvrir. Le modèle attribue à chaque document un mélange de sujets, reflétant ainsi la structure thématique des données. Enfin, on interprète les résultats en extrayant les mots-clés caractéristiques de chaque sujet, permettant ainsi de comprendre les thèmes sous-jacents dans la collection de documents.

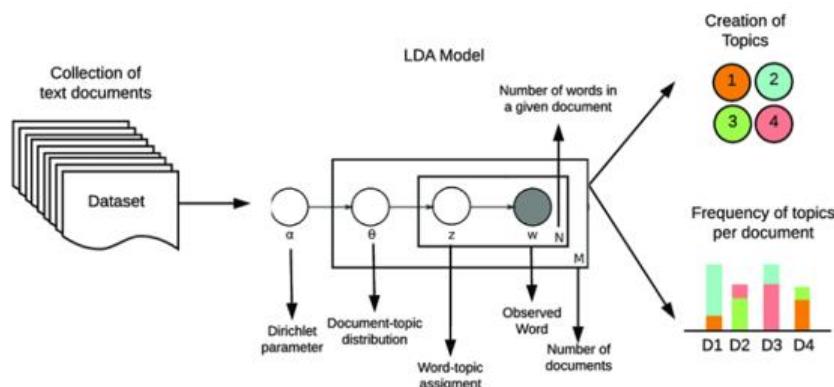


Figure 13 : Représentation graphique du modèle LDA

4.4.3 Le résumé de sujet

Il existe deux principaux types de résumé de texte : le résumé extractif et le résumé abstractif. Le résumé extractif consiste à identifier les sections importantes du texte et à les générer textuellement, ce qui produit un sous-ensemble de phrases à partir du texte original. En revanche, les méthodes de résumé abstractives tentent de saisir le sens global du texte en créant de nouveaux mots et phrases, les combinant de manière significative et en ajoutant les faits les plus essentiels du texte et de ce fait est plus complexe. Pour la création de notre modèle nous utiliserons un algorithme de résumé extractif : BERT Extractive Summarizer (BES). Cette approche garantit que les résumés créés conservent la cohérence sémantique des articles sources améliorant ainsi la qualité de l'extraction.

4.4.3.1 Bert Extractive Summarizer

Le BERT Extractive Summarizer (BES) [W7] est un algorithme de résumé extractif qui repose sur les puissantes capacités de BERT à comprendre les relations complexes entre les mots et les phrases. L'algorithme fonctionne en plusieurs étapes. Tout d'abord, il incorpore le texte d'origine en utilisant BERT, créant ainsi des incorporations pour chaque phrase du document.

Ensuite, il utilise ces incorporations pour évaluer l'importance de chaque phrase dans le contexte du document complet. Les phrases qui portent les informations les plus cruciales sont extraites en fonction de ces scores d'importance, formant ainsi le résumé.

Ainsi pour chaque sujet, BES permettra de créer un résumé concis indispensable pour l'étape suivante qui est la mesure de similarité entre les articles d'un même sujet et le résumé qui lui est associé.

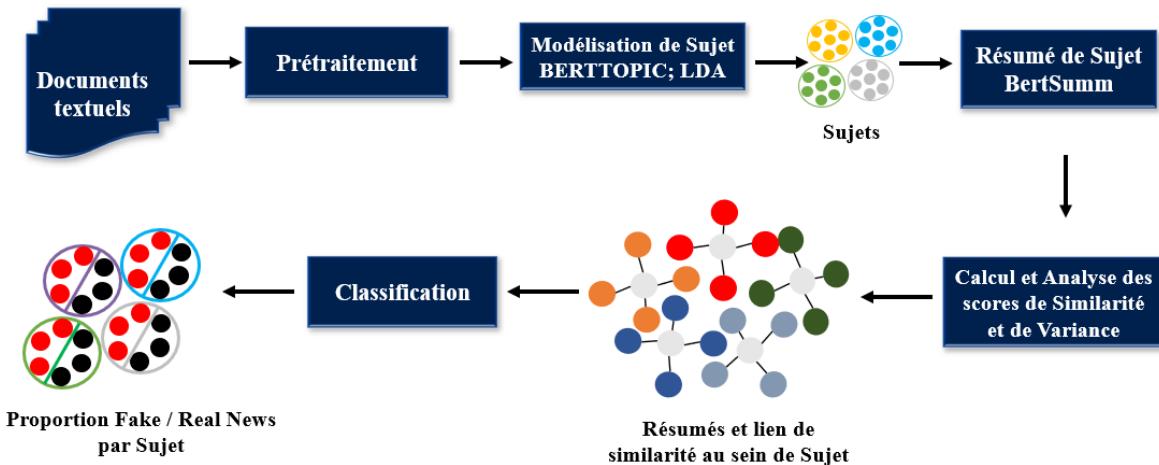
4.4.4 Scores de similarité

Une fois que les résumés, renfermant le maximum d'informations contextuelles, ont été créés, l'étape suivante consiste à évaluer la similarité entre chaque document appartenant à un même sujet et le résumé qui lui est associé. Pour se faire, nous utiliserons la similarité cosinus, une méthode fondamentale qui mesure la similitude entre deux vecteurs en évaluant l'angle cosinus entre eux. Cette approche est particulièrement adaptée pour comparer équitablement des documents de longueurs variables.

Pour la conception de notre modèle, nous adoptons doc2vec pour mesurer les scores de similarité, privilégiant cette méthode en raison de sa pertinence dans la comparaison contextuelle des articles. Par la suite, la variance de similarité est calculée pour chaque sujet, permettant d'émettre des hypothèses sur la diversité des documents au sein d'un sujet. Une variance élevée suggère une diversité élevée, tandis qu'une faible variance indique une similarité accrue. Ces mesures visent à faciliter la distinction entre les fausses informations et informations authentiques au sein de chaque sujet.

4.4.5 Classification des articles

L'objectif est d'attribuer une classe à chaque article de notre jeu de données. Deux classes sont possibles : « pas fake » et « fake ». Ces classes sont assignées après le calcul des scores de similarité et de variance, en fonction de la probabilité qu'un article soit factuel ou non.



Conclusion

Dans ce chapitre, nous avons présenté une approche en deux étapes pour la détection de Fake News, combinant des techniques de clustering et de similarité sémantique. Nous utilisons BertTopic et LDA pour le clustering. Ensuite, nous générerons des résumés contextuels de chaque sujet via BERT Extractive Summarizer. Enfin, le calcul de similarité entre les articles d'un même sujet et le résumé associé permet d'identifier les articles potentiellement fallacieux. Cette combinaison de techniques vise à produire un modèle performant.

Chapitre 5 : Implémentation de la solution

5.1. Présentation du jeu de données

Notre base de données est au format JSON, et pèse environ 31 Mo. La structure du fichier est basée sur une liste d'articles, où chaque article est représenté par un objet JSON. Chaque objet correspondant à un article différent est identifié par un identifiant unique, une URL, et comporte plusieurs champs d'information. Les données incluent des détails tels que le titre de l'article, l'auteur, la date de publication, le nombre de likes, lectures, commentaires et partages, ainsi que la date de collecte de l'article (voir Annexe 1).

Parmi les articles inclus, on peut trouver des sujets variés, tels que des informations sur des événements sportifs, des articles sur des sujets politiques, ainsi que des sujets divers tels que des accidents impliquant des équipes universitaires ou des opérations de police contre le proxénétisme.

Chaque article est également associé à des images d'illustration, bien que certains articles n'en contiennent pas. Il est à noter que certains champs comme la thématique de l'article, la sous-catégorie, la date de mise à jour, les vidéos, les commentaires, les mots-clés, et l'auteur peuvent être vides pour certains articles. Ces articles proviennent de plusieurs sources tels que Dakaractu, Xibaaru, yerimpost, walf-groupe et ont été obtenus par web scraping. On dénombre au total 9073 articles.

5.2. Prétraitement des données

La base de données initiale comportait un ensemble de caractères et de format incorrect ne permettant pas sa lecture par les modules de traitement de fichier JSON en Python. Le prétraitement des données s'est déroulé en deux étapes majeures. Tout d'abord, nous avons entrepris une étape de structuration du jeu de données afin de corriger les erreurs de formatage présentes dans les données. Ensuite, nous avons procédé à une phase d'harmonisation des données dans le but d'assurer la qualité et la cohérence des données traitées. L'ensemble de ces traitements ont été réalisé en Python par usage de la puissante bibliothèque Pandas et d'expressions régulières⁷.

5.2.1 L'extraction et correction des valeurs numériques

Dans certains articles, les champs tels que "nombreLikesArticle", "nombreLecturesArticle", "nombreCommentairesArticle" et "nombrePartagesArticle" présentaient des données incohérentes, y compris des chaînes de caractères et des espaces blancs. Pour résoudre ce problème, nous avons utilisé des expressions régulières pour extraire les valeurs associées à ces champs pour chaque article. Ce processus nous a permis de normaliser les données numériques dans ces champs, garantissant ainsi leur cohérence et leur utilité pour l'analyse et le traitement ultérieur.

⁷ Une expression régulière (ou regex) est une séquence de caractères qui définit un modèle de recherche utilisé pour identifier des correspondances dans du texte.

5.2.2 La suppression des caractères de non-breaking space

Un caractère de non-breaking space est un caractère spécial utilisé dans la typographie et la mise en page pour indiquer un espace qui ne doit pas être rompu par un saut de ligne ou de page. En Unicode, ce caractère est généralement représenté par le code U+00A0. La suppression de ces caractères était importante puisqu'ils étaient inclus dans des valeurs numériques et entraînaient des problèmes d'analyse de nos données.

5.2.3 Harmonisation des Dates

Nous avons harmonisé les valeurs des colonnes 'datePublicationArticle', 'dateMiseJourArticle' et 'dateCollectArticle' pour assurer leur cohérence. Cette harmonisation a permis de normaliser les dates dans ces colonnes vers un format standard ISO 8601 (AAAA-MM-JJHH:MM:SS). Cette opération garantit que toutes les dates dans notre base de données sont désormais uniformes et cohérentes, ce qui facilite la manipulation et l'analyse ultérieure des données.

5.2.4 Traitement des Commentaires

Le champ 'commentairesArticle' a été réorganisé pour remédier à son problème de structure. Auparavant, il contenait une liste de commentaires non structurés sous forme de texte, ce qui compliquait son utilisation pour des tâches de NLP. Désormais, il adopte une structure améliorée avec une liste de dictionnaires de commentaire, représentant individuellement chaque commentaire avec les informations extraites.

