République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université de Mohamed El Bachir El Ibrahimi de Borj Bou Arréridj
Faculté des Mathématiques et d'Informatique
Département d'informatique

MEMOIRE

Présenté en vue de l'obtention du diplôme

Master en Informatique

Spécialité : Ingénierie de l'Informatique Décisionnelle

THEME

La modélisation thématique pour la caractérisation des fausses informations sur les médias sociaux

Présenté par :

GUEMRAOUI ZINEB

BEHIH DALILA

Soutenu publiquement le : jj/mm/aaaa

Devant le jury composé de :

Président : …….

Examinateur : ……

Encadreur : MOHDEB Djamila. M.C.A à l'Université de Bordj Bou Arréridj

2022/2023

Dédicace

Je dédie ma graduation à ceux qui ont été la raison de ma joie

La bougie de mon chemin, ma mère et mon père

Et à mon frère et ma sœur

Et à tous mes amis qui ont partagé ma joie.

Zineb

Dédicace

Je dédie ce travail :

A ma chère mère

A mon cher père

Qui n'ont jamais cessé de formuler des prières à mon égard, de me soutenir et de m'épauler pour que je puisse atteindre mes objectifs.

A mes frères Hakim, Mounir et Mohammed

A mes sœurs Fahima et Nozha

Pour leurs soutiens moraux tout au long de mes études, et qui m'ont toujours encouragé, et à qui je souhaite plus de succès.

À tous mes amis

A toute personne qui occupe une place dans mon cœur

Dalila

Remerciement

Nous remercions Dieu le Tout-Puissant pour la santé et la volonté qui nous ont permis de terminer ce mémoire.

Nous sommes profondément reconnaissants envers Madame MOHDEB Djamila pour son encadrement exceptionnel, sa patience, sa rigueur et sa disponibilité.

Nous exprimons notre gratitude envers les membres du jury pour avoir accepté d'évaluer notre travail.

Enfin, nous remercions toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce mémoire.

Résumé

Les théories du complot sont un sujet courant sur les sites de médias sociaux tels que Facebook. Ce type important de fausses informations incluent de nombreuses affirmations fausses et trompeuses qui encouragent la remise en question des faits et événements officiels. Dans cette étude, nous avons extrait depuis le réseau social Facebook des publications textuelles en langue arabe dont le contenu est lié au sujet des théories du complot. La collection des documents collectée a été ensuite sujet d'un certain nombre d'expérimentations qui visent principalement à évaluer le potentiel des techniques classiques et avancées de la modélisation thématique dans la tâche d'identification des thèmes communs aux théories de complots qui se propagent en langue arabe sur Facebook.

Les outils de modélisation thématique implémentés dans cette recherche comprennent LDA, LSI, NMF, Top2Vec et BerTopic. Ces outils utilisent diverses approches algorithmiques pour analyser et extraire les thèmes communs des textes. L'efficacité de ces outils pour le texte arabe a été évaluée selon certain nombre de critères dans l'objectif d'améliorer la précision des résultats et améliorer le rapport de cohérence entre les résultats fournis par chaque technique. Les résultats obtenus montrent que la performance des modèles avancés Top2Vec et BerTopic surpassent celle des modèles classiques LDA, LSI et NMF.

Mots-clés : Réseaux sociaux, fausse information, modélisation thématique, traitement automatique de la langue arabe.

Abstract

Conspiracy theories are a common topic on social media sites such as Facebook. This prominent type of fake news includes many false and misleading claims that encourage questioning of official facts and events. In this study, we extracted from the social network Facebook textual publications in Arabic language whose content is related to the subject of conspiracy theories. The collected collection of documents was then the subject of a number of experiments aimed primarily at evaluating the potential of classical and advanced techniques of thematic modeling in the task of identifying common themes in conspiracy theories that are propagated in Arabic language on Facebook.

Topic modeling tools implemented in this research include LDA, LSI, NMF, Top2Vec and BerTopic. These tools use various algorithmic approaches to analyze and extract common themes from texts. The effectiveness of these tools for Arabic text has been evaluated according to a number of criteria with the aim of improving the accuracy of the results and improving the consistency ratio between the results provided by each technique. The results obtained show that the performance of the advanced models Top2Vec and BerTopic surpass that of the classical models LDA, LSI and NMF.

Keywords: Social networks, false information, topic modeling, Arabic Natural Language Processing.

ملخص

نظريات المؤامرة موضوع شائع على مواقع التواصل الاجتماعي مثلFacebook ، يتضمن هذا النوع البارز من الأخبار الكاذبة العديد من الادعاءات الكاذبة والمضللة التي تشجع على التشكيك في الحقائق والأحداث الرسمية. استخرجنا في هذه الدراسة من شبكة التواصل الاجتماعي منشورات نصية على Facebook باللغة العربية ، يرتبط محتواها بموضوع نظريات المؤامرةكانت مجموعة الوثائق التي تم جمعها موضوعًا لعدد من التجارب التي تهدف في المقام الأول إلى تقييم إمكانات التقنيات الكلاسيكية والمتقدمة للنمذجة الموضوعية في مهمة تحديد الموضوعات المشتركة في نظريات المؤامرة التي يتم نشرها باللغة العربية على .Facebook

تتضمن أدوات نمذجة الموضوعات المطبقة في هذا البحث LDA و LSI و NMF و Top2Vec و BerTopic. تستخدم هذه الأدوات أساليب حسابية مختلفة لتحليل واستخراج الموضوعات المشتركة من النصوص. تم تقييم فعالية هذه الأدوات للنص العربي وفقًا لعدد من المعايير بهدف تحسين دقة النتائج وتحسين نسبة التناسق بين النتائج المقدمة من كل تقنية. أظهرت النتائج التي تم الحصول عليها أن أداء الطرازين المتقدمين Top2Vec و BerTopic يفوق أداء النماذج الكلاسيكية LDA و LSI وNMF.

الكلمات المفتاحية: شبكات التواصل الاجتماعي، معلومات كاذبة، نمذجة موضوعية، معالجة آلية للغة العربية.

Table des matières

Liste des abréviations

AraBERT	Arabic Bidirectional Encoder Representations from Transformers
BOW	Bag Of Words
CSV	Comma-Separated Values
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
IDF	Inverse Document Frequency
IR	Information Retrieval
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
ML	Machine Learning
MT	Modélisation Thématique
NLP	Natural Language Processing
NMF	Non-negative Matrix Factorization
NLTK	Natural Language Toolkit
TF	Term Frenquency
UMAP	Uniform Manifold Approximation and Projection

Liste des figures

Liste des tableaux

Liste des algorithmes

Introduction générale

1. Contexte

L'avènement des réseaux sociaux a révolutionné la manière dont l'information est diffusée et consommée, offrant aux gens des opportunités sans précédent de se connecter et de partager des idées. Cependant, cette démocratisation de l'information s'accompagne de ses propres défis, car les fausses informations, notamment celles de type « **théories du complot** », ont trouvé un terreau fertile pour se propager et proliférer sur ces plateformes. Les fausses informations du type théorie du complot font référence à des récits trompeurs ou fabriqués qui proposent des explications alternatives ou des agendas cachés derrière des événements, des phénomènes ou des actions importantes. Les théories du complot manquent généralement de preuves substantielles et reposent sur des spéculations, des erreurs logiques et une interprétation sélective des faits pour étayer leurs affirmations. Ce phénomène a suscité une attention considérable en raison de ses implications potentielles pour l'opinion publique, la cohésion sociale et les processus démocratiques.

La modélisation thématique joue un rôle crucial dans l'identification et l'analyse des fausses informations, notamment les théories du complot, circulant sur les réseaux sociaux. En automatisant l'identification et l'analyse des sujets et des modèles connexes, la modélisation thématique facilite une compréhension plus approfondie de la dynamique derrière la propagation de la désinformation. Ces connaissances, à leur tour, permettent aux chercheurs, aux décideurs et aux plateformes de médias sociaux de développer des interventions et des contre-mesures ciblées pour atténuer l'impact des théories du complot et favoriser une société plus informée et résiliente.

2. Problématique

Le principal problème de la modélisation thématique est la difficulté des systèmes automatisés à comprendre et à analyser les langues qui reposent sur des caractères non latins tels que l'arabe. Les langues avec des lettres latines dépendent de l'organisation des lettres et des mots d'une manière qui facilite le processus d'analyse et de compréhension, tandis que d'autres langues ont besoin de méthodes et de techniques avancées pour comprendre les textes et en extraire des informations importantes. Par conséquent, ce problème rend difficile pour les systèmes

automatisés de comprendre le contenu écrit en arabe, d'autant plus qu'il a plusieurs dialectes différents, et cela peut donc conduire à la non-divulgation d'informations importantes liées aux théories du complot qui circulent sur les sites de réseaux sociaux.

3. Objectif et contribution

En évaluant et en testant les performances des algorithmes de modélisation thématique pour découvrir des modèles et des sujets cachés dans une collection de textes arabes extrait de Facebook, cette recherche vise à fournir un aperçu du potentiel de ces algorithmes pour identifier et catégoriser efficacement les théories du complot dans le contenu en ligne en langue arabe. En outre, cette étude reconnaît la prévalence et l'impact des théories du complot sur Facebook, l'une des plateformes de médias sociaux les plus populaires et les plus influentes au monde, soulignant l'importance de relever les défis spécifiques de la désinformation dans ce contexte spécifique.

4. Structure du rapport

La suite de ce mémoire est organisée en quatre chapitres :

- Le premier chapitre fournit un exposé général sur le thème des théories du complot.
- Le deuxième chapitre présente le sujet de la modélisation thématique, en expliquant ses objectifs et en identifiant certaines de ses techniques.
- Le troisième chapitre décrit l'objectif du projet et la méthodologie utilisée pour concevoir et structurer sa partie pratique.
- Le quatrième et dernier chapitre, décrit l'environnement matériel et logiciel dans lequel notre cas d’étude a été implémenté, et présente puis discute les résultats obtenus.

Chapitre 01 : Les théories de complots : concepts de base

1.1. Introduction

Ce chapitre présente un aperçu général sur les concepts théoriques qui sont en relation avec le terme « théorie de complot ». Nous fournirons dans les sections suivantes la définition académique de la théorie de complot et mentionnerons ses caractéristiques les plus importantes. Nous aborderons ensuite les raisons qui incitent les gens à adhérer à ces théories, et ses impacts positifs ou négatifs sur la société.

1.2. Définition du terme « fausse information »

Les fausses informations sont des renseignements mensongers ou inventés dont l'authenticité est ambiguë ou jamais confirmée. Elles portent sur une situation, un évènement, une personnalité publique ou inconnue, une organisation, un gouvernement voire un état.

Les informations erronées sont délivrées via les médias traditionnels ou les médias sociaux non institutionnels tels les blogs et les réseaux sociaux en ligne. L'objectif se varie entre manipuler et tromper l'opinion publique ou un auditoire bien déterminé sur des sujets spécifiques ; modifier leurs décisions et actions ; faire du mal aux opposants (organisations, partis, politiciens, intellectuels …etc.) ou bien pour le but d'augmenter le profit et la popularité [1].

1.3. Définition du terme « théorie de complot »

En s'appuyant sur la définition d'Uscinski et de ses collègues (Uscinski et al., 2016), nous désignons par les théories du complot (*en anglais*, conspiracy theories) « des tentatives d'expliquer des événements et des circonstances sociaux et politiques importants en affirmant de manière invraisemblable qu'ils sont l'effet ultime et consciemment poursuivi d'une secrète conspiration menée par un groupe d'acteurs puissants et malveillants » [2].