5.2.5 Utilisation du champ 'sousCategorieArticle'

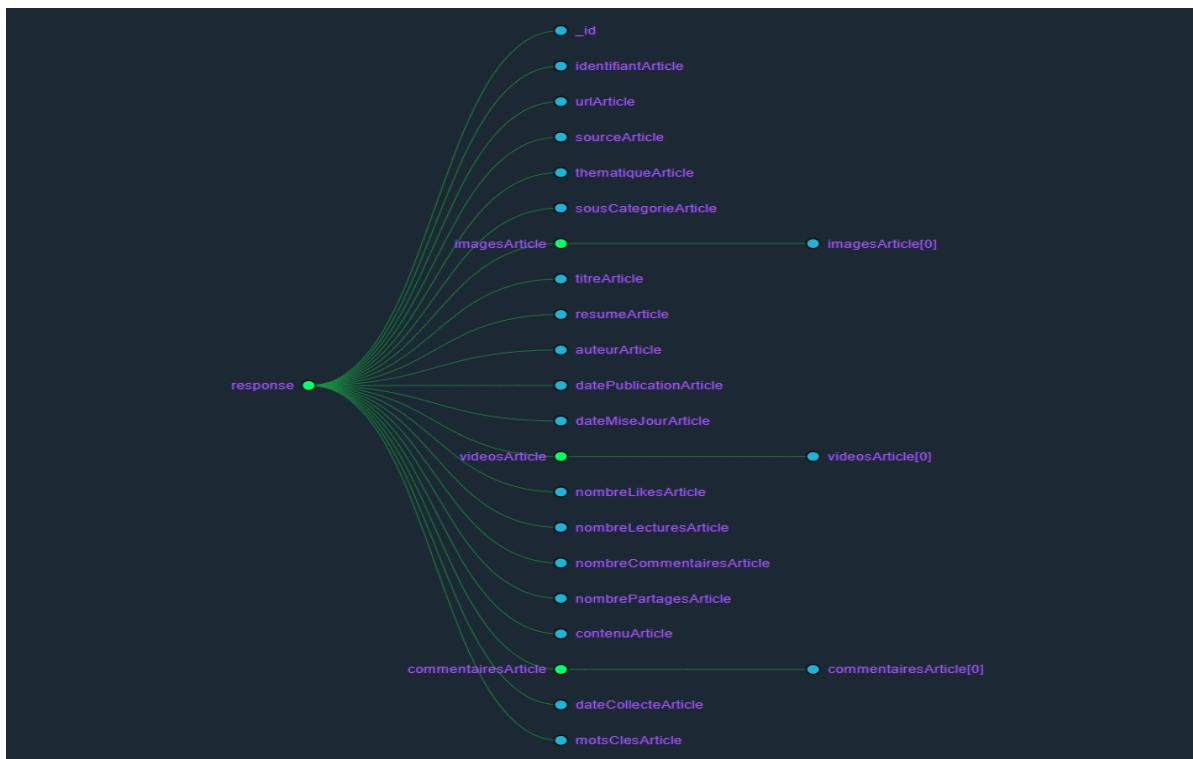
Pour mieux distinguer les articles vidéo des articles texte ou texte et vidéo, nous avons utilisé le champ 'sousCategorieArticle'. Ce champ était vide pour tous les articles, mais il est devenu un moyen utile de classifier nos articles en fonction de leur contenu. Nous avons identifié trois cas principaux :

- **Articles sans contenu texte ni lien vidéo** : Nous avons trouvé 1343 articles pour lesquels le champ 'sousCategorieArticle' est resté vide. Cela signifie qu'ils ne contiennent aucun contenu d'article ni de lien vidéo.
- **Articles vidéo sans contenu article** : Nous avons également trouvé 679 articles vidéo pour lesquels le champ 'sousCategorieArticle' est renseigné à 'video', indiquant qu'il s'agit de vidéos.
- **Articles avec contenu article** : Enfin, nous avons dénombré un total de 7041 articles (77,6% des données initiales) ayant un contenu d'article. Cette catégorie englobe les articles texte uniquement et les articles qui combinent du texte et des vidéos.

Ces distinctions permettront de sélectionner que les articles comportant du texte pour la phase de création du modèle de détection de Fake News, puisque notre approche se base sur le contenu textuel.

5.2.6 Traitement du champ ‘auteurArticle’

Le traitement du champ ‘auteurArticle’ a consisté en plusieurs étapes, afin de nettoyer et normaliser les informations relatives aux auteurs des articles. Ces étapes comprennent la suppression de dates, d’adresses email, de titres d’auteur, et d’informations superflues. Le processus a également inclus la gestion des cas d’auteurs multiples, la mise en majuscules des premières lettres, et la suppression des espaces inutiles.



Capture 3 : Schéma de la base de données

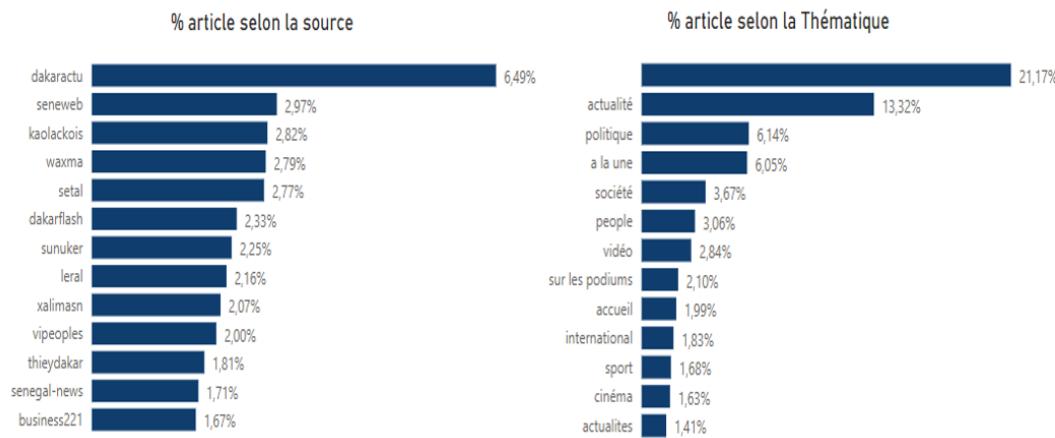
5.3. Analyse exploratoire

L'objectif de cette analyse, comme son nom l'indique, est d'explorer notre jeu de données afin de découvrir ses tendances et ses caractéristiques. Toute analyse de données débute par la phase de formulation des questions, visant à identifier les aspects et les éléments que l'on souhaite observer.

Dans notre cas, nous nous sommes principalement posé quatre questions clés : '*Quelle est la proportion des articles selon leur source ?*', '*Quelles sont les thématiques les plus abordées par les articles ?*', '*Quelles sont les sources d'articles qui attirent le plus de lecteurs ?*', et enfin, '*Quelle est la distribution des dates de publication de nos articles ?*'

5.3.1 Proportion des articles selon leur source et leur thématique

En examinant la figure ci-dessous, nous constatons que la majorité de nos articles proviennent de sites tels que Dakaractu, Senweb, Kaolackois Waxma, etc. Les thématiques les plus abordées comprennent l'actualité à 13,32 %, la politique à 6,14 %, les faits de société à 3,67 % et les sujets people à 3,06 %.

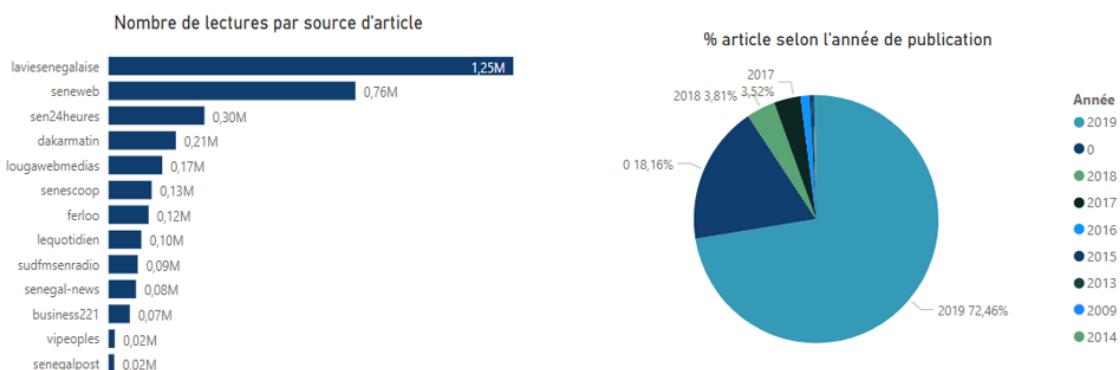


Capture 4 : Proportion des articles selon la source et la thématique

5.3.2 Analyse des sources les plus influentes et de la répartition temporelle des articles

La figure illustre que les sources d'articles les plus attractives sont Laviesenegalaise, avec plus de 1,25 million de lectures, suivie de Seneweb (0,76 million), Sen24heures et Dakarmatin (0,21 million). Ces sources bénéficient d'une audience considérable, les plaçant potentiellement comme des canaux propices à la diffusion de Fake News en raison de leur capacité à toucher un large public.

Concernant la chronologie des publications, la majorité de nos articles ont été publiés en 2019, représentant 72,46 % du total, suivi par ceux de 2018 à 3,75 %.



Capture 5 : Proportion des articles selon la source et la thématique

5.4. Environnement de Travail

Nous avons mis au point notre modèle de clustering sur un système d'exploitation Windows 10, utilisant une machine HP dotée d'un processeur Intel Core i7 cadencé à 2,80 GHz et d'une mémoire RAM de 12,00 Go. Bien que l'utilisation d'environnements Linux soit généralement privilégiée en production pour les applications de Deep Learning, notre contexte nous offrait des ressources de calcul locales suffisantes pour mener à bien toutes nos opérations. Afin d'assurer la stabilité de notre environnement, nous avons choisi d'utiliser Python 3.9. Nous avons également mis en place des environnements virtuels pour prévenir tout conflit potentiel entre les versions des modules de traitement que nous avons employés.

5.5. Logiciels utilisés

5.5.1 Jupyter Notebooks

Pour le développement et la documentation de notre projet, nous avons utilisé Jupyter Notebooks, une application web interactive qui permet d'intégrer du code, des visualisations et du texte explicatif dans un même document.

Il a joué un rôle crucial dans notre processus de travail en nous offrant un environnement convivial pour l'exploration de données, la création de modèles, et la présentation des résultats. Son intégration avec la distribution Anaconda a simplifié la gestion des environnements virtuels et des dépendances, facilitant ainsi le développement.

5.5.2 Néo4j Sandbox

Neo4j Sandbox est une plateforme cloud gratuite permettant de tester et de visualiser des bases de données graphes sans installation locale. Nous l'avons utilisé pour représenter les relations entre les articles et les sujets classifiés. Grâce à son interface intuitive et ses options de personnalisation, il a facilité l'exploration des données et la mise en évidence des résultats de notre modèle.

5.6. Librairies utilisées

5.6.1 Pandas

La librairie Pandas a été au cœur de notre manipulation et analyse de données. Pandas nous a permis de charger, nettoyer et transformer nos données efficacement grâce au dataframe (une de ses structures de données). Ses fonctionnalités pour l'agrégation, le filtrage et la fusion des données ont facilité la préparation des données pour l'entraînement du modèle.

5.6.2 LDA et BertTopic

LDA a été utilisée pour la modélisation de sujets basée sur la probabilité, tandis que BertTopic s'est appuyée sur des modèles linguistiques pré-entraînés pour extraire des sujets. Ces librairies ont grandement contribué à notre capacité à comprendre et à regrouper les sujets dans un ensemble volumineux d'articles.

5.6.3 Spacy et Gensim

Spacy et Gensim ont été des piliers essentiels, contribuant de manière significative à la réussite de notre projet. Spacy a permis le prétraitement linguistique de nos documents. Ses capacités avancées de tokenisation, lemmatisation et suppression des stopwords ont facilité la préparation des textes avant l'application de divers modèles, tels que Doc2Vec et LDA. D'autre part, la bibliothèque open source de traitement NLP en Python, Gensim, a permis la mise en œuvre des algorithmes LDA et Doc2Vec.

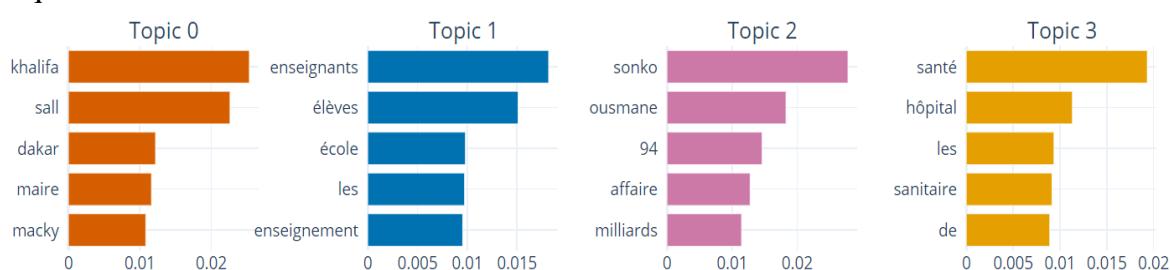
5.7. Modélisation de sujet

5.7.1 BerTopic

La modélisation des sujets avec BertTopic a été réalisée à l'aide de sa bibliothèque dédiée, une puissante ressource basée sur BERT. Un modèle BertTopic a été initialisé en spécifiant la langue française, garantissant ainsi une compréhension appropriée du texte.

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.

Parmi les sujets les plus saillants, on retrouve des thématiques liées à la politique, telles que « *Ousmane Sonko et l'affaire des 94 milliards* », « *les événements liés à la nomination du directeur général de l'IPRES* » et « *l'affaire avec l'ex-maire de Dakar Khalifa Sall* ». Le modèle a également mis en lumière des sujets variés, allant de la santé publique, au football, à l'éducation avec la rentrée académique de 2019, aux télécommunications avec l'entrée de la marque Free dans le secteur.

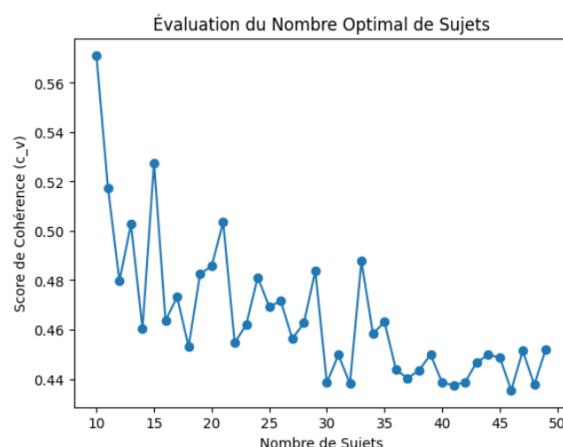


Capture 6 : Top 4 des sujets obtenus avec BertTopic

5.7.2 LDA

Les 1943 articles identifiés comme n'appartenant à aucun sujet ont été soumis au modèle de modélisation de sujet probabiliste LDA. Avant cette étape, une préparation des articles était nécessaire, impliquant la suppression des stopwords, la lemmatisation des mots, et la conversion de l'ensemble du texte en minuscules. En effet LDA étant un algorithme probabiliste, il est sensible au bruit et à la variabilité des données textuelles. Un prétraitement rigoureux était donc essentiel pour optimiser les performances du modèle. Par la suite, la création d'un dictionnaire de données était impérative pour la création du modèle.

Les paramètres du modèle, y compris le nombre optimal de sujets, ont été déterminés en utilisant une méthode de recherche en grille⁸, et la mesure de cohérence : « coherence score » a été utilisée pour évaluer la qualité des sujets. Le modèle lui-même a été construit en conséquence, en tenant compte des paramètres optimums. C'est ainsi qu'avec le modèle optimisé nous avons obtenu 10 sujets avec lesquels le score de cohérence est le plus élevé **0.56** comme le montre la capture ci-dessous.



Capture 7 : Détermination du nombre de sujet optimal

Tableau 1: Top sept des sujets obtenus avec LDA

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 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	économique	aliou
aller	Sall	wade	savoir	religieux	projet	joueur
venir	Contrat	bien	fois	homme	africain	brésil

⁸ La méthode de recherche en grille est une technique d'optimisation qui explore systématiquement un espace de paramètres prédéfini afin de trouver la combinaison optimale selon des critères définis.

5.8. Le résumé de sujet

Comme mentionné au chapitre précédent, le résumé de groupe a pour objectif essentiel de condenser l'essence et le contexte des documents appartenant à chaque sujet identifié. Cette étape cruciale vise à créer une représentation concise et informative des sujets, permettant ainsi une évaluation plus efficace de la similarité entre les articles d'un même groupe et le résumé associé.

Le résumé de sujet s'est fait grâce à l'usage de la bibliothèque Bert Extractive Summarizer. À l'aide de la fonction ***resume_extraction***, chaque sujet est parcouru, extrayant les contenus des articles qui lui sont associés. Ensuite, ces contenus sont agrégés pour former un seul texte représentant l'ensemble du groupe d'articles. Par la suite le nombre optimal de phrases pour le résumé (nb_sentences) est déterminé grâce à la méthode du coude (elbow). Les résumés ainsi créés sont ensuite intégrés aux données, prêts à être utilisés dans la phase ultérieure de mesure de similarité.

```
def resume_extraction(data):
    for i, topic in enumerate(data) : # Pour chaque sujet
        docs = topic['article'] # on récupère le contenu articles associés
        topic_article = ""
        for doc in docs:
            topic_article = topic_article + " " + doc['contenuArticle'] # contenu de tous les articles associés

        model = summarizer() # initialisation du modèle
        nb_sentences = model.calculate_optimal_k(topic_article[:1000000], k_max=10) # determination du nombre de phrases optimales
        print(nb_sentences)
        resume = model(topic_article[:1000000], num_sentences=nb_sentences) # génération du résumé
        data[i][ "resume" ] = resume # ajout du résumé
```

Capture 8 : Script pour la génération des résumés de sujet

5.9. Calcul de Similarité

Le processus de calcul des similarités entre les articles d'un même groupe et le résumé associé est réalisé à l'aide du modèle Doc2Vec. Tout d'abord, les documents textuels de chaque sujet sont tagués et utilisés pour entraîner le modèle Doc2Vec. Ensuite, pour chaque sujet, le contenu de chaque article est comparé au résumé de sujet en termes de similarité cosinus.

Enfin, la variance des similarités entre les articles du même sujet et le résumé est calculée. Les résultats sont intégrés aux données, fournissant ainsi des informations sur la diversité des articles au sein de chaque sujet. Ce processus facilitera la distinction entre les fausses nouvelles et les informations authentiques au sein de chaque sujet, contribuant ainsi à l'objectif global de d'identification des fausses nouvelles.

```

def train_model(docs):
    # Création des documents tagués pour Doc2Vec
    tagged_data = [TaggedDocument(words=nlp(_d.lower()), tags=[str(i)]) for i, _d in enumerate(docs)]
    # Entraînement du modèle Doc2Vec
    doc2vec_model = Doc2Vec(vector_size=100, window=2, min_count=1, workers=4, epochs=200)
    doc2vec_model.build_vocab(tagged_data)
    doc2vec_model.train(tagged_data, total_examples=doc2vec_model.corpus_count, epochs=doc2vec_model.epochs)

    return doc2vec_model

def measure_similarity(doc1, doc2, model):
    token1 = [token.text.lower() for token in nlp(doc1)]
    token2 = [token.text.lower() for token in nlp(doc2)]

    # Obtention des vecteurs de chaque document
    vector1 = model.infer_vector(token1)
    vector2 = model.infer_vector(token2)

    # Calcul de la similarité cosinus
    similarity = cosine_similarity([vector1], [vector2])[0][0]

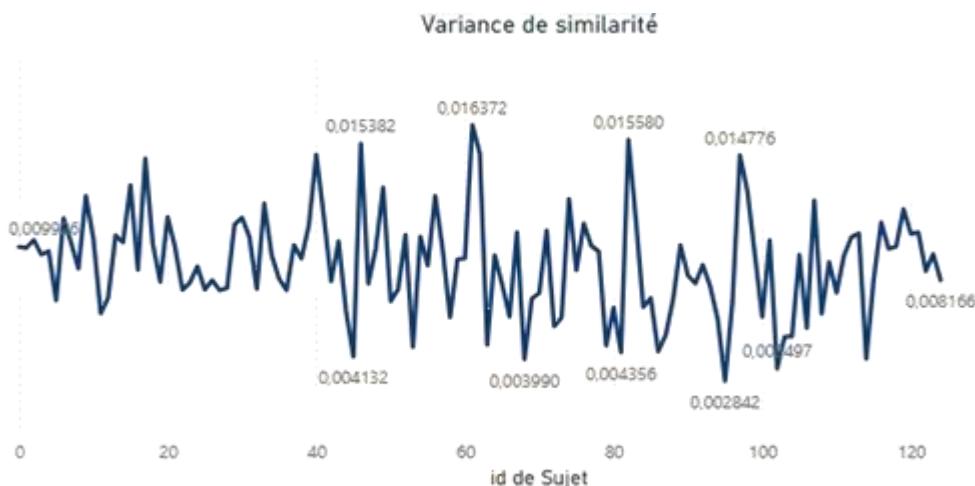
    return similarity