1.4. Caractéristiques d’une fausse information de type théorie de complot

Bien qu'il n'y ait pas de consensus uniforme dans la littérature, les théories du complot se répandent particulièrement dans les situations où l’information est faiblement disponible et le niveau de confiance aux sources officielles de l’information est très réduit [3].

- **Caractéristiques centrales :** Les théories de complots invoquent habituellement des actions ou des opérations secrètes qui sont menées par un petit groupe de conspirateurs [4].
- **Caractéristiques psychologiques :** Les théories du complot se caractérisent par leur tendance à utiliser des idées négatives et à approfondir les questions de mort et de religion [5].
- **Caractéristiques politiques :** Plusieurs études indiquent que les théories du complot forment une association positive avec les activités politiques, ce qui encourage les gens à s'engager en politique [6]. De ce fait, le complotisme est considéré comme un « type de discours politique » qui se définit par trois caractéristiques : (1) une propension à « localiser la source de phénomènes sociaux et politiques inhabituels dans des forces invisibles, intentionnelles et malveillantes », (2) une propension à interpréter les événements « en termes de lutte manichéenne entre le bien et le mal », et (3) l’implication que « les récits traditionnels des événements politiques sont mensongères » [3].
- **Caractéristiques sociales :** L’audience ciblé par les théories du complot ne se limite pas à un groupe spécifique de la société, mais s'étend à tous les groupes, des intellectuels aux éduqués aux ignorants [7].

1.5. Les facteurs qui influencent la propagation des théories de complot

- Les théories de complot se répandent en raison de multiples facteurs, dont le facteur socio-politique comme la monopolisation du pouvoir, en plus du facteur d'inégalité sociale, ainsi que le facteur psychologique comme le raisonnement naïve et non scientifique des publics influencé par les conspirationnistes.
- Internet représente l'un des outils les plus importants pour la propagation des fausses informations. Les théories du complot ont proliféré en ligne, en particulier sur les médias sociaux, ce qui signifie que plus les individus passent de temps en ligne, plus ils adhèrent à ces théories [8].
- Le temps contribue à révéler la validité des théories du complot, car certaines d'entre elles s'affaiblissent, comme la « rumeur du vaccin Covid 19 », ou meurent, comme la « rumeur que le Jour du Jugement sera le 12/12/ 2012 ». Mais certaines de ces théories existent toujours avec la même force, comme « la conspiration des francs-maçons et les extraterrestres » [9].

1.6. Les raisons d'adhérer aux théories de complots

De nombreuses études se sont concentrées sur les relations entre divers traits psychologiques et la probabilité de croire à des théories de complots. En particulier :

- L'adhésion aux théories du complot est dû à la perte de confiance dans les informations officielles, qu'elles soient politiques, scientifiques ou médiatiques [10].
- La croyance aux théories du complot est liée à un ensemble de variables individuelles qui reflètent de mauvaises performances personnelles, telles que la folie interpersonnelle, le narcissisme, l'insatisfaction, l'attachement précaire et le machiavélisme [11].
- La littérature a montré que les faibles niveaux d'estime de soi, de pensée analytique, les idées paranoïaques, la schizotypie et les niveaux élevés de besoin de fermeture cognitive ont tendance à augmenter les croyances aux théories du complot [2].
- L'orientation politique peut également être un principal facteur alimentant les croyances

dans le complotisme et renforçant le scepticisme envers les autorités pour ceux qui ont une faible confiance dans les institutions [2].

1.7. Conséquences de croire aux théories des complots

1.7.1. Conséquences positives

- La croyance dans les théories du complot conduit à un sentiment de communauté partagé par d'autres qui soutiennent les mêmes théories, satisfaisant ainsi un besoin social.
- Ces théories peuvent inspirer une action collective et tenter un changement social, en particulier la réaction à des événements spécifiques, et ont donc potentiellement la capacité de satisfaire des besoins existentiels.
- Les théories du complot peuvent révéler de véritables anomalies dans les interprétations et les discours dominants.
- Les théories du complot peuvent permettre aux gens de remettre en question les hiérarchies sociales, ce qui peut encourager les gouvernements à être plus transparents [12].

1.7.2. Conséquences négatives

- Des recherches empiriques émergentes ont montré que les théories du complot peuvent entraîner une augmentation des sentiments d'impuissance, de déception, d'indifférence, de méfiance et d'anormalité plutôt que de les réduire.
- Les théories du complot peuvent être politiquement pernicieuses, comme organiser une manifestation, s'impliquer dans des actions politiques illégales ou l'inaction plutôt que l'action.
- Les théories du complot peuvent conduire à un désengagement scientifique, comme la falsification de données scientifiques pour obtenir un soutien ou un financement, et également influencer des choix médicaux importants concernant les maladies et les médicaments.
- Les théories du complot influencent la façon dont les gens se comportent dans leur travail quotidien et leur vie sociale, ainsi que s'écartent des normes sociales, ce qui les

rendent plus susceptibles de s'engager dans des comportements contre-normatifs (crimes) [12].

1.8. Les théories du complot les plus répandues

- **11 septembre 2001 :** Il y a beaucoup de critiques sur le fait que les attentats du 11 septembre n'ont pas été perpétrés par Al-Qaïda mais plutôt par Israël ou les États-Unis [13].
- **Terre plate :** De nombreuses théories circulent autour de l'idée que la terre est plate et non ronde, bien qu'il existe de nombreuses preuves qui le nient [14].
- **Vaccin contre le covid19** : Beaucoup de gens remettent en question la validité de ce vaccin, car ils pensent qu'il menace leur vie, et ils prétendent qu'il s'agit d'un projet de développement de micro puces traçables [15].
- **Les Juifs sont derrière chaque crime :** Certains récits décrivent les Juifs comme l'organisation moderne qui rassemble toutes les forces du mal et de la haine dans le monde, qui font l'impossible pour mettre en œuvre leurs plans destructeurs contre les sociétés et les civilisations [16].

1.9. Conclusion

Les théories du complot font référence à la mise en place de complots sataniques pour atteindre certains objectifs. Dans ce chapitre, nous avons traité ce type de fausses informations avec ses caractéristiques, ses impacts et les raisons de s'y engager.

Chapitre 02 : La modélisation thématique pour le texte arabe

2.1. Introduction

Avec le volume croissant d'informations textuelles en langue arabe circulant sur Internet, en particulier les réseaux sociaux, il est devenu important de traiter le contenu arabe en ligne pour satisfaire les différentes fins de l'automatisation.

Dans ce chapitre, nous allons découvrir les objectifs et le processus de la modélisation thématique en plus de ses techniques les plus importantes.

2.2. Définition de la modélisation thématique

La modélisation thématique, en anglais Topic Modeling, est une technique d'exploration de texte qui utilise des techniques statistiques ou probabilistes pour découvrir les sujets latents (cachés) dans les grandes collections de documents [17].

2.3. Objectifs de la modélisation thématique

La modélisation thématique prend une grandeur accrue dans de nombreux domaines tels que le traitement automatique du langage naturel (NLP), la recherche d'informations (IR) et l'apprentissage automatique (ML). Les objectifs poursuivis par la modélisation thématique sont les suivants :

- Détecter le contenu et analyser les informations dans les réseaux sociaux en ligne.
- Résumer les longs documents tels que des actualités, des articles et des livres.
- Proposer des méthodes et des techniques de traitement de textes courts [18].
- Découvrir la structure cachée d'un ensemble de documents [19].
- Vérifier et estimer l'utilité d'un texte [20].

2.4. Processus de la modélisation thématique pour texte arabe

La nature morphologique de la langue arabe est complexe. La plupart des mots arabes proviennent de milliers de racines. Le mécanisme de travail pour modéliser les sujets arabes comprend :

2.4.1. Collection de documents

Les documents sont collectés à partir de diverses sources « PDF, texte, fichiers Web… », Qu'il s'agisse de données textuelles courtes ou longues.

2.4.2. Nettoyage du texte

Le nettoyage des documents textuels en langue arabe consiste généralement à :

- Supprimer les documents qui contiennent moins que le nombre de mots spécifié.
- Supprimer l'extension de ligne entre les lettres arabes. Par exemple l'extension dans le mot : " جميـــــــل ".
- Supprimer la ponctuation, les lettres non arabes, les chiffres et les signes diacritiques.
- Supprimer les mots à une seule lettre et les lettres répétées.
- Appliquer la normalisation sur quelques lettres arabes. Par exemple le remplacement du dernier caractère " ة " par " ه ".

2.4.3. Processus de prétraitement

1. La suppression des mots vides

Les mots vides (*en anglais*, Stopwords) sont les mots qui apparaissent fréquemment et qui n'ont pas d'apport sémantique au processus de modélisation thématique, comme les prépositions, les conjonctions « ...الى ,من ». Ils doivent être éliminés avant la modélisation afin de réduire la taille finale du vocabulaire issu du corpus textuel à modéliser les thématiques [21].

2. La segmentation

La segmentation (*en anglais*, Tokenisation) est la première étape dans le processus d'analyse d'un texte. Elle consiste à identifier les unités constituant le texte, de ce fait le module de segmentation permet de fractionner le texte arabe en trois (03) niveaux, qui sont :

- La segmentation au niveau du texte : c'est de décomposer le texte en phrases par rapport aux signes de ponctuations.
- La segmentation au niveau de la phrase : consiste à décomposer la phrase en segments en éliminant les blancs et les virgules.
- La segmentation au niveau du mot (élimination des affixes) : est l'opération la plus délicate en terme de découpage, parce qu'elle consiste à enlever d'un mot toutes les composantes lexicales et grammaticales [22].

3. Le racinisation

Le racinisation (*en anglais*, stemming) consiste à extraire l'étymologie ou la racine des mots. Le but de cette étape est de réduire le nombre de termes uniques (taille du vocabulaire) en réduisant les mots dérivés ou réflexifs à leurs formes racines et aussi d'éviter les mêmes mots avec des formes différentes.

Le tableau ci-dessous représente le résultat de stemming d'un ensemble de mots en utilisant différents stemmers arabes [21].

Tableau 1. Exemples de processus de Stemming pour le texte arabe.

Mots originaux	الاقتصاد	علماء	يتعلمون	الام	مقالاته	يكلمني
Mots après stemmer Khoja	قصد	علم	علم	لوم	قلي	كلم
Mots après light 10 stemmer	اقتصاد	علماء	يتعلم	ام	مقالات	يكلمن
Mots après le radical ARL	قتصد	علماء	بتعلم	الم	مقال	كلم
Mots après stemmer Farasa	اقتصاد	عالم	تعلم	ام	مقال	كلمني

4. La lemmatisation du texte arabe

La lemmatisation est une opération qui consiste à transformer un mot éventuellement agglutiné ou possédant des marqueurs de dérivation à sa forme canonique (lemme ou racine).

- **Lemme :** est un mot graphique dont les affixes (préfixes, infixe et suffixes) ont étés supprimés. Exemple : soit le mot (" العلم ", La science), ce dernier est décomposé en préfixe " الـ " et lemme " علم ".
- **Racine :** le plus souvent trilitère (dans le cas de l'arabe), est une suite de consonnes forment le radical du mot. Exemple : soit le mot (" المدرسة ", Ecole) dont la racine est (" درس ", D+R+S) [22].