```

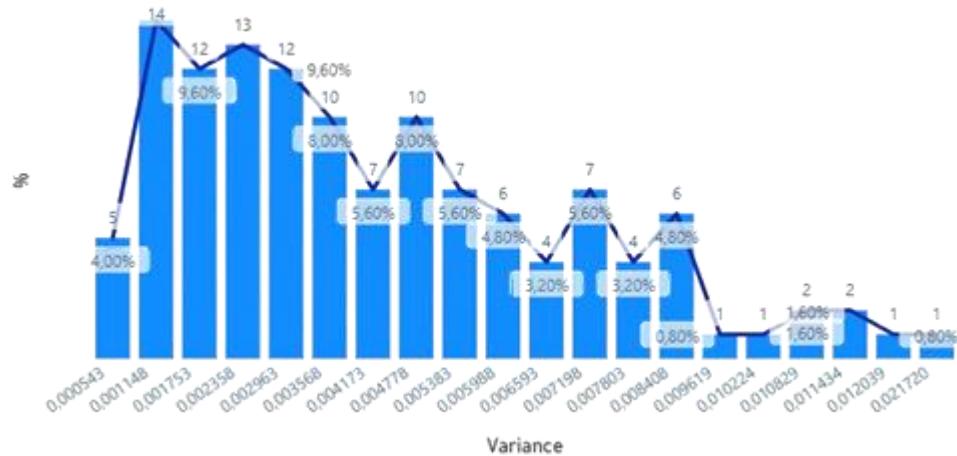
Capture 9 : Script pour calcul de similarité avec Doc2Vec

5.10. Classification des articles

Afin de classifier nos articles, nous procérons à l'analyse des densités de variances et de similarités obtenues à partir de l'étape précédente. Comme illustré dans la figure ci-dessous, la majorité de nos sujets présente une faible variance de similarité. En effet, 80% des sujets ont une variance comprise entre 0,000543 et 0,006593. L'identification des sujets dans cette plage révèle que cela représente 65,1 % de nos articles, indiquant ainsi que la majorité de nos articles abordent les mêmes thématiques au sein des sujets.



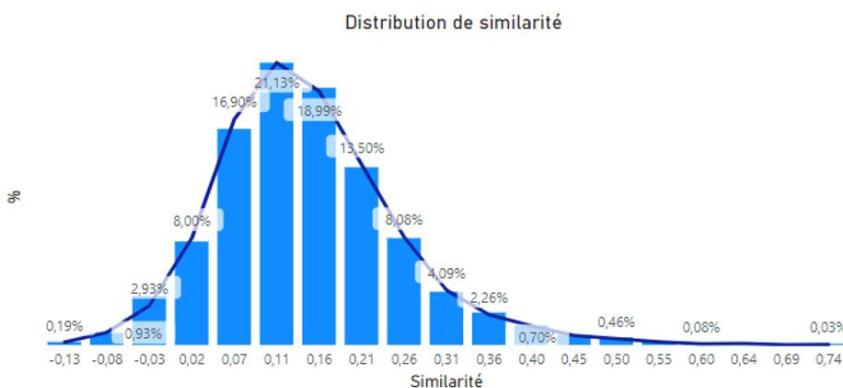
Capture 10 : Variance de similarité dans chaque sujet



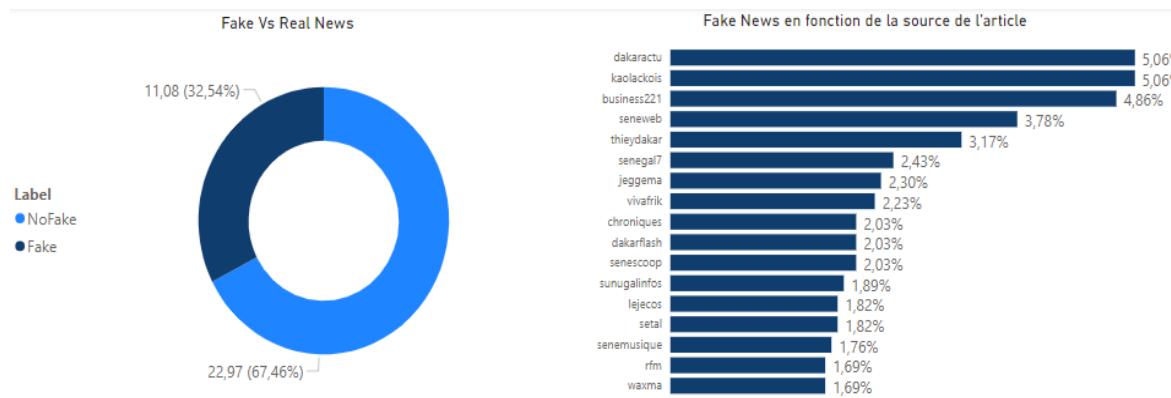
Capture 11 : Courbe de distribution de variance

En ce qui concerne la distribution de similarité, elle présente une forme de cloche centrée au niveau de l'écart type (0,105). La majorité de nos articles (68,05 %) affiche une similarité comprise entre 0,11 et 0,36 ; seuls 3 % d'entre eux dépassent une similarité de 0,36.

Sachant que la plupart de nos articles abordent la même thématique au sein de chaque sujet, d'après l'analyse de la distribution de la variance, et que la majorité de nos articles ont une similarité supérieure à 0,11 ; nous en déduisons que le seuil de similarité pour la classification des articles est fixé à 0,11. Cette approche permet de distinguer les sujets présentant une diversité thématique et des sujets plus homogènes, contribuant ainsi à une labélisation significative des articles en fonction de leur similarité.



Capture 12 : Courbe de distribution de similarité

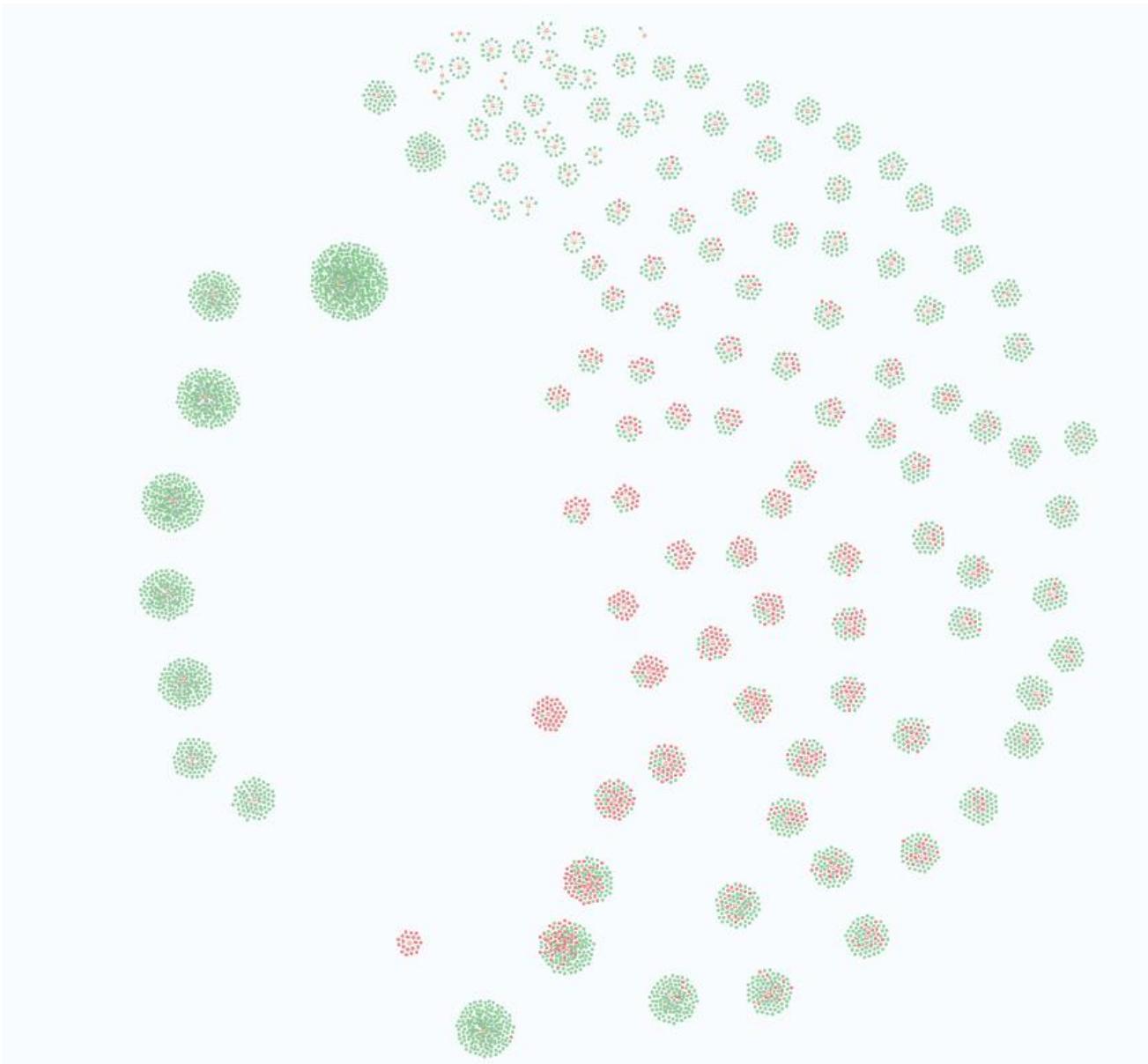
*Capture 13 : Répartition des Fake News*

5.11. Résultats et analyse des articles classifiés

5.11.1 Résultats

La figure ci-dessous illustre différents clusters contenant l'ensemble des articles de la base de données. Ces clusters représentent les 125 sujets ou thématiques identifiés à l'aide des algorithmes de topic modeling BERTopic et LDA. Chaque cluster regroupe un ensemble d'articles cohérents ayant un lien direct avec le sujet abordé.