Un exemple plus général sur la lemmatisation du mot arabe est illustré dans la figure 1.

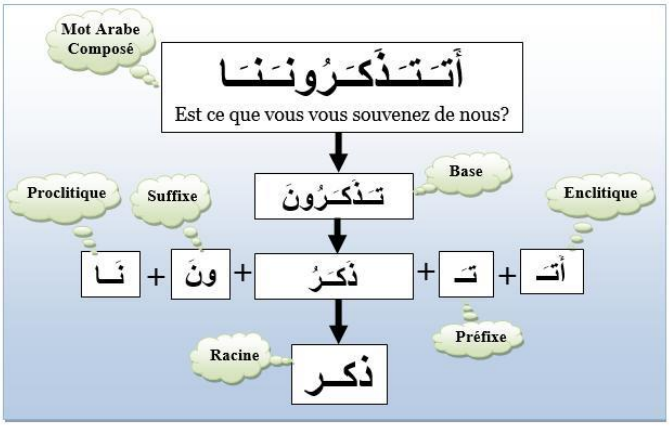


Figure 1. Résultat de la lemmatisation du mot " أتتذكروننا " [22].

2.4.4. Vectorisation du texte

Dans le traitement du langage naturel, les données (mots et phrases) sont représentées comme des vecteurs (une structure de données semblable à une liste ou une matrice de valeurs) pour faciliter leur traitement, l'une des méthodes les plus courantes de convertir le texte est TF-IDF et le sac-de-mots (Bag-Of-Words) [43].

2.5. Techniques de la modélisation thématique

Il existe de nombreux algorithmes destinés à la modélisation thématique :

2.5.1. Latent Dirichlet Allocation (LDA)

LDA est une technique courante pour modéliser et extraire les « thèmes cachés » d'un corpus donné. LDA se base sur deux hypothèses principales :

- Les documents sont composés d'un mélange de sujets.
- Les sujets sont composés d'un mélange de mots.

Les deux hypothèses mentionnées ci-dessus s'appliquent au jeu de données textuelles spécifié. Supposons que nous ayons cinq ensembles de documents $D1$, $D2$, $D3$, $D4$ et $D5$. Le texte est d'abord nettoyé, traité et encodé en mots pour obtenir une matrice Termes/Documents, qui se présente sous la forme d'une ligne pour chaque document et d'une colonne pour chaque mot, puis la matrice est convertie en deux autres matrices (la matrice Sujets/Documents et la matrice Termes/Sujets).

Le tableau ci-dessous représente la fréquence des mots dans tous les documents sous forme de matrice :

Tableau 2. Matrice Termes/Documents

	M1	*M2*	*M3*	*M4*	*M5*	*M6*	*M7*	*M8*
D1	*0*	*0*	*0*	*1*	*0*	*1*	*0*	*0*
D2	*0*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
D3	*1*	*0*	*0*	*0*	*1*	*1*	*0*	*0*
D4	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
D5	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*

L'ensemble de l'espace vectoriel LDA et son jeu de données sont représentés par le schéma ci-dessous dans la figure 2 :

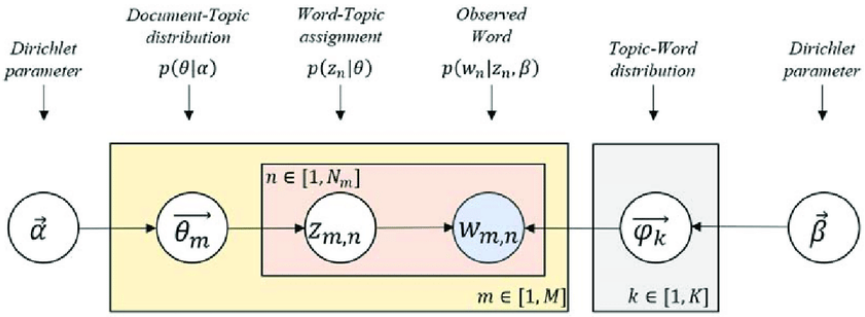


Figure 2. Espace vectoriel de LDA [25].

- La case jaune renvoie à tous les documents du corpus.
- La case couleur pêche correspond au nombre de mots d'un document.
- À l'intérieur de cette boîte en couleur pêche, il peut y avoir beaucoup de mots. L'un de ces mots est *w*, qui se trouve dans le cercle de couleur bleu.
- Alpha (α) contrôle la distribution des sujets par document, et bêta (□) contrôle la distribution de mots par sujet.
- M : est le nombre total de documents dans le corpus.
- N : est le nombre de mots dans le document.
- W : est le mot dans un document.
- Z : est le thème latent attribué à un mot.
- Thêta (θ) : est la distribution thématique [25].

2.5.2. Negative Matrix Factorization (NMF)

NMF est une méthode d'apprentissage automatique non supervisée. Le noyau principal de l'apprentissage non supervisé est la quantification de la distance entre les éléments. La distance peut être mesurée par différentes méthodes :

1. **Divergence Kullback-Leibler généralisée :** Il s'agit d'une mesure statistique utilisée pour quantifier les différences entre une distribution et une autre. Plus la valeur de la divergence

Kullback-Leibler est proche de zéro, plus la proximité des mots correspondants augmente. En d'autres termes, la valeur de divergence est inférieure.

$$Kl\_div\ (x,y)\ = \begin{cases} x\log\left(\frac{x}{y}\right) - x + y & x > 0, y > 0 \\ y & x = 0, y \geq 0 \\ \infty & otherwise \end{cases}$$

2. **Frobenius Norm :** En utilisant la norme de Frobenius. Il est défini par la racine carrée de la somme des carrés absolus de ses éléments. Elle est également connue sous le nom de norme euclidienne [26].

$$\|A\|_F \equiv \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$$

2.5.3. La méthode Top2Vec

Top2Vec est un algorithme qui détecte les sujets présents dans le texte et génère conjointement des vecteurs de sujet, de document et de mot intégrés. À un niveau élevé, l'algorithme effectue les étapes suivantes pour découvrir des thèmes dans une liste de documents.

1. **Générer des vecteurs d'intégration pour les documents et les mots :** Un vecteur d'intégration est un vecteur qui permet de représenter un document mot ou texte dans un espace multidimensionnel. La création de vecteurs d'intégration pour chaque document permet de traiter chaque document comme un point dans un espace multidimensionnel et crée également des vecteurs de mots intégrés conjointement, ce qui permet de déterminer ultérieurement les mots-clés du sujet.

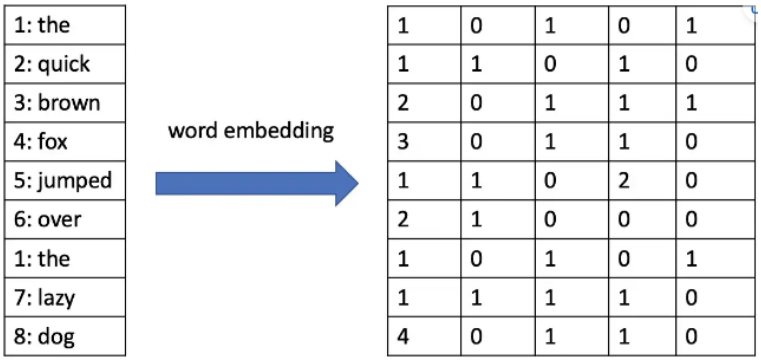


Figure 3. Génération des vecteurs d'intégration [27].

2. **Effectuer une réduction de dimensionnalité :** Une fois les vecteurs pour chaque document sont prêts, l'étape suivante consiste à les diviser en clusters à l'aide d'un algorithme de clustering, Top2Vec utilise un algorithme appelé UMAP (Uniform Manifold Approximation and Projection) pour générer des vecteurs d'intégration de dimension inférieure pour chaque document.
3. **Regrouper les vecteurs :** Top2Vec utilise HDBSCAN, un algorithme de clustering hiérarchique basé sur la densité, pour trouver des zones denses de documents.
4. **Attribuer des sujets à chaque cluster :** Une fois les clusters pour chaque document sont extraits, on peut simplement traiter chaque cluster de documents comme un sujet distinct dans le modèle de sujet. Chaque sujet peut être représenté sous la forme d'un vecteur de sujets qui est essentiellement le centroïde des documents originaux appartenant à ce groupe de sujets. Afin d'étiqueter le sujet à l'aide d'un ensemble de mots-clés, on peut calculer les mots les plus proches du vecteur centroïde du sujet. Une fois que les mots-clés pour chaque sujet sont déterminés, le travail de l'algorithme est terminé, et c'est à nous, en tant qu'humains, d'interpréter ce que ces sujets signifient vraiment [27].

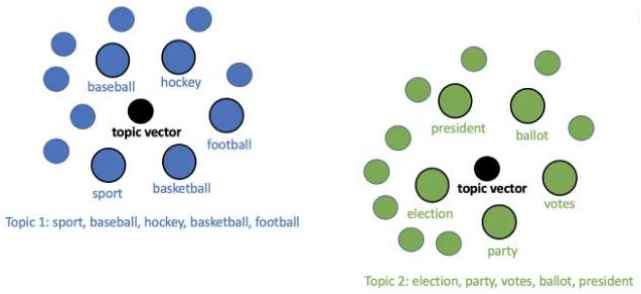


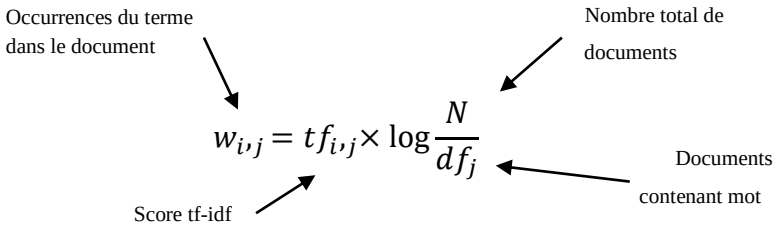
Figure 4. Attribuer des sujets à chaque cluster [27].

2.5.4. Latent Semantic Indexing (LSI)

L'analyse sémantique latente, ou LSA, est l'une des principales techniques de modélisation thématique. L'idée de base est de prendre un tableau de ce que nous avons - documentation et terme - et de le décomposer en une matrice de sujet de document et une matrice de terme de sujet distinctes.

Dans la première étape, nous créons la matrice Termes/Documents. Etant donné qu'il y a m documents et n mots dans notre vocabulaire, nous pouvons construire une matrice A de dimension $m \times n$ où chaque ligne représente un document et chaque colonne représente un mot. Dans la version la plus simple de LSA, chaque valeur peut être le nombre d'occurrences du mot j dans le document i.

Par conséquent, le modèle LSA remplace généralement les simples énumérateurs (Compte le nombre de fois que chaque mot apparaît dans chaque document) du tableau de Termes/Documents par le résultat de TF-IDF, ou fréquence du terme-fréquence inverse du document, attribue un poids au terme j dans le document i comme suit :



Intuitivement, un terme à un grand poids quand il se produit fréquemment à travers le document, mais rarement à travers le corpus. Le mot « M1 » peut apparaître souvent dans un document, mais comme il est probablement assez commun dans le reste du corpus, il n'aura pas un score TF-IDF élevé. Cependant, si le mot « M2 » apparaît souvent dans un document, parce qu'il est plus rare dans le reste du corpus, il aura un score tf-idf plus élevé [28].