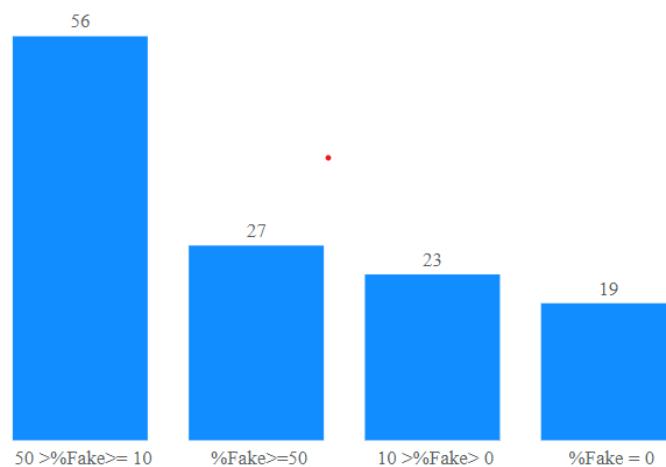
Le degré de similarité entre les articles et leur sujet est évalué sur une échelle allant de -1 à 1, ce qui permet de mesurer leur cohérence. Conformément à la méthodologie décrite précédemment, tout article présentant une similarité inférieure à 0,11 par rapport à son sujet est classé comme une *Fake News*. Sur le graphique ci-dessous, les 125 résumés de sujets sont représentés en jaune, tandis que les articles factuels apparaissent en vert et les articles potentiellement fallacieux en rouge.



Capture 14 : Visualisation des clusters articles de la base de données

5.11.2 Analyse des articles classifiés

Après la classification des articles, nous avons constaté que seuls 19 sujets ont l'intégralité de leurs articles identifiés comme **étant factuels**. Ces sujets seront donc classés dans la catégorie des **sujets sans Fake News** dans la suite de notre analyse. Ensuite, 23 sujets présentent moins de 10 % de leurs articles classifiés comme Fake News. Ces sujets seront regroupés dans la catégorie des **sujets avec peu de Fake News**. Par ailleurs, la majorité des sujets, soit 56, ont entre 10 % et 50 % de leurs articles identifiés comme étant des Fake News. Ceux-ci seront catégorisés dans le groupe des **sujets avec une proportion modérée de Fake News**. Enfin, 27 sujets comptent plus de 50 % de leurs articles classés comme probablement Fake News. Ces derniers seront intégrés dans la catégorie des **sujets avec une forte proportion de Fake News**. Ci-dessous, une illustration montrant la répartition des Fake News par sujet, conformément aux catégories décrites.



Capture 15 : Catégories de sujets selon la proportion potentielle de Fake News

5.11.2.1 Les sujets sans Fake News

Les sujets totalement exempts de Fake News représentent le summum de la crédibilité informationnelle. Il s'agit principalement de domaines **juridiques** et **administratifs** extrêmement formels, tels que les procédures officielles, les documentations administratives et les rapports institutionnels. Les questions de sécurité et de diplomatie internationale constituent un autre domaine où la désinformation est totalement absente. Les relations internationales, les documents officiels et les procédures de contrôle bénéficient d'une traçabilité et d'une vérification absolues.

5.11.2.2 Les sujets avec peu de Fake News

Les domaines présentant peu de Fake News se caractérisent par leur **nature technique et spécialisée**. Les sujets institutionnels, juridiques et diplomatiques dominent cette catégorie, avec une forte traçabilité et vérifiabilité des informations. Les événements culturels, tels que les distinctions, les personnalités culturelles et les événements religieux comme le Magal, semblent moins sujets à la manipulation informationnelle. Cette résilience peut s'expliquer par une communauté plus informée et un rapport plus critique à l'information.

Les aspects économiques et financiers, notamment les budgets, crédits et programmes économiques, présentent également une forte immunité contre la désinformation. La nature chiffrée et documentée de ces sujets limite significativement les possibilités de manipulation.

5.11.2.3 Les sujets avec une proportion modérée de Fake News

Cette catégorie se distingue par un équilibre subtil entre information et désinformation. Les sujets politiques et institutionnels y occupent une place centrale, avec des thématiques autour des processus électoraux, des personnalités politiques comme Macky Sall ou Ousmane Sonko, et des enjeux institutionnels complexes.

Les problématiques de santé et de société émergent également comme des domaines significatifs. Les systèmes de santé, les questions migratoires et les problématiques sociales présentent une vulnérabilité modérée à la désinformation. Cette proportion mesurée suggère que ces sujets, bien que sensibles, bénéficient d'un certain niveau de vérification et de contextualisation.

Les enjeux économiques et de développement, notamment les projets économiques africains, le secteur privé et les financements internationaux, montrent une tendance à la désinformation partielle. La nature technique et complexe de ces sujets permet une propagation limitée mais non négligeable de fausses informations.

5.11.2.4 Les sujets avec une forte proportion de Fake News

Les sujets présentant une proportion élevée de Fake News se caractérisent par une dynamique de désinformation massive et particulièrement agressive. Le domaine sportif domine largement cette catégorie, avec une concentration significative autour des matchs internationaux, notamment ceux impliquant le Brésil et le Sénégal. Les compétitions de football, comme la Coupe d'Afrique des Nations (CAN) et l'Union des fédérations ouest-africaines de football (UFOA), semblent être des terrains particulièrement propices à la propagation de fausses informations.

Les contenus médiatiques numériques constituent un autre vecteur important de désinformation. Les podcasts, clips vidéo et contenus de streaming audio représentent des supports où l'information peut être rapidement déformée et manipulée. La viralité de ces contenus, combinée à leur aspect émotionnel, favorise la diffusion rapide de fausses informations.

La sphère politique et territoriale n'est pas en reste, avec des sujets impliquant des relations entre ministres et préfets, des enjeux de collectivités territoriales et des problématiques énergétiques qui génèrent également un nombre significatif de Fake News. La complexité de ces sujets et leur potentiel émotionnel contribuent à leur vulnérabilité face à la désinformation.

5.11.3 Vue sur le sujet 111 : « louga_plage_potou_morgue »

L'ensemble des onze (11) articles du **sujet 111** converge autour d'un événement tragique : **la noyade de plusieurs jeunes hommes à Louga, au Sénégal**. Cet événement, largement relayé par les médias locaux, a généré un volume significatif d'articles en l'espace de quelques jours. Parmi les médias ayant couvert cette tragédie figurent notamment **Seneweb, Senescoop, Metro Dakar, Leral, Ferloo, Bestinfos, Thierry Dakar, Senegal7, Le Soleil, Sunugalinfos et Lougawebmedias**.

Ces articles se distinguent par leur forte cohérence narrative. Tous relatent un **événement unifié**, reprenant des détails constants tels que les **âges des victimes, le lieu du drame et l'intervention des sapeurs-pompiers**. La plupart citent **Radio Sénégal** comme source principale, ce qui renforce la crédibilité des faits rapportés. L'absence de contradictions entre les différentes versions de l'événement témoigne également d'une couverture fiable et rigoureuse.

Malgré une base commune, les récits présentent des variations intéressantes. Certains articles apportent des précisions supplémentaires, comme les noms ou les origines des victimes, tandis que d'autres adoptent un angle plus humain, évoquant la tristesse des familles et de la communauté touchée.

Enfin, l'absence d'article potentiellement fausse dans ce corpus est attribuable à plusieurs facteurs : la vérifiabilité de l'événement, la fiabilité des sources médiatiques impliquées, et l'absence de biais ou de tentatives de manipulation.



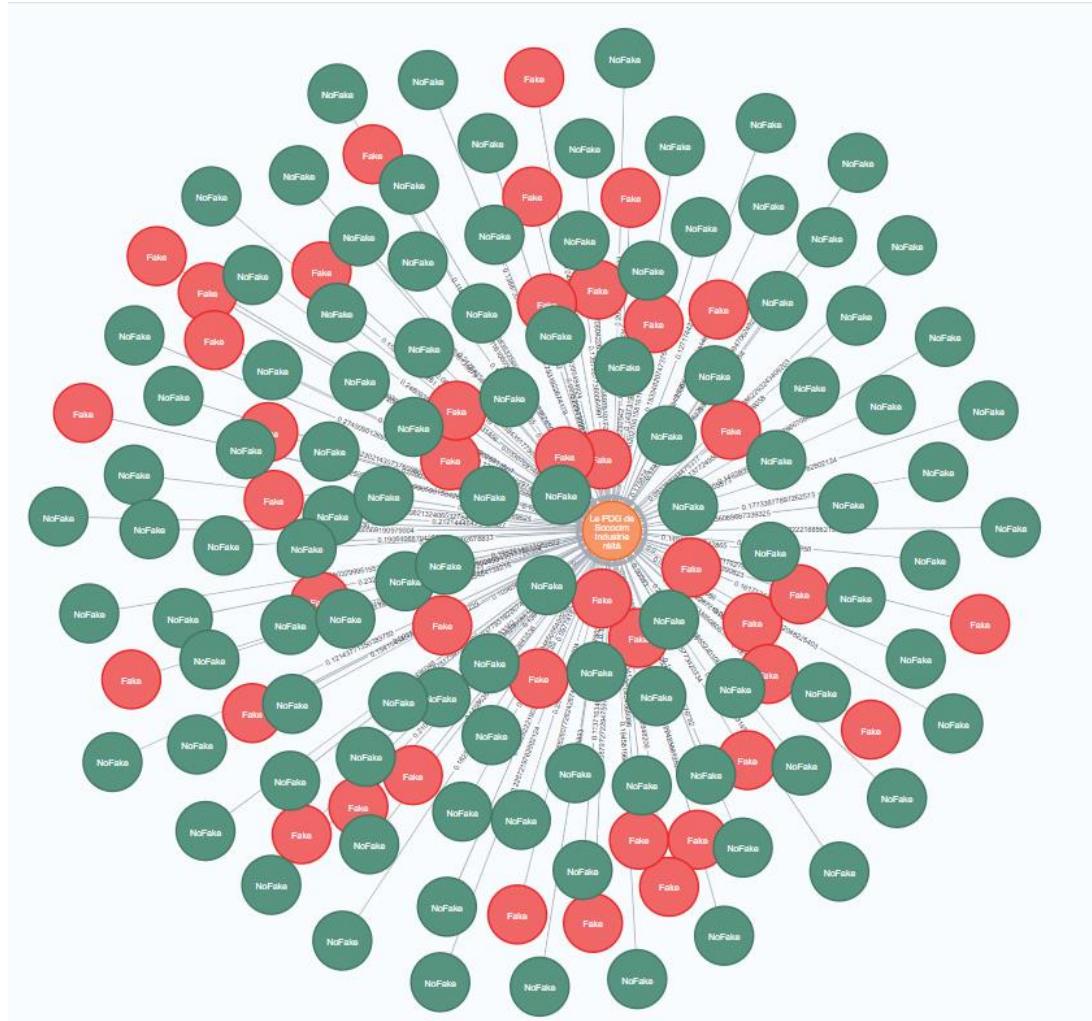
Capture 16 : Représentation graphique du sujet 111