2.5.5. La méthode BerTopic

BerTopic est une technique de modélisation de sujet qui utilise des transformateurs (intégrations BERT) et TF-IDF basé sur des classes pour créer des clusters denses. Il vous permet également de visualiser facilement les sujets créés.

L'algorithme BerTopic comporte trois étapes :

1. **Intégrer les données textuelles (documents) :** Dans cette étape, l'algorithme extrait les intégrations de documents avec BERT, ou il peut utiliser toute autre technique intégration.

2. **Documents groupés :** Il utilise l'UMAP pour réduire la dimensionnalité des intégrations et la technique HDBSCAN pour regrouper les intégrations réduites et créer des grappes de documents sémantiquement similaires.
3. **Créer une représentation de sujet :** La dernière étape consiste à extraire et à réduire les sujets avec TF-IDF basé sur les classes, puis à améliorer la cohérence des mots avec la pertinence marginale maximale [29].

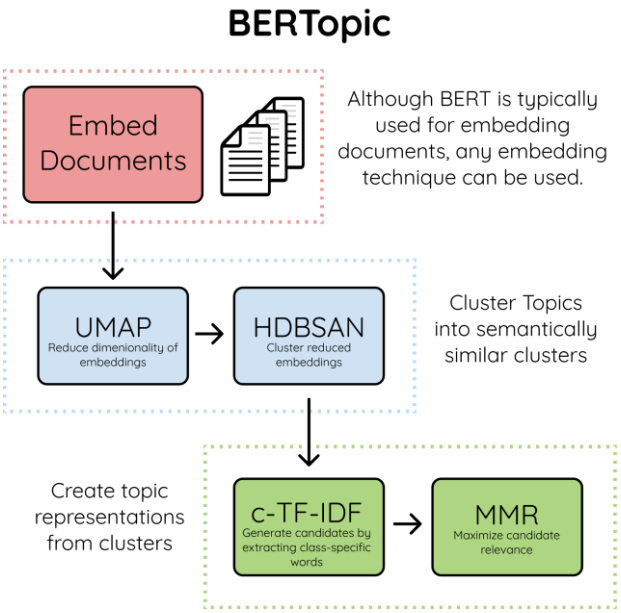


Figure 5. Créer une représentation de sujet [29].

2.6. Conclusion

Dans ce chapitre, nous nous sommes familiarisés avec le concept de la modélisation thématique. Nous avons par la suite expliqué le processus global de la modélisation et décrit certaines de ses techniques les plus utilisées.

Chapitre 03 : Conception et réalisation

3.1. Introduction

Dans ce chapitre, nous découvrirons le projet et comment il procède depuis l'étape de la collecte et du traitement des données jusqu'à l'application de méthodes de modélisation thématique et leur évaluation.

3.2. Description du projet

L'objectif de ce projet est d'exploiter la technique de la modélisation thématique pour caractériser les cas de désinformation de type « théories de complots » dans le contenu arabe sur le réseau social « Facebook ». A cette fin, plusieurs publications qui promeuvent les théories du complot en langue arabe ont été extraites puis analysées pour découvrir leurs principaux thèmes en utilisant les algorithmes de modélisation thématiques les plus connues par leurs efficacités dont LDA et LSI et NMF, Top2Vec et BerTopic.

3.3. Description de la méthodologie de conception

Pour atteindre l'objectif du projet, nous avons suivi une série d'étapes :

- Extraire et collecter les données textuelles à partir d'un ensemble de pages et de groupes Facebook faisant la promotion de théories du complot en langue arabe.
- Construire un jeu de données préliminaire qui contient tous les textes extraits.
- Nettoyage et prétraitement de données collectées.
- Application des techniques de modélisation thématique aux données finales pour la découverte des sujets latents.
- Evaluation des techniques appliquées.

La figure suivante montre les étapes ci-mentionnées :

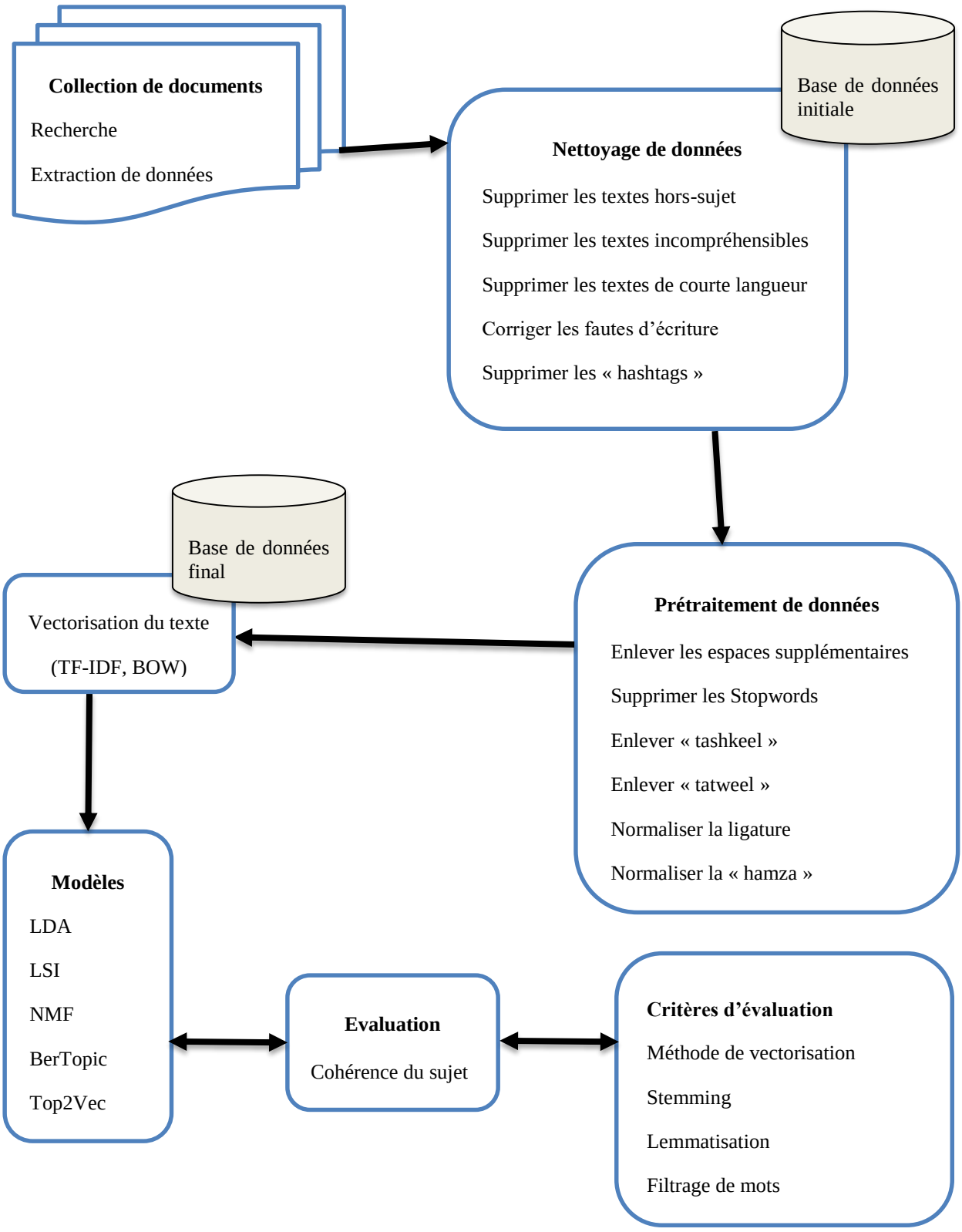


Figure 6. Architecture du système de modélisation thématique.

3.4. Méthodologie

3.4.1. Description du processus de collecte de données

Nous avons déjà abordé le sujet des théories du complot dans le premier chapitre et les raisons de leur propagation, qui sont les sites de réseaux sociaux. Au sujet de notre étude, nous nous sommes intéressés particulièrement par Facebook. Ci-dessous une description concise du processus que nous avons suivi pour la collecte des données :

Tableau 3. Description du processus de collection de données.

Projet: Modélisation thématique pour la caractérisation des fausses informations sur les réseaux sociaux
✓ **Objectif** : Collecter des données à partir d'autant de pages et groupes que possible sur Facebook qui promeuvent en langue arabe des fausses informations de type « théories du complot ». ✓ **Sources** : Cibler les pages et les groupes ouverts et publics pour extraire des données. Environ 38 pages et groupes publics ont été sélectionnés. ✓ **Outils** : Pour extraire les données des sources spécifiées, nous avons utilisé des scrapers Facebook, qui sont des outils conçus pour extraire les données des pages Facebook publiques. Les données extraites peuvent inclure des publications, des commentaires, des réactions, des comptages de partages sur la publication. Il convient de noter que le scraping peut être sujet à des préoccupations juridiques et éthiques, en particulier si les données sont utilisées d'une manière qui viole les conditions d'utilisation de Facebook. ✓ **Qualité des données** : Les données extraites sont de format textuelle, écrites en arabe, et dont le sujet est relié aux théories du complot. ✓ **Organisation des données** : Les données extraites ont été organisées et sauvegardées au format CSV.

3.4.2. Description du jeu de données final

La base de données finale comprend des textes en langue arabe (arabe standard et arabe dialectale) qui soulèvent et promeuvent divers sujets de théories du complot. Elle contient 2048 documents collectés à partir de 38 pages et groupes actifs sur Facebook à l'aide des outils de scraping développés pour cette plateforme sociale. Les principaux sujets qui se sont discutés dans le corpus final incluent :

- **Pilule rouge (***en anglais,* **Red Pill) et anti-féminisme** : Ce sujet explore le concept de la « pilule rouge », issu du film « Matrix » et fait référence à la prise de conscience des vérités cachées. Dans le contexte des théories du complot, cela implique souvent un scepticisme à l'égard des récits traditionnels, en particulier en ce qui concerne la dynamique des sexes et le féminisme. Certains théoriciens du complot dans ce domaine soutiennent qu'il existe un effort organisé pour supprimer certaines perspectives ou manipuler les structures sociétales.

- **Terre plate (***en anglais,* **Flat Earth)** : La théorie de la Terre plate affirme que la Terre est plate plutôt que sphérique, contrairement au consensus scientifique. Les théories du complot liées à la Terre plate impliquent souvent des revendications d'une dissimulation mondiale massive orchestrée par les gouvernements et les institutions scientifiques. Les partisans de cette théorie remettent en question l'authenticité de l'exploration spatiale et des images satellites, affirmant qu'elles font partie d'un grand complot visant à tromper le grand public.

- **Organisations et sociétés secrètes** : Ce sujet explore les théories du complot entourant les sociétés secrètes, tels que la franc-maçonnerie ou les Illuminati. Les théoriciens du complot spéculent sur leurs agendas cachés et leur influence sur les événements mondiaux, attribuant souvent des motifs néfastes et un vaste contrôle à ces organisations. De telles théories suggèrent que ces groupes manipulent la politique, l'économie et même façonnent les événements mondiaux dans les coulisses.

- **Politique** : Les théories du complot liées à la politique englobent un large éventail d'idées. Ils tournent souvent autour de la croyance que les gouvernements, les politiciens ou les individus influents se livrent à des activités secrètes pour maintenir le pouvoir, contrôler l'opinion publique ou saper les processus démocratiques. Ces théories peuvent impliquer de prétendues

opérations sous fausse bannière, la surveillance gouvernementale ou l'existence de structures de pouvoir cachées tirant les ficelles des affaires politiques.