5.11.4 Vue sur le sujet 1 « enseignants_élèves_école_les »

Le sujet 1, regroupant **144 articles**, est indéniablement centré sur l'éducation au Sénégal. Bien que les thèmes abordés soient diversifiés, un fil conducteur lié aux enjeux éducatifs se dégage nettement. Les articles explorent différents angles politiques, sociaux et économiques soulignant ainsi l'interconnexion entre l'éducation et la politique. En effet, les décisions dans ce secteur ont des répercussions majeures sur le plan socio-économique, souvent issues de compromis politiques.

Les articles mettent en lumière plusieurs aspects clés du système éducatif sénégalais. L'accès à l'**éducation constitue** un enjeu central, illustré par l'ouverture de nouveaux établissements, tels que les ISEPS, visant à élargir les opportunités d'enseignement supérieur. Par ailleurs, les **conditions d'enseignement** suscitent des préoccupations récurrentes, notamment en raison du manque d'enseignants, de la vétusté des infrastructures, et des difficultés rencontrées lors des rentrées scolaires. Les **politiques éducatives** occupent également une place importante, avec des réformes gouvernementales portant sur la revalorisation des enseignants et l'ajustement des programmes scolaires. Enfin, la **vie universitaire** est souvent évoquée à travers les mouvements sociaux dans les établissements d'enseignement supérieur et les relations parfois tendues entre les étudiants et les autorités académiques.

Parmi ces 144 articles, **29,17 %** ont été identifiés comme **potentiellement fallacieux**, ce qui classe ce sujet dans la catégorie des thématiques à risque modéré de Fake News. Les articles proviennent de sources variées et reconnues dans le paysage médiatique sénégalais, telles que **Seneweb**, **Senescoop**, **Seneplus**, **SenegalDirect**, **SenegalActus**, **Sen360**, **Leral**, et **Le Quotidien**.



Capture 17 : Représentation graphique du sujet 1

5.12. Limites de la solution

Le modèle probabiliste LDA utilisé pour la formation des sujets présente des limites, car il ne capte pas toujours bien les relations sémantiques complexes. Cette approche peut ainsi manquer de flexibilité face à des articles très courts ou très longs, ainsi qu'à des sujets plus nuancés, ce qui réduit la précision des résultats obtenus.

Ensuite, bien que **l'hypothèse selon laquelle le résumé de chaque groupe d'articles serait factuel**, l'écart par rapport au résumé principal dans un sujet spécifique n'est pas suffisant pour mesurer le « caractère faux » des articles de presse en ligne. Les fausses nouvelles peuvent se manifester par des distorsions subtiles des faits, que l'analyse thématique seule ne parvient pas toujours à capturer. Sur des sujets controversés, par exemple, les fausses informations peuvent dominer, alors que sur d'autres sujets, des informations factuelles sont plus prévalentes.

Par ailleurs, un des inconvénients de la solution proposée est la **nécessité de réitérer l'ensemble du processus** pour classifier de nouveaux articles. Cela implique de reprendre le prétraitement des données, la modélisation des sujets, le calcul des similarités et l'identification des clusters. Cette contrainte peut constituer une limite importante, car elle exige des ressources computationnelles de plus en plus élevées à mesure que la taille des données augmente, ce qui entraîne des temps de calcul prolongés et potentiellement inefficaces.

Enfin, la solution proposée est principalement limitée par **l'indisponibilité d'articles de presse en ligne sénégalais étiquetés (Fake, pas Fake)**. Sans ces données labellisées, il est difficile de garantir une évaluation précise et robuste du modèle, ce qui compromet son applicabilité dans un contexte global.

5.13. Perspectives d'amélioration

⊕ Tout d'abord, une labélisation manuelle d'une partie de la base de données permettrait de disposer de données bien étiquetées pour tester et améliorer le modèle.

⊕ Par ailleurs, l'implication des institutions publiques serait cruciale dans la création de plateformes de fact-checking et d'organismes spécialisés dans la détection des Fake News, fournissant ainsi des jeux de données fiables pour alimenter les modèles de Machine Learning.

⊕ Enfin, la sensibilisation et la formation des populations sur la détection et les effets des Fake News permettraient de lutter activement contre la désinformation à une échelle plus large, complétant ainsi les solutions technologiques.

Conclusion

Ce chapitre a détaillé les différentes étapes nécessaires à l'implémentation de notre solution. Du prétraitement des données au choix des algorithmes de Machine Learning, en passant par leur paramétrage et évaluation, chaque point a été abordé en profondeur. Les limites de notre solution ont été présentées et des perspectives d'amélioration du modèle ont été proposées.

CONCLUSION

L'objectif de ce mémoire était de proposer une approche pour la détection les Fake News dans les articles de presse sénégalaise en ligne. Pour ce faire, nous avons adopté une méthode basée sur l'apprentissage non supervisé, en tirant parti des capacités d'un modèle de langage avancé : BERT.

Nous avons commencé par un prétraitement rigoureux des données pour garantir leur harmonisation et leur qualité. Ensuite, nous avons regroupé les articles en sujets cohérents en utilisant BertTopic, un modèle robuste de modélisation thématique. Face aux limites rencontrées avec certains textes atypiques, l'algorithme LDA fut aussi utilisé. Pour chaque sujet, des résumés ont été générés à l'aide de BertSum, permettant ainsi le calcul de scores de similarité entre les articles. L'écart observé par rapport à ces résumés a servi à évaluer le potentiel de désinformation au sein de chaque groupe de sujets.

Cependant, la présente approche n'est pas exempte de limites. L'écart de similarité, à lui seul, ne permet pas de classifier de manière fiable les articles. Il est en effet possible que le résumé d'un sujet contienne lui-même des informations erronées, ou que des fausses nouvelles soient concentrées autour d'un même sujet. De plus, le manque de données de référence rend difficile une évaluation robuste et précise du modèle.

Plusieurs perspectives d'amélioration peuvent être envisagées. Parmi elles, la labellisation manuelle d'une partie des données, l'adoption d'une approche semi-supervisée s'appuyant sur ces données labellisées, et la participation active d'institutions publiques à la création de plateformes de fact-checking et d'organismes spécialisés dans la détection des Fake News. Ces initiatives pourraient fournir des jeux de données fiables pour alimenter les modèles de machine learning. Enfin, la sensibilisation et la formation des populations sur les Fake News contribueraient à lutter activement contre la désinformation à une plus grande échelle, complétant ainsi les solutions technologiques proposées.

BIBLIOGRAPHIE

- [B1] Institut de recherche pour le développement (IRD) (2021), « Les motifs des réticences vis-à-vis du vaccin anti-covid-19 et les espaces de progression des opinions au Sénégal ». Note 1 CORAF Réticences Vaccin Sénégal.
- [B2] Xichen Zhang, Ali A Ghorbani (2019). « An Overview of Online Fake News: Characterization, Detection, and Discussion ». Information Processing & Management.
- [B3] Gaglani, J., Gandhi, Y., Gogate, S. et Halbe, A. (2020). «Unsupervised WhatsApp fake news detection using semantic search ». 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 285-289). IEEE
- [B4] Gangireddy, S.C.R., Long, C. et Chakraborty, T. (2020). « Unsupervised detection of fake news: a graph-based approach ». Acts of the 31st ACM Conference on Hypertext and Social Media, (pp. 75-83).
- [B5] Li, D., Guo, H., Wang, Z. et Zheng, Z. (2021), « Unsupervised fake news detection based on autoencoder», In IEEE Access, 9, (pp.29356-29365).
- [B6] Jwa, H., Oh, D., Park, K., Kang, JM, and Lim, H. (2019), « exBAKE: Automatic Fake News Detection Model Based on Bidirectional Encoder Representations from Transformers (BERT) ». In Applied Sciences, 9(19), 4062
- [B7] Shin, Yucheol; Sojdehei, Yvan; Zheng, Limin; and Blanchard, Brad (2023) « Content-Based Unsupervised Fake News Detection on Ukraine-Russia War ». SMU Data Science Review: Vol. 7: No. 1, Article 3.
- [B8] Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N; Kaiser, Łukasz; Polosukhin, Illia (2017). "Attention is All you Need". Advances in Neural Information Processing Systems. 30. Curran Associates, Inc. arXiv:1706.03762.
- [B9] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). « Efficient Estimation of Word Representations in Vector Space ». ResearchGate, (pp.1-7)
- [B10] Ercong Nie (2021). « fastText: An improved character-level modeling for word representation ». LMU München (pp.2-5)
- [B11] Mémoire ESMT : Régis Donald Bilélé DAMOUE (2021), « Conception et réalisation d'un système de vidéo analytique : détecteur de véhicules et de foule, lecteur de plaques d'immatriculation et reconnaissance faciale ».

WEBOGRAPHIE

- [W1] [Elsa Negre. Comparaison de textes: quelques approches](#); consulté le 20/10/2023
- [W2] [Vectorization Techniques in NLP \[Guide\]](#) ; consulté le 23/10/2023
- [W3] [Word Embedding & NLP : définition, exemples](#) ; consulté le 16/11/2023
- [W4] [Les Modèles Génératifs \(Partie 3\) | Les Transformers](#) ; consulté le 23/07/2024
- [W5] [Hands-on Guide To Extractive Text Summarization](#); consulté le 02/08/2024
- [W6] [Topic Modeling in Python: Latent Dirichlet Allocation \(LDA\)](#); consulté le 03/08/2024
- [W7] [Bert Extractive Summarizer GitHub Repository](#); consulté le 03/08/2024
- [W8] [Modélisez des sujets avec des méthodes non supervisées](#) ; consulté le 03/08/2024
- [W9] [Doc2Vec in NLP: GeeksForGeeks](#); consulté le 03/08/2024
- [W10] [How does transformers Work](#); consulté le 05/10/2024
- [W11] [BERTopic: Leveraging BERT embeddings for topic modeling](#); consulté le 10/10/2024

TABLE DES MATIERES

SIGLES ET ABREVIATIONS	4
LISTE DES FIGURES	5
LISTE DES CAPTURES	5
LISTE DES TABLEAUX	6
LISTE DES ÉQUATIONS.....	6
SOMMAIRE	7
INTRODUCTION.....	1
Chapitre 1 : Présentation Du Sujet	2
1.1. Problématique	2
1.2. Objectifs du mémoire	2
1.3. Démarche et méthodologie	3
1.4. Délimitations ..	3
Chapitre 2 : Introduction aux Fakes News	4
2.1. Définition	4
2.2. Les éléments constitutifs des Fake News	4
2.3. Les différents types de Fake News	5
2.3.1 <i>Désinformation en politique</i>	5
2.3.2 <i>Désinformation en santé</i>	5
2.3.3 <i>Les Fake News de divertissement</i>	6
2.4. Les mécanismes de propagation des Fake News	6
2.4.1 <i>Titres sensationnels et accrocheurs</i>	6
2.4.2 <i>Techniques de référencement</i>	6
2.4.3 <i>Les commentaires et interactions du lecteur</i>	6
2.4.4 <i>Les Réseaux sociaux</i>	7
2.5. Approches pour la détection des Fake News.....	7
2.5.1 <i>Approches basées sur les caractéristiques textuelles</i>	7
2.5.2 <i>Approches Fondées sur les Métadonnées</i>	7

2.5.3	<i>Approches Axées sur la Vérification des Faits</i>	7
2.5.4	<i>Approches basées sur l'apprentissage automatique.....</i>	8
Chapitre 3 : Introduction à l'Intelligence Artificielle et au Traitement du Langage Naturel		9
3.1.	Apprentissage automatique.....	9
3.1.1	<i>L'apprentissage supervisé.....</i>	9
3.1.2	<i>L'apprentissage non supervisé</i>	10
3.1.3	<i>L'apprentissage par renforcement</i>	12
3.2.	L'apprentissage profond	12
3.3.	Les applications de l'Intelligence Artificielle	13
3.3.1	<i>La médecine.....</i>	13
3.3.2	<i>Les transports</i>	13
3.3.3	<i>Le service client.....</i>	13
3.4.	Fondement du NLP	14
3.5.	Le prétraitement des données	14
3.6.	La vectorisation des données	16
3.6.1	<i>Vectorisation basée sur la syntaxe</i>	16
3.6.2	<i>Vectorisation basée sur la sémantique</i>	18
3.7.	Les Transformers	20
3.7.1	<i>Architecture des Transformers</i>	20
3.7.2	<i>Concepts clés</i>	21
3.7.3	<i>Cas d'un modèle de langage utilisant les Transformers : BERT.....</i>	21
3.8.	Applications du NLP	22
3.9.	Machine Learning, NLP et détection de Fake News	22
3.9.1	<i>Détection non supervisée des Fake News.....</i>	23
Chapitre 4 : Conception du modèle de détection des Fake News		24
4.1.	Présentation de la solution.....	24
4.2.	Collecte des données : par Web Scraping	24
4.2.1	<i>Fonctionnement</i>	24

4.3.	Modélisation de sujets	26
4.4.1	<i>BertTopic</i>	26
4.4.2	<i>Latent Dirichlet Allocation</i>	28
4.4.3	<i>Le résumé de sujet</i>	30
4.4.4	<i>Scores de similarité</i>	30
4.4.5	<i>Classification des articles</i>	31
	Chapitre 5 : Implémentation de la solution	32
5.1.	Présentation du jeu de données.....	32
5.2.	Prétraitement des données	32
5.2.1	<i>L'extraction et correction des valeurs numériques</i>	32
5.2.2	<i>La suppression des caractères de non-breaking space</i>	33
5.2.3	<i>Harmonisation des Dates</i>	33
5.2.4	<i>Traitemet des Commentaires</i>	33
5.2.5	<i>Utilisation du champ ‘sousCategorieArticle’</i>	33
5.2.6	<i>Traitemet du champ ‘auteurArticle’</i>	34
5.3.	Analyse exploratoire.....	34
5.3.1	<i>Proportion des articles selon leur source et leur thématique</i>	35
5.3.2	<i>Analyse des sources les plus influentes et de la répartition temporelle des articles</i>	35
5.4.	Environnement de Travail	36
5.5.	Logiciels utilisés	36
5.5.1	<i>Jupyter Notebooks</i>	36
5.5.2	<i>Néo4j Sandbox</i>	36
5.6.	Librairies utilisées.....	36
5.6.1	<i>Pandas</i>	36
5.6.2	<i>LDA et BertTopic</i>	37
5.6.3	<i>Spacy et Gensim</i>	37
5.7.	Modélisation de sujet.....	37
5.7.1	<i>BerTopic</i>	37

5.7.2	<i>LDA</i>	38
5.8.	Le résumé de sujet	39
5.9.	Calcul de Similarité	39
5.10.	Classification des articles	40
5.11.	Résultats et analyse des articles classifiés	42
5.11.1	<i>Résultats</i>	42
5.11.2	<i>Analyse des articles classifiés</i>	44
5.11.3	<i>Vue sur le sujet 111 : « louga_plage_potou_morgue »</i>	46
5.11.4	<i>Vue sur le sujet 1 « enseignants_élèves_ecole_les »</i>	47
5.12.	Limites de la solution	49
5.13.	Perspectives d'amélioration.....	49
	CONCLUSION	51
	BIBLIOGRAPHIE	52
	WEBOGRAPHIE.....	53
	TABLE DES MATIERES	54
	ANNEXES	58
	Annexe I. Champs de la Base de Données	58
	Annexe II. Ensemble des sujets obtenus avec BertTopic.....	59
	Annexe III. Ensemble des sujets obtenus avec LDA	62
	Annexe IV. Hyperparamètres du modèle BERTopic.....	63
	Annexe V. Hyperparamètres du modèle LDA	64

ANNEXES

Annexe I. Champs de la Base de Données

Champ	Type de Données Associé	Description du Champ
identifiantArticle	Chaine de caractères	Identifiant de l'article.
urlArticle	Chaine de caractères	L'URL de l'article sur le site Web.
sourceArticle	Chaine de caractères	La source de l'article.
thematiqueArticle	Chaine de caractères	La thématique de l'article.
sousCategorieArticle	Chaine de caractères	La sous-catégorie de l'article.
imagesArticle	Tableau de chaine de caractères	Les URLs des images associées à l'article.
titreArticle	Chaine de caractères	Le titre de l'article.
resumeArticle	Chaine de caractères	Un résumé ou une description brève de l'article.
auteurArticle	Chaine de caractères	L'auteur de l'article.
datePublicationArticle	Chaine de caractères	La date de publication de l'article.
dateMiseJourArticle	Chaine de caractères	La date de mise à jour de l'article.
videosArticle	Tableau de chaine de caractères	Les URLs des vidéos associées à l'article.
nombreLikesArticle	Entier	Le nombre de likes ou de j'aime pour l'article.
nombreLecturesArticle	Entier	Le nombre de lectures de l'article.
nombreCommentairesArticle	Entier	Le nombre de commentaires sur l'article.
nombrePartagesArticle	Entier	Le nombre de partages de l'article.
contenuArticle	Chaine de caractères	Le contenu complet de l'article.
commentairesArticle	Tableau d'objets JSON	Les commentaires associés à l'article.
dateCollecteArticle	Chaine de caractères	La date à laquelle l'article a été collecté ou extrait.
motsClesArticle	Chaine de caractères	Les mots-clés associés à l'article.

Annexe II. Ensemble des sujets obtenus avec BertTopic

Numéro de Sujet	Libellé	Nombre d'article	% Fake News
-1	je_et_la_de	1943	N.A
0	khalifa_sall_dakar_maire	172	0.58
1	enseignants_élèves_école_les	144	29.17
2	sonko_ousmane_94_affaire	104	3.85
3	santé_hôpital_les_sanitaire	100	46
4	ipres_conseil_général_directeur	86	11.63
5	dollar_français_peso_dinar	82	18.29
6	aéroport_avion_passagers_securiport	72	16.67
7	et_des_la_les	67	22.39
8	brésil_singapour_lions_match	62	61.29
9	free_wari_opérateur_mobile	59	28.81
10	parti_socialiste_tanor_ps	59	25.42
11	pays_migrants_espagne_migration	57	35.09
12	équipe_finale_sénégal_can	54	59.26
13	niang_mbaye_singapour_blessure	51	9.8
14	ans_enfant_été_son	50	4
15	revue_presse_octobre_2019	48	29.17
16	je_moi_tu_elle	48	50
17	presse_ligne_site_carte	44	38.64
18	mosquée_wade_massalikoul_sonko	43	4.65
19	iran_khodro_viande_ibk	43	4.65
20	ins_data_ad_td	41	97.56
21	magal_safar_touba_serigne	40	7.5
22	balla_eumeu_combat_modou	40	12.5
23	caf_zamalek_foot_caire	40	15
24	préfet_arrondissement_matricule_nommé	40	45
25	énergie_solaire Électricité_agriculture	40	67.5
26	clip_regardez_vidéo_goor	39	76.92
27	pèlerinage_pèlerins_la_des	38	21.05
28	seck_sidy_été_soukèye	38	2.63
29	album_musique_photos_viviane	38	57.89
30	foncière_collectivités état_chef	37	45.95
31	wade_pds_sall_macky	37	0
32	el_diouf_avocat émission	35	5.71
33	ndoumbélé_on_est_qui	34	8.82
34	projets_et_des_sénégal	32	59.38
35	accident_blessés_camion_route	31	19.35
36	mort_film_il_ses	31	12.9
37	maisons_immeubles_salam_darou	29	3.45