- **Religion et prophéties** : Les théories du complot dans des contextes religieux impliquent souvent des interprétations de textes sacrés, de prophéties ou de personnalités religieuses. Ils peuvent suggérer des connaissances cachées sur des plans divins, des scénarios apocalyptiques ou la manipulation d'enseignements religieux par des entités puissantes. De telles théories peuvent entremêler les croyances religieuses avec des éléments politiques, sociaux ou surnaturels pour étayer leurs revendications.

- **Économie** : Les théories du complot liées à l'économie se concentrent sur les activités secrètes présumées d'institutions financières, de sociétés multinationales ou d'entités gouvernementales. Ces théories prétendent souvent que les crises économiques, les fluctuations boursières ou les disparités de richesse sont des résultats intentionnels orchestrés par une élite cachée cherchant à consolider le pouvoir, à contrôler les ressources ou à mettre en œuvre un programme spécifique.

- **Santé** : Les théories du complot concernant la santé touchent à des sujets tels que les vaccins, l'industrie pharmaceutique et la recherche médicale. Ces théories propagent souvent des explications alternatives ou des soupçons sur l'efficacité, la sécurité ou les intentions cachées derrière certaines interventions médicales. Certains théoriciens affirment que les grandes organisations de santé collaborent avec les gouvernements pour dissimuler des effets nocifs ou promeuvent intentionnellement certains traitements pour des arrière-pensées.

3.4.3. Fonctions de nettoyage et de prétraitement de données

Avant d'exploiter les données extraites pour l'objectif de notre projet, elles doivent passer par les étapes de nettoyage et de prétraitement qui les rendent exploitable pour les algorithmes de modélisation thématique.

Les fonctions que nous avons appliquées pour prétraiter les données sont :

- Supprimer les textes hors-sujet : C-à-d les textes dont les sujets qui ne sont pas reliés aux théories de complot.
- Supprimer les textes incompréhensibles.
- Supprimer les textes de courte langueur : les textes contenant moins de quatre mots.
- Corriger les fautes d'écriture.
- Supprimer les « hashtags ».
- Enlever les espaces supplémentaires.
- Supprimer les Stopwords : comme " في ", " ثم ".
- Supprimer la ponctuation arabe et non arabe.
- Supprimer les caractères spéciaux.
- Supprimer les URLs.
- Supprimer les Emojis.
- Supprimer les mentions (@).
- Enlever « tashkeel ».
- Enlever « tatweel » : supprimer l'extension de ligne entre les lettres arabes.
- Normaliser la ligature : le remplacement de " ء " par " ئ ".
- Normaliser la « hamza » : le remplacement de la lettre finale " ة " par " ه " et la lettre finale " ي " avec " ى ".

3.4.4. La lemmatisation

Afin d'appliquer la lemmatisation aux textes arabes, nous avons utilisé le **package Qalsadi**. Il s'agit d'un package qui comprend des outils d'analyse logique des textes arabes appliqués dans la lemmatisation et d'analyse morphologique. Le lemmatiseur Qalsadi convertit les mots dans leur forme de base à l'aide du dictionnaire linguistique joint au package, par exemple en supprimant le " ي " de la racine du mot " يكتب " pour devenir " كتب " ou remettre le mot à son poids d'origine comme كتبَ - يَكْتُبُ - كَتَابَةٌ ← كتب [31].

3.4.5. La racinisation

En ce qui concerne la langue arabe, les techniques implémentées de racinisation des textes qui existent sont essentiellement la technique basée sur la racine (root-based) et la technique basée sur la racinisation légère (light stemmer).

1. **Les stemmers basés sur la racine (Root-based stemmers) :** ce type de stemmers décolle tout d'abord les préfixes et les suffixes, puis il vérifie la liste des formats et des racines pour déterminer si le mot resté après l'élimination de ces deux parties correspond à une racine ou un format connu. Si c'est le cas, il restaure la racine. Sinon, il renvoie le mot réel, pas le modificateur.
2. **Les stemmers légers (Light stemmer) :** C'est le processus de supprimer les préfixes et les suffixes des mots en se basant sur une liste spécifique sans savoir si le résidu est une racine, ce qui peut se traduit dans certains cas par une racine incorrecte telle que le mot " بستان " qui peut devenir " بستا ".

Pour notre projet, nous avons utilisé quatre stemmers développés pour la langue arabe [42]. Ils sont montrés dans le Tableau 4.

Tableau 4. Comparaison des racines arabes utilisées.

Stemmers	**Techniques**	**Fonctionnement**
Tashaphyne	Light stemmer	Il fonctionne en normalisant d'abord les mots (forme standard) en préparation des tâches de recherche et d'indexation qui nécessitent une dérivation. Deuxièmement, la directive de syllabe (extraction de racine) est implémentée à l'aide d'une liste de recherche par défaut de suffixes arabes (كم...,ـة) qui permet différents niveaux de dérivation.
ARLStem	Light stemmer	Il s'appuie sur de nouvelles règles pour éliminer intelligemment les préfixes et les suffixes pour obtenir la racine. Cela commence par la normalisation des caractères, la suppression des suffixes et des préfixes. Ensuite, le traitement des formes de pluriel, le féminin et le verbe pour extraire la racine du mot.

Khoja	Root-based	Il divise le mot en un préfixe, une racine et un suffixe, et le compare à une liste de suffixes et de préfixes, puis analyse et extrait la racine en la comparant aux modèles stockés dans le dictionnaire racine.
ISRI Stemmer	Root-based	C'est un dérivé basé sur la grammaire qui découle du mot selon des règles spécifiques pour trouver sa racine et est similaire au dérivé de Khoja mais sans l'utilisation du dictionnaire racine.

3.4.6. Vectorisation du texte (BOW & TF-IDF)

Nous avons exploité deux techniques pour la vectorisation du texte :

1. **TF-IDF :** Est une mesure statistique numérique utilisée pour évaluer l'importance d'un mot (terme) dans n'importe quel contenu d'une collection de documents en fonction des occurrences de tous les mots, et analyse également le niveau de pertinence des mots-clés utilisés dans le contenu donné. Termes/Documents Fréquence considère non seulement la fréquence mais induit également une information discriminante pour chaque terme. Comme mentionné dans l'équation Termes/Documents Fréquence fait référence au nombre de fréquences d'un mot particulier dans un document est divisé par la somme totale des mots dans le document sélectionné [23].
   → $\boldsymbol{TF}_{\boldsymbol{i,j}}$= nombre d'occurrences de mot dans les documents / nombre de mots dans tous les documents.
   → $\boldsymbol{IDF}_{\boldsymbol{i,j}}$= log (nombre de documents / nombre de documents contenant le mot).
   → $(\boldsymbol{TF\_IDF})_{\boldsymbol{i,j}} = \boldsymbol{TF}_{\boldsymbol{i,j}} \times \boldsymbol{IDF}_{\boldsymbol{i,j}}$ **[2**4].
2. **BOW :** (*en anglais*, Bag-Of-Words) Le modèle de sac de mots représente chaque document texte comme un vecteur de nombre de mots. Cela signifie que chaque document est représenté comme un ensemble de nombres de mots, chaque mot étant une dimension dans le vecteur. Le modèle est appelé un "sac de mots" car il traite chaque mot comme une entité distincte, sans tenir compte de la grammaire ou de l'ordre des mots [32].

t_i, d est la fréquence (le nombre d'occurrences) du terme t_i dans le document d.

Cette représentation peut être modélisée comme étant une projection d'un document d dans un espace de haute dimension :

$\rightarrow \boldsymbol{d} \leftrightarrow \emptyset(\boldsymbol{d}) = \left(\boldsymbol{tf}\ (\boldsymbol{t_1}, \boldsymbol{d}), \boldsymbol{tf}\ (\boldsymbol{t_2}, \boldsymbol{d}), \ldots, \boldsymbol{tf}\ (\boldsymbol{t_n}, \boldsymbol{d})\right) \boldsymbol{\epsilon}\ \boldsymbol{R^n}$ [24].

3.5. Méthodes utilisées pour la modélisation thématique

3.5.1. Méthodes de base

Nous avons implémenté trois méthodes de base pour la modélisation thématique : LDA, LSA et NMF. Les algorigrammes de ces techniques sont indiqués ci-dessous :

- **Modèle LDA**

Algorithme 1: Algorithme du Modèle LDA [25]

1. Initialiser le nombre de sujets K, le paramètre Dirichlet α et le paramètre Dirichlet β.
2. Pour chaque document d :
    a. Initialiser les proportions de sujets θ_d.
    b. Pour chaque mot w dans le document d :
        i. Assigner un sujet z au mot w.
3. Répéter jusqu'à convergence :
    a. Pour chaque document d :
        i. Mettre à jour les proportions de sujets θ_d en fonction des sujets assignés aux mots dans d.
    b. Pour chaque sujet k :
        i. Mettre à jour la distribution de sujets β_k en fonction des mots assignés au sujet k dans tous les documents.
4. Retourner les distributions de sujets θ_d et de mots β_k.

- **Modèle LSI**

Algorithme 2: Algorithme du Modèle LSI [28]

1. Créer une matrice terme-document où chaque ligne représente un terme et chaque colonne représente un document.
2. Normaliser les lignes de la matrice pour avoir une longueur unitaire.
3. Effectuer une décomposition en valeurs singulières (SVD) sur la matrice.
4. Choisir le nombre de dimensions à conserver.
5. Créer de nouveaux vecteurs de document en multipliant la matrice tronquée par la transposition de la matrice singulière droite.
6. Créer de nouveaux vecteurs de termes en multipliant la transposition de la matrice tronquée par la matrice singulière gauche.

- **Modèle NMF**

Algorithme 3: Algorithme du Modèle NMF [26]

1. Créer une matrice non-négative A de dimensions $m * n$, où m est le nombre de termes et n est le nombre de documents.
2. Initialiser deux matrices non-négatives W et H de dimensions $m * k$ et $k * n$, respectivement, où k est le nombre de facteurs latents choisi.
3. Répéter les étapes 4 à 6 pour un nombre fixe d'itérations ou jusqu'à la convergence.
4. Mettre à jour la matrice W en utilisant la formule suivante :
$W <- W * ((A * H') ./ (W * (H * H')))$
5. Mettre à jour la matrice H en utilisant la formule suivante :
$H <- H * ((W' * A) ./ ((W' * W) * H))$
6. Normaliser les colonnes de la matrice H pour qu'elles aient une somme de 1.
7. Retourner les matrices W et H, qui représentent les poids des termes et des documents pour chaque facteur latent.

3.5.2. Méthodes avancées

Nous avons implémenté deux méthodes de modélisation thématique avancées : BerTopic et Top2Vec. Les pseudocodes de ces algorithmes sont indiqués ci-dessous :

- **Modèle BerTopic**

Algorithme 4: Algorithme du Modèle BerTopic [29]

Entrée : Liste des documents
Sortie : Liste des sujets avec leurs documents correspondants
1. Charger le modèle et le tokenizer BERT pré-formés
2. Créer des intégrations de documents à l'aide du modèle BERT
3. Effectuer une réduction de dimensionnalité à l'aide d'UMAP
4. Regrouper des documents à l'aide de HDBSCAN
5. Obtenir la représentation du sujet pour chaque document à l'aide de l'intégration UMAP et de l'étiquette de cluster HDBSCAN
6. Calculer la cohérence du sujet pour chaque sujet
7. Trier les sujets par score de cohérence
8. Affecter les documents au meilleur sujet

- **Modèle Top2Vec**

Algorithme 5: Algorithme du Modèle Top2Vec [27]

Entrée : Corpus de documents
1. Tokéniser les documents
2. Former un modèle Word2Vec sur les documents tokenisés
3. Utiliser le modèle Word2Vec formé pour générer des vecteurs de document pour chaque document du corpus
4. Appliquer l'algorithme Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) pour regrouper les vecteurs de document
5. Pour chaque cluster, extraire les documents les plus représentatifs tels que déterminés par la similarité cosinus avec le centroïde du cluster
6. Calculer les incorporations des documents les plus représentatifs et regroupez-les avec HDBSCAN pour obtenir l'ensemble final de sujets
Sortie : Liste de sujets, où chaque sujet est représenté par un ensemble de documents les plus représentatifs

3.6. Métrique d'évaluation

L'évaluation quantitative et l'évaluation qualitative sont deux approches distinctes utilisées pour évaluer les résultats des méthodes de modélisation thématique.

L'évaluation quantitative implique l'utilisation de métriques et de mesures numériques pour évaluer la qualité et la cohérence des sujets générés. Ces métriques peuvent inclure des mesures

telles que les scores de cohérence, la perplexité ou des mesures statistiques de la distribution des sujets. L'évaluation quantitative fournit une évaluation systématique et objective des performances de l'algorithme sur la base de critères prédéfinis. Il permet des comparaisons quantitatives entre différents modèles ou variations de paramètres, permettant aux chercheurs d'identifier le modèle qui obtient les scores les plus élevés en fonction de la métrique sélectionnée.

D'autre part, **l'évaluation qualitative** implique un jugement humain et une évaluation subjective des sujets générés. Cette approche d'évaluation implique généralement des experts humains ou des évaluateurs qui évaluent les sujets en fonction de leur interopérabilité, de leur pertinence et de leur cohérence. L'évaluation qualitative se concentre sur la compréhension humaine inhérente des sujets et leur signification dans le contexte de l'ensemble de données ou du domaine de recherche analysé. Il fournit des informations sur l'interopérabilité humaine et la convivialité des sujets générés, capturant des aspects qui peuvent ne pas être capturés par des mesures quantitatives seules.

Les mesures quantitatives fournissent des évaluations objectives de la cohérence et de l'homogénéité, tandis que l'évaluation qualitative garantit que les sujets correspondent à la compréhension humaine et répondent aux objectifs visés de l'analyse. La combinaison des deux approches aide les chercheurs à prendre des décisions éclairées et à affiner le processus de modélisation des sujets pour des résultats optimaux [45].

3.6.1. La cohérence

La cohérence est une mesure quantitative généralement utilisée pour évaluer les modèles d’analyse thématique en mesurant le degré de similarité sémantique des mots dans une thématique. Dans ce cas, les sujets sont représentés par les N premiers mots ayant la probabilité la plus élevée d’appartenir à ce sujet particulier. En bref, le score de cohérence mesure à quel point ces mots sont similaires les uns aux autres [24].

Il existe deux types très utilisés de cohérence :

1. **UCI :** Dans la mesure UCI, chaque terme unique est associé à chaque autre terme. Il utilise des informations mutuelles ponctuelles :

$$\boldsymbol{UCI = \log \frac{P(w_i, w_j)+1}{P(w_i)+P(w_j)}}.$$

Où :

— $P(w)$ Représente la probabilité que w soit présent dans un document aléatoire.

— $P(w_i, w_j)$ Représente la probabilité que w_i et w_j soient présents dans le même document.

2. **Umass :** Elle sert à comparer seuls les termes précédents et suivants d'un terme donné :

$$\boldsymbol{Umass = \log \frac{D(w_i, w_j)+1}{D(w_i)}}.$$

Où :
— $D(w_i)$ représente le nombre de documents contenant le terme w_i .
— $D(w_i, w_j)$ représente le nombre de documents contenant les termes w_i et w_j.

Si le score Umass est bas et le score UCl en cohérence est élevé, meilleur est le modèle [33].

3. **C_V :** L'une des mesures de cohérence les plus populaires est appelée CV. Il crée des vecteurs de contenu de mots en utilisant leurs cooccurrences et, après cela, calcule le score en utilisant les informations mutuelles ponctuelles normalisées (NPMI) et la similarité cosinus [41].

3.7. Conclusion

Dans ce chapitre, nous avons défini les objectifs de notre projet et la méthodologie utilisée pour y parvenir, puis l'étape de modélisation thématique, qui consiste à appliquer des modèles de base et avancés aux données préparées, et enfin l'étape de l'évaluation de la qualité de modélisation.

Chapitre 04 : Implémentation et expérimentations

4.1. Introduction

Dans ce chapitre, nous présentons l'environnement de travail et les outils utilisés pour mettre en œuvre le projet. Nous discutons également et passons en revue les différentes sorties des résultats de la recherche.

4.2. Environnement de travail et outils d'implémentation

4.2.1. Matériel

Tableau 5. Caractéristiques du matériel utilisé.

Caractéristiques	**Poste de travail N°01**	**Poste de travail N°02**
PC	Acer	HP
Système d'exploitation	Windows 10 Professionnel	Windows 10 Professionnel
Processeur	Intel(R) Core(TM) i3-5005U CPU @ 2.00GHz 2.00 GHz	Intel(R) Celeron(R) N4020 CPU @ 1.10GHz 1.10 GHz
RAM	8,00 Go	8,00 Go
Type de système	SE 64 bits	SE 64 bits

4.2.2. Environnement de programmation

Dans la partie implémentation, nous avons utilisé le langage de programmation Python version 3.11.2 afin d'extraire les données de Facebook et implémenter notre système de modélisation thématique. **Python** est un langage de programmation open source de haut niveau, développé pour être utilisé avec une grande variété de systèmes d'exploitation. Il se caractérise comme un langage orienté objet, qui est utilisé dans de nombreux domaines, y compris le traitement du langage naturel (NLP) [34].

Le programme principal de notre projet a été codé à l'aide des éditeurs Python suivants :

- **Jupyter Notebook :** Jupyter Notebook est l'application Web originale pour créer et partager des documents informatiques Python. Il offre une expérience simple, rationalisée et centrée sur les documents [35].
- **Google Colab :** C'est un produit de Google Research. Colab permet à quiconque d'écrire et d'exécuter rapidement du code Python via le navigateur, et est particulièrement adapté à l'apprentissage automatique, à l'analyse de données et à l'éducation [36].

4.2.3. Les principaux packages python utilisés

Nous avons utilisé plusieurs packagent pour réaliser ce projet, parmi lesquels on cite :

- **Gensim :** Gensim est une bibliothèque Python libre et open source permettant de représenter des documents sous forme de vecteurs sémantiques. Gensim est conçu pour traiter du texte numérique brut et non structuré à l'aide d'algorithmes d'apprentissage automatique non supervisés [37].
- **Matplotlib** : est une bibliothèque du langage de programmation Python destinée à tracer et à visualiser des données sous formes de graphiques.
- **PyArabic** : une bibliothèque de langue arabe spécifique pour Python. Elle fournit des fonctions de base pour manipuler les lettres et le texte arabes, comme détecter les lettres arabes, les groupes et les caractéristiques de lettres arabes, supprimer les signes diacritiques, etc.
- **PyLDAvis** : une bibliothèque Python pour la visualisation interactive des modèles thématiques. PyLDAvis est conçu pour aider les utilisateurs à interpréter les thématiques dans un modèle de sujet qui a été adapté à un corpus de données textuelles. Le package extrait des informations d'un modèle de sujet LDA adapté pour informer une visualisation Web interactive [24].
- **Package Facebook-scraper :** c'est un outil qui permet d'extraire des données de plusieurs sources Facebook comme les pages, les groupes et les profils qui sont accessibles en public [30].

- **Package Facebook-page-scraper :** un outil conçu pour extraire des données de pages spécifiques sur Facebook, accessibles via le Web sans utiliser de cookies [30].
- **NLTK :** NLTK est une plate-forme leader pour la création de programmes Python pour travailler avec des données de langage humain. Il fournit une suite de bibliothèques de traitement de texte pour la classification, la tokenisation, la radicalisation, le balisage, l'analyse et le raisonnement sémantique, des wrappers pour les bibliothèques NLP de niveau industriel [38].

4.3. Informations générales sur le jeu de données

4.3.1. Caractéristiques du jeu de données

Après collecte, analyse et nettoyage des données, nous avons obtenu une base de données dont les caractéristiques sont montrées dans le Tableau 6.

Tableau 6. Description du jeu de données final.

Nom de l'ensemble de données	Facebook_CT_Data
Nombres de lignes	2048
Nombre de colonnes	3
Type de données	Textuelles
Noms des colonnes	Text, Langue, Topic
Utilisation de la mémoire	48.1+ KB
Nombre de valeurs nulles	0
Le nombre de mots	525936
Sources	38 pages et groupes Facebook
Les sujets	Politique, économie, terre plate, pilule rouge, franc-maçonnerie, juifs, religion, féminisme, guerre, occultisme
La langue (ou le dialecte)	Arabe standard, algérien, égyptien, marocain, syrien

4.3.2. Le nuage de mots de données

Le nuage de mots (*en anglais*, Word Cloud) est une technique qui nous permet d'afficher les mots les plus courants dans un corpus de textes donné. La Figure 7 présente le nuage de mots de notre jeu de données, où plus les mots sont fréquents, plus leur taille de police est grande.

Figure 7. Les mots les plus fréquents dans la base de données.

Nous remarquons que le nuage de mots extrait reflète la diversité des concepts et des sujets dans la base de données étudiée. Les mots les plus fréquents apparaissent comme " الرجل ", " النساء", " الحقيقة ". La présence de ces mots indique que les sujets des débats portent sur le féminisme et les problèmes d'égalité des genres, en raison de l'intérêt de la société pour ces questions. Il existe également un groupe de mots associés à la sphère politique, tels que " الشعب ", " المنطقة ", " الدولة ". Dans l'ensemble, ce nuage de mots donne un aperçu des principaux concepts et sujets couverts par la recherche et contribue à la compréhension des principales tendances et modèles dans les données identifiées.

4.3.3. La distribution des longueurs des documents

La Figure 8 montre que le nombre moyen de mots dans un document typique dans notre jeu de données est de 257, soit plus de 1200 documents d'une longueur de 150 mots. Evidemment, plus le nombre de mots est grand, plus le nombre de documents est faible.

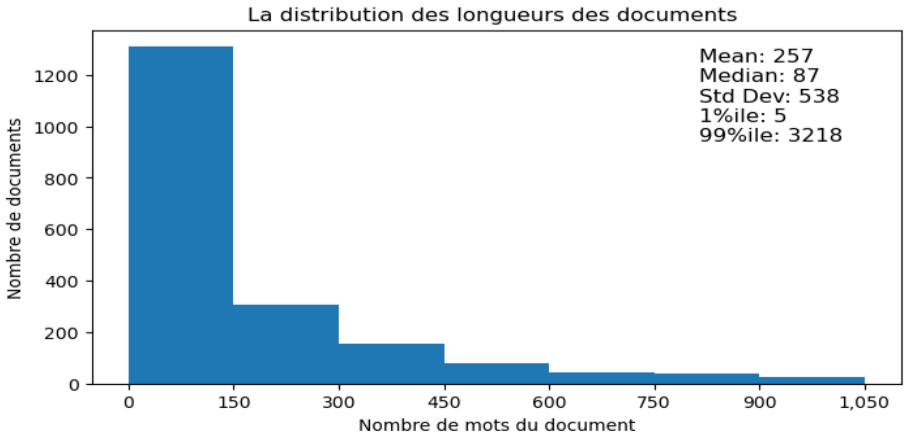


Figure 8. La distribution des longueurs des documents.