Numéro de Sujet	Libellé	Nombre d'article	% Fake News
38	perspiciatis_unde_iste_mque	29	86.21
39	firmino_brésil_score_match	29	51.72
40	électoral_élections_opposition_processus	29	24.14
41	syrie_turque_turquie_offensive	28	46.43
42	rts1_radio_streamaudio_direct	28	71.43
43	assemblée_commission_députés_règlement	28	21.43
44	messi_lionel_fc_barcelone	27	29.63
45	drogue_cocaïne_contôleurs_documents	27	0
46	timis_frank_sall_aliou	27	3.7
47	sall_pétrole_macky_prix	27	0
48	diack_iaaf_lamine_fils	26	0
49	trump_donald_démocrates_enquête	26	3.85
50	sylla_mbacké_serigne_ahma	26	0
51	real_madrid_zidane_club	26	19.23
52	pluies_nuit_fortes_vent	25	32
53	commissaire_urbaine_sûreté_pharmacien	25	0
54	liverpool_mané_league_sadio	24	45.83
55	chirac_jacques_mugabe_président	23	8.7
56	monsieur_ministre_république_du	23	4.35
57	ra_serigne_cheikh_mbacké	23	0
58	sadio_mané_liverpool_brésil	23	34.78
59	dpee_2019_prévisions_prévision	23	73.91
60	ipres_environnement_du_de	22	13.64
61	ufoa_finale_thiès_demi	22	50
62	bus_saed_véhicule_louis	22	22.73
63	dramé_watt_coumba_hawa	22	0
64	finale_ufoa_mali_thiès	22	77.27
65	puits_gaz_pays_pétrole	21	52.38
66	adama_gaye_justice_journaliste	21	4.76
67	dikk_dem_véhicules_transport	21	14.29
68	adamu_boko_haram_shekau	21	0
69	cni_élections_scrutin_les	21	42.86
70	senghor_léopold_il_sédar	20	20
71	sonko_leader_il_ousmane	20	30
72	yeene_maux_commentaire_aïssata	20	100
73	guinéens_condé_constitution_mandat	20	20
74	eau_forages_magal_touba	20	5
75	police_elèves_awa_ndiaye	20	25
76	mamie_actrice_mhd_awards	19	26.32
77	armée_hélicoptère_crash_hélicoptères	19	42.11
78	sde_suez_suprême_recours	19	10.53
79	logement_logements_fofana_urbanisme	19	0

Numéro de Sujet	Libellé	Nombre d'article	% Fake News
80	cancer_cancers_sein_col	19	47.37
81	serigne_mbacké_khalife_mountakha	18	22.22
82	dialogue_national_sall_famara	18	16.67
83	attaque_soldats_boulkessi_gsim	18	50
84	halle_fusillade_police_synagogue	18	0
85	aliou_football_cissé_lions	18	27.78
86	episode_vues_astra_episode	17	94.12
87	1451_tf_fcfa_sofico	15	0
88	diack_iaaf_qatar_fils	17	5.88
89	ambassadeur_france_auprès_lalliot	17	0
90	daoud_lagn_fii_invitation	16	25
91	collectivités_territoriales_maires_décentralisation	16	56.25
92	neymar_sélections_match_brésil	16	56.25
93	pape_diouf_prince_label	16	6.25
94	en_qui_sur_cela	15	6.67
95	femmes_der_récupérateurs_fonciers	14	50
96	gana_psg_but_idrissa	13	46.15
97	permis_conduire_support_rose	15	20
98	joueurs_favori_liste_pape	13	69.23
99	budget_finances_crédits_programme	13	7.69
100	dabakh_maodo_niasse_sy	13	0
101	touba_cheikh_père_serigne	13	0
102	climatique_climat réchauffement_carbone	13	61.54
103	no_ffffff_podcast_télécharger	13	84.62
104	ziguinchor_président_macky_sall	13	7.69
105	physique_chimie_gueye_snes	12	0
106	sonko_justicier_ousmane_plateau	12	16.67
107	alcaly_saoudite_arabie_morsi	12	0
108	com_sites_retrouvez_pornographiques	11	63.64
109	Min_Vi_Ali_Zionist	11	18.18
110	trésor_incendie_bureau_dispositif	11	0
111	louga_plage_potou_morgue	11	0
112	ministre_demandé_préfet_arrondissement	11	72.73
113	kfc_restaurant_brvm_ngom	11	9.09
114	football_match_brésil_sénégal	10	80

Annexe III. Ensemble des sujets obtenus avec LDA

Numéro de Sujet	Libellé	Nombre d'article	% Fake News
1	karim_latifah_bien_mohamed_aller_venir	95	37.89
2	priver_secteur_secteur priver_etat_sall_contrat	51	23.53
3	sall_macky_politique_pouvoir_wade_bien	488	14.75
4	prendre_malick_voir_aicha_savoir_fois	163	30.67
5	cheikh_serigne_général_touba_religieux_homme	246	17.07
6	afrique_développement_milliard_economique_projet_africain	364	47.8
7	2019_tweet_match_aliou_joueur_brésil	202	35.64
8	article_électoral_droit_sonko_code_affaire	146	13.01
9	dakar_kaolack_tweet_acte_ziguinchor_jour	96	30.21
10	afrique_personne_mort_contre_nigérian_journaliste	92	23.91

Annexe IV. Hyperparamètres du modèle BERTopic

Hyperparamètre	Description	Valeur
embedding_model	Pre-trained embedding model used for document embeddings (e.g., SBERT, all-MiniLM-L6-v2).	"paraphrase-multilingual-MiniLM-L12-v2"
top_n_words	Nombre de mots clés pour représenter chaque thème.	20
min_topic_size	Nombre minimum de documents nécessaires pour constituer un thème valide.	10
n_gram_range	Nombre de n-grammes à prendre en compte lors de la formation des sujets (par exemple, bigrammes, trigrammes).	(1, 1)
calculate_probabilities	Calculer ou non les probabilités des sujets par document, ce qui est plus lent mais plus détaillé.	False
hdbscan_min_cluster_size	Taille minimale des clusters lors de l'utilisation de l'algorithme de clustering HDBSCAN.	10
verbose	Affichage ou non de la progression et des informations intermédiaires pendant l'apprentissage du modèle.	False
reduction_model	Modèle de réduction de la dimensionnalité pour les incorporations avant le regroupement (par exemple, ACP ou UMAP).	UMAP
low_memory	Réduction de l'utilisation de la mémoire grâce au traitement des documents par lots.	False
seed_topic_list	Liste prédéfinie de sujets pour guider la formation de nouveaux sujets pendant la formation.	None

Annexe V. Hyperparamètres du modèle LDA

Hyperparamètre	Description	Valeur
num_topics	Le nombre de sujets à extraire des données.	10
num_words	Le nombre de mots les plus pertinents utilisés en cas de distance.	20
distance	La métrique de distance pour calculer la différence avec	jaccard
random_state	Nombre aléatoire pour la reproduction des résultats.	1
id2word	Correspondance entre les identifiants de mots et les mots. Elle est utilisée pour déterminer la taille du vocabulaire, ainsi que pour le débogage et l'impression des sujets.	Vocabulaire
passes	Nombre de passages au sein du corpus au cours de la formation.	5
chunksize	Nombre de documents à utiliser dans chaque itération.	1000

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



Nom et Prénoms : SANOGO Moussa Steve Belvin

Titre du mémoire : Conception d'un modèle de Clustering pour la détection de Fausses Informations au sein de la presse sénégalaise en ligne

Directeur de mémoire : Pr. BOUSSO Mamadou

Codirecteur de mémoire : M. PREIRA Jean-Marie

Résumé

Face à la problématique de la prolifération des Fake News au sein des articles de presse en ligne au Sénégal, nous avons proposé une approche de détection basée sur l'apprentissage non supervisé. Cette approche repose sur l'usage d'outils de NLP et de Larges Modèles de Langage : BertTopic, BertSum afin de regrouper les articles en sujets cohérents. Par la suite, une analyse de similarité a permis la classification des données.

Une analyse des articles classifiés a permis d'identifier quatre grands groupes de sujets, classés selon la proportion potentielle de Fake News observée en leur sein. Par exemple, certains sujets, comme les domaines juridiques et administratifs formels, sont totalement exempts de désinformation, tandis que d'autres, tels que les compétitions sportives internationales ou certains débats politiques, présentent une forte prévalence de Fake News. Ces observations soulignent une variabilité dans la vulnérabilité des thématiques à la désinformation, influencée par des facteurs tels que la complexité, l'émotion, ou encore la traçabilité des informations.

Bien que l'approche proposée présente des limites, elle constitue un point de départ pour développer des méthodes plus efficaces, notamment en l'absence de données étiquetées. Dans l'ensemble, cette étude contribue aux efforts en cours pour lutter contre la propagation des fausses nouvelles et de la désinformation à l'ère numérique.

Mots clés : *Fake News, IA, Machine Learning, NLP, Topic Modeling, Bert, BertTopic.*



THESIS FOR TELECOMMUNICATIONS DESIGN ENGINEERING DEGREE

Full Name: SANOGO Moussa Steve Belvin

Thesis Title: Design of a Clustering model for detecting Fake News in the Senegalese online press

Supervisor: M. BOUSSO Mamadou

Co-Supervisor: M. PREIRA Jean-Marie

ABSTRACT

Faced with the proliferation of Fake News within online press articles in Senegal, we have proposed a detection approach based on unsupervised learning. This approach relies on the use of NLP tools and Large Language Models: BertTopic, BertSum to group articles into coherent topics. Subsequently, a similarity analysis was used to classifier our data.

An analysis of the classified articles identified four main groups of topics, categorized based on the potential proportion of Fake News observed within them. For instance, some topics, such as formal legal and administrative domains, are entirely free from misinformation, while others, such as international sports competitions or certain political debates, exhibit a high prevalence of Fake News. These findings highlight a variability in the susceptibility of topics to misinformation, influenced by factors such as complexity, emotional appeal, and the traceability of information.

Although the proposed approach has its limitations, it provides a starting point for developing more effective methods, particularly in the absence of labeled data. Overall, this study contributes to ongoing efforts to combat the spread of fake news and misinformation in the digital age.

Keywords: *Fake News, AI, Machine Learning, NLP, Topic Modeling, Bert, BertTopic.*