4.4. Processus de modélisation thématique

4.4.1. Paramétrage des méthodes de base « LDA, LSA, NMF »

Les paramètres principaux des modèles thématiques LDA, LSA et NMF du package Gensim sont indiqués ci-dessous :

- **Corpus** : Flux de vecteurs de documents ou matrice creuse Termes/Documents.
- **Nombre de thématique** (num_topics) : Le nombre de sujets latents demandés à extraire du corpus d'entraînement.
- **Id2word** : Mappage des identifiants de mots aux mots. Il est utilisé pour déterminer la taille du vocabulaire, ainsi que pour le débogage et l'impression des sujets.
- **Passes (passages)** : Le nombre de fois qu'on doit déterminer pour parcourir ou passer sur la totalité du corpus pendant la phase d'apprentissage [24].

Pour une expérimentation de base, nous avons appliqué les modèles thématiques LDA, NMF et LSA avec num_topics = 04 et passes = 01. Sachant que la méthode LSA du package Gensim n'exige pas un nombre bien déterminé pour les passages. Elle a plutôt le paramètre booléen **onepass** qui doit être fixé à la valeur False pour forcer le multi-passage sur le corpus textuel.

4.4.2. Paramétrage des méthodes avancées Topic2Vec et BerTopic

Pour **BerTopic**, les paramètres ont été définis comme suit [39] :

- **Language :** permet de spécifier la langue des textes à traiter dans Bertopic. Ces informations sont utilisées pour sélectionner le modèle pré-formé approprié qui prend en charge la langue sélectionnée. Lorsque on définit une valeur pour 'langue', Bertopic charge automatiquement le modèle approprié pour analyser les sujets dans la langue sélectionnée, ce qui simplifie le processus d'utilisation du cadre et nous fournit les résultats attendus basés sur le modèle approprié [39].
- **Embedding model :** BerTopic commence par convertir les documents d'entrée en représentations numériques en utilisant des convertisseurs de phrases, car ils sont parfaitement capables de capturer la similitude sémantique entre les documents. Un modèle d'intégration est choisi pour convertir les documents en représentations numériques [40].
- **Nr_topics :** Représente le nombre de sujets à extraire.

Nous avons testé deux variantes de BerTopic ; BerTopic de base avec le paramètre de Language = Arabe, et BerTopic avec le modèle de language pré-entrainé AraBERT qui est destiné spécialement à la langue Arabe.

AraBERT, est un modèle de représentation de la langue arabe pour améliorer l'état de l'art dans plusieurs tâches de la NLP arabe. Ce paradigme est largement considéré comme la base des résultats les plus récents dans diverses tâches de la NLP dans plusieurs langues [47].

Pour **Top2Vec**, les paramètres ont été définis comme suit :

- **Documents** : Corpus d'entrée, doit être une liste de chaînes.
- **Topic_nums** : Le nombre de sujets à récupérer [44].

4.5. Résultats obtenus

4.5.1. Les thématiques présentes dans le corpus

Les tableaux suivants représentent les résultats obtenus après avoir appliqué LDA, LSI, NMF, Top2Vec et BerTopic respectivement. Chaque tableau montre les cinq premiers thématiques avec leurs mots clés et leurs distributions.

1. Modèle LDA

Tableau 7. Les résultats obtenus après l'application de LDA.

Topic00		Topic01		Topic02		Topic03		Topic04	
Mot clé	**Dist**	**Mot clé**	**Dist**	**Mot clé**	**Dist**	**Mot clé**	**Dist**	**Mot clé**	**Dist**
النساء	0.001	الرجل	0.001	النساء	0.001	الرجال	0.001	النسوية	0.001
الرجل	0.001	نسوية	0.001	النسوية	0.001	النساء	0.001	الرجال	0.001
تنتهك	0.000	امراه	0.001	الحب	0.001	الرجل	0.001	النساء	0.001
جاذبيتها	0.000	الرجال	0.001	تدرك	0.001	لحماية	0.001	وتارة	0.001
تستطيع	0.000	يتحكم	0.000	الرجل	0.001	اهتمام	0.001	الرجل	0.001
النضج	0.000	التقليدي	0.000	معايير	0.001	غريزة	0.001	هرمون	0.001
الصداقة	0.000	تفكر	0.000	ابنك	0.001	يجدها	0.001	معدلات	0.000
علمت	0.000	الجنسية	0.000	للعثور	0.000	المستقبل	0.000	الخصوبة	0.001

2. Modèle LSI

Tableau 8. Les résultats obtenus après l'application de LSI.

Topic00		Topic01		Topic02		Topic03		Topic04	
Mot clé	**Dist**	**Mot clé**	**Dist**	**Mot clé**	**Dist**	**Mot clé**	**Dist**	**Mot clé**	**Dist**
العربي	0.075	النساء	0.275	النساء	0.324	عاجل	0.458	عاجل	0.458
النساء	0.068	الرجال	0.227	الرجال	0.275	نووية	0.148	تركيا	0.190
الثالثة	0.067	الرجل	0.180	الرجل	0.183	الصين	0.145	فيروس	0.175
الاوسط	0.063	النسوية	0.130	السماء	0.180	فيروس	0.139	زلزال	0.165
المسيح	0.062	الشمس	0.126	النسوية	0.170	النووية	0.131	كورونا	0.162
الرجل	0.062	السماء	0.122	الشمس	0.170	كورونا	0.127	الزلزال	0.145
الرجال	0.059	عاجل	0.113	تعالى	0.124	زلزال	0.126	الصين	0.135
الجيش	0.059	امرأة	0.098	والارض	0.113	تركيا	0.124	ووهان	0.093

3. Modèle NMF

Tableau 9. Les résultats obtenus après l'application de NMF.

Topic00		Topic01		Topic02		Topic03		Topic04	
Mot clé	**Dist**	**Mot clé**	**Dist**	**Mot clé**	**Dist**	**Mot clé**	**Dist**	**Mot clé**	**Dist**
رئيس	0.001	النساء	0.001	النساء	0.001	الجيش	0.001	الشعب	0.003
السعودية	0.001	الرجل	0.001	هتلر	0.001	المصري	0.001	العربي	0.002
العربي	0.001	السماء	0.001	الرجال	0.001	ارض	0.001	رسول	0.001
نووي	0.001	الرجل	0.001	عاجل	0.001	روتشيلد	0.001	اليهودي	0.001
عبد	0.001	رجل	0.001	القمر	0.001	الطاقة	0.001	الصين	0.001
بوتين	0.001	الحقيقة	0.001	هرمز	0.001	العين	0.001	الدولة	0.001
الثالثة	0.001	شخص	0.001	نووية	0.001	اليهودية	0.001	الاوسط	0.001
بريطانيا	0.001	المجتمع	0.001	اوروبا	0.001	الروسي	0.001	العرب	0.001

4. Modèle Top2Vec

Tableau 10. Les résultats obtenus après l'application de Top2Vec.

Topic00		Topic01		Topic02		Topic03		Topic04	
Mot clé	**Scroll**	**Mot clé**	**Scroll**	**Mot clé**	**Scroll**	**Mot clé**	**Scroll**	**Mot clé**	**Scroll**
السماء	0.6333	النساء	0.7145	الإيرانية	0.6338	المساواة	0.7006	النووية	0.8060
مسطحة	0.6005	الرجال	0.7018	الربيع	0.5869	الاجتماعية	0.6658	سلاح	0.6318
كروية	0.5927	النسوية	0.6620	الامريكية	0.5849	الافكار	0.5991	النووي	0.6306
القران	0.5620	الزواج	0.6224	القوات	0.5805	المجتمع	0.5600	الزلازل	0.5935
السماوات	0.5398	الذكور	0.6023	العسكرية	0.5616	القوانين	0.5311	قنبلة	0.5285
القيامة	0.5311	الحب	0.5632	العربي	0.5421	الطبيعة	0.5004	هارب	0.5189
ولأرض	0.5094	التواصل	0.5466	القذافي	0.5413	الشمولية	0.4820	الجوي	0.5143
آيات	0.5044	المجتمع	0.4977	الاسرائيلي	0.5396	الحرية	0.4724	التقنية	0.4970

5. Modèle BerTopic de base

Tableau 11. Le résultat obtenu après l'application de BerTopic.

Topic00		Topic01		Topic02		Topic03		Topic04	
Mot clé	**Scroll**	**Mot clé**	**Scroll**	**Mot clé**	**Scroll**	**Mot clé**	**Scroll**	**Mot clé**	**Scroll**
الاوسط	0.0056	النساء	0.0074	روسيا	0.1474	زلزال	0.0572	الصين	0.2376
المنطقة	0.0054	الطاقة	0.0060	اكرانيا	0.1417	تركيا	0.0372	الصيني	0.1154
العربي	0.0052	الرجال	0.0059	الروسي	0.1134	حركة	0.0324	المال	0.0833
تعالى	0.0048	الرجل	0.0059	بوتين	0.0556	القران	0.0267	اسياد	0.0715
العرب	0.0047	الشمس	0.0053	الروسية	0.0538	الخسف	0.0230	المنطاد	0.0574
الوصف	0.0047	الشيطان	0.0051	الاكراني	0.0467	سورة	0.0221	البحرية	0.0554
الدولار	0.0047	الشعب	0.0050	الاتحاد	0.0447	رسول	0.0218	الحزب	0.0408
رئيس	0.0046	الدولة	0.0047	اوروبا	0.0411	الزلزال	0.0558	المتجسس	0.0368

6. Modèle BerTopic avec l'intégration de AraBERT

Tableau 12. Les résultats obtenus après l'application de BerTopic AraBERT.

Topic00		Topic01	
Mot clé	**Scroll**	**Mot clé**	**Scroll**
النساء	0.0156	العربي	0.0053
الرجال	0.0152	المسيح	0.0053
الرجل	0.0138	الدولة	0.0053
بوتين	0.0116	تعالى	0.0052
الشمس	0.0114	العرب	0.0051
السماء	0.0113	الطاقة	0.0050
الدولار	0.0047	الاوسط	0.0049
تركيا	0.0099	الشيطان	0.0049

4.5.2. Etude de performance des modèles thématiques

L'influence du prétraitement de données et du choix de paramètres sur les performances des algorithmes de modélisation thématique est cruciale. Un prétraitement approprié des données, y compris le nettoyage du texte, la tokenisation, la suppression des mots vides et la racinisation (stemming), joue un rôle essentiel dans la définition de la qualité et de la cohérence des sujets résultants. La façon dont les données sont prétraitées peut avoir un impact sur l'identification de

sujets significatifs et l'élimination du bruit ou des informations non pertinentes. De plus, le choix des paramètres, tels que le nombre de sujets, affecte de manière significative les résultats de la modélisation des sujets. Par conséquence, la sélection de valeurs de paramètres appropriées en fonction des caractéristiques du jeu de données et des objectifs de recherche est essentielle pour obtenir des résultats précis et interprétables. L'examen attentif et l'optimisation des étapes de prétraitement des données et des choix de paramètres sont fondamentaux pour maximiser l'efficacité et la fiabilité des algorithmes de modélisation de sujets.

4.5.2.1. Influence de pré-traitement de données sur la performance des modèles thématiques

Nous présentons dans la section suivante les scores de cohérence des modèles thématiques LDA, NMF, LSA, Top2Vec et BerTopic par rapport à cinq différentes critères d'évaluation : la technique de vectorisation du texte (TF-IDF ou BOW), la lemmatisation, la racinisation (le stemming), la lemmatisation et la racinisation ensemble, et le filtrage de mots moins fréquents.

→ La cohérence par rapport à la technique de vectorisation

Tableau 13. Les scores de cohérence obtenus pour les méthodes de base.

Modèle	Cohérence de base	Cohérence « BOW »
LDA	0.33	0.43
LSI	0.46	0.38
NMF	0.34	0.29
Top2Vec	0.61	0.61
BerTopic_Base	0.27	0.27
BerTopic_AraBERT	0.85	0.85

→ **La cohérence en appliquant la lemmatisation avec le lemmatiseur « Qalsadi »**

Tableau 14. Les scores de cohérence obtenus en appliquant la lemmatisation.

Modèles	**Cohérence de base**	**Cohérence « Lemmatisation »**
LDA	0.33	0.34
LSI	0.46	0.56
NMF	0.34	0.35
Top2Vec	0.61	**0.62**
BerTopic_Base	0.27	0.47
BerTopic_ AraBERT	0.85	0.71

→ **La cohérence en appliquant la racinisation avec les quatre stemmers : Tachaphyne, ISRIStemmer, ARLStemmer et Khoja Stemmer**

Tableau 15. Les scores de cohérence obtenus en appliquant la racinisation.

Modèles	**Cohérence de base**	**Cohérence Tachaphyne**	**Cohérence ISRIStemmer**	**Cohérence ARLStemmer**	**Cohérence Khoja**
LDA	0.33	0.34	0.29	**0.35**	0.33
LSI	0.46	0.47	0.45	**0.53**	0.52
NMF	**0.34**	0.29	0.31	0.31	0.31
Top2Vec	0.61	0.69	0.50	**0.70**	0.64
BerTopic_Base	0.27	0.37	0.08	**0.66**	0.46
BerTopic_ AraBERT	**0.85**	0.79	0.30	0.73	0,77

→ **La cohérence en appliquant la racinisation avec ARLStemmer et la lemmatisation avec Qalsadi**

Tableau 16. Les scores de cohérence obtenus en appliquant la racinisation et la lemmatisation.

Modèles	**Cohérence de base**	**Cohérence «ARLStemmer & Lemmatisation »**
LDA	0.33	0.36
LSI	0.46	0.47
NMF	0.34	0.33
Top2Vec	0.61	**0.65**
BerTopic_Base	0.27	0.54
BerTopic_ AraBERT	0.85	0.70

→ **La cohérence en appliquant le filtrage des mots moins fréquents**

Tableau 17. Les scores de cohérence obtenus en appliquant le filtrage des mots moins fréquents.

Modèles	**Cohérence de base**	**Cohérence « filtrage »**
LDA	0.33	0.30
LSI	0.46	0.40
NMF	0.34	0.23
Top2Vec	0.61	**0.55**
BerTopic_Base	0.27	0.27
BerTopic_ AraBERT	0.85	0.66

4.5.2.2. Influence du choix de nombre de thématiques sur la performance des modèles thématiques

Les figures 9, 10, 11, 12, 13 et 14 montrent les performances des modèles LDA, LSA, NMF et Top2Vec en fonction du nombre de sujets sélectionné selon les critères d'évaluation mentionnés précédemment. Notons qu'en BerTopic, on n'a pas l'option de choisir un nombre de sujets en préalable vue que ce paramètre est déterminé automatiquement par l'algorithme de BerTopic. Pour

chaque modèle, la valeur de cohérence la plus élevée correspond au nombre optimal de sujets à extraire.

→ **La performance de base des modèles thématiques en fonction du nombre de sujets**

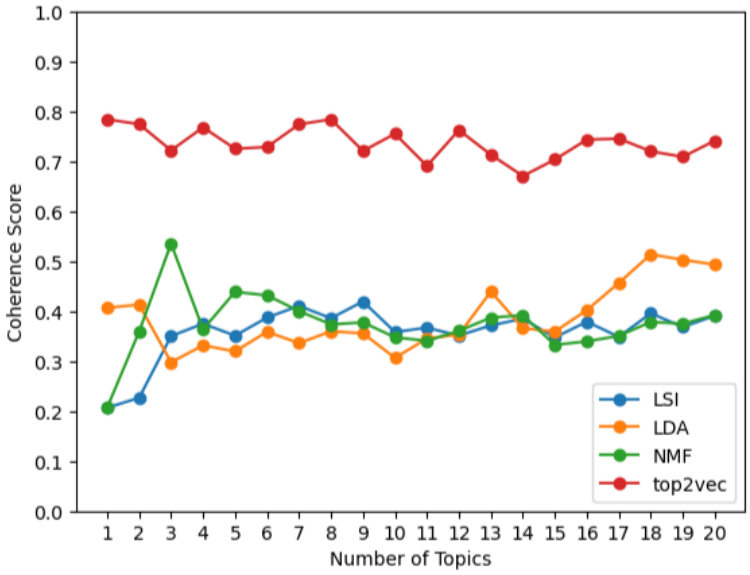


Figure 9. Le nombre optimal de thématique pour LDA, LSI, NMF, Top2Vec.

→ **La performance des modèles thématiques en fonction du nombre de sujets après avoir appliqué la lemmatisation Qalsadi**

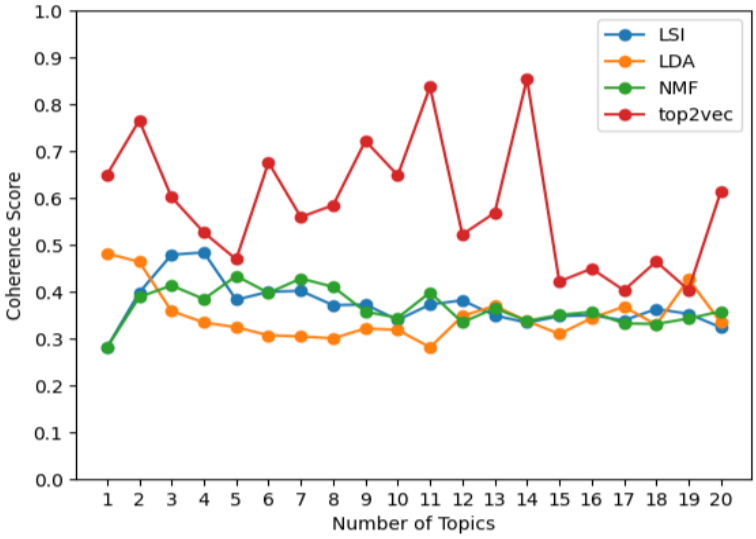


Figure 10. Le nombre optimal de thématique après avoir appliqué la lemmatisation.

→ **La performance des modèles thématiques en fonction du nombre de sujets après avoir appliqué la racinisation**

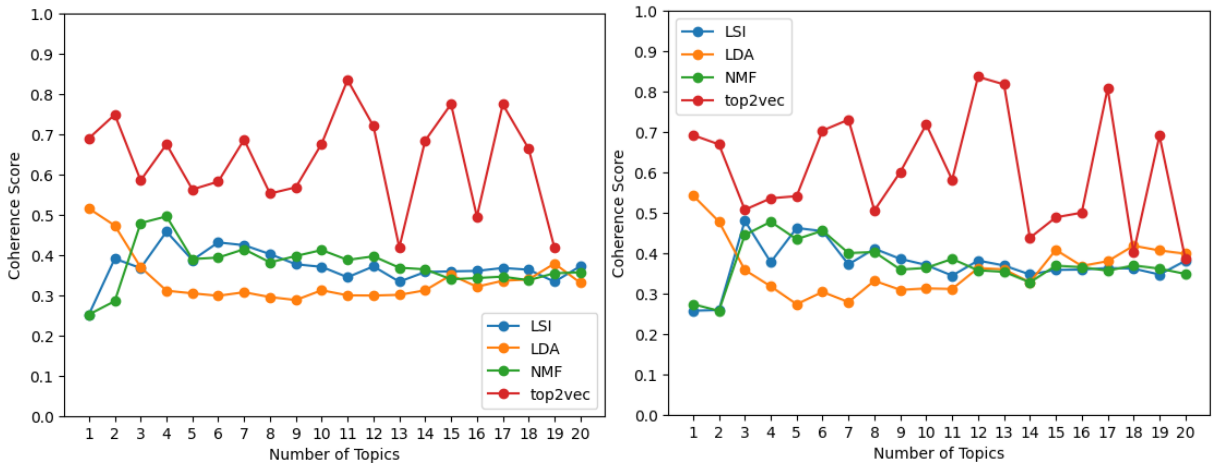


Figure 11. Le nombre optimal de thématique après avoir appliqué la stemming tashaphyne (gauche) et khoja (droite).

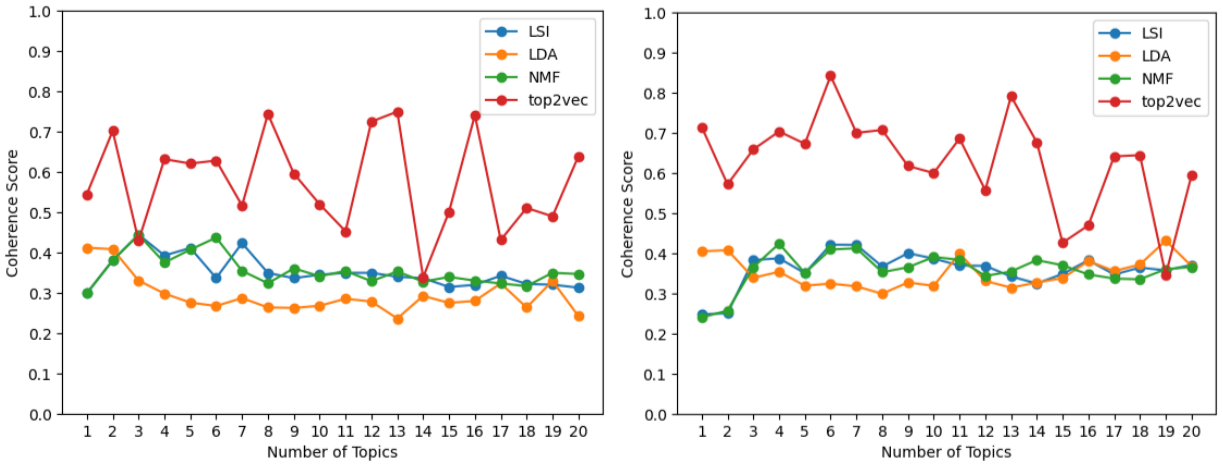


Figure 12. Le nombre optimal de thématique après avoir appliqué la stemming ISRIStemmer (gauche) et ARLStemmer (droite).

→ **La performance des modèles thématiques en fonction du nombre de sujets après avoir appliqué la stemming ARLStemmer & lemmatisation Qalsadi**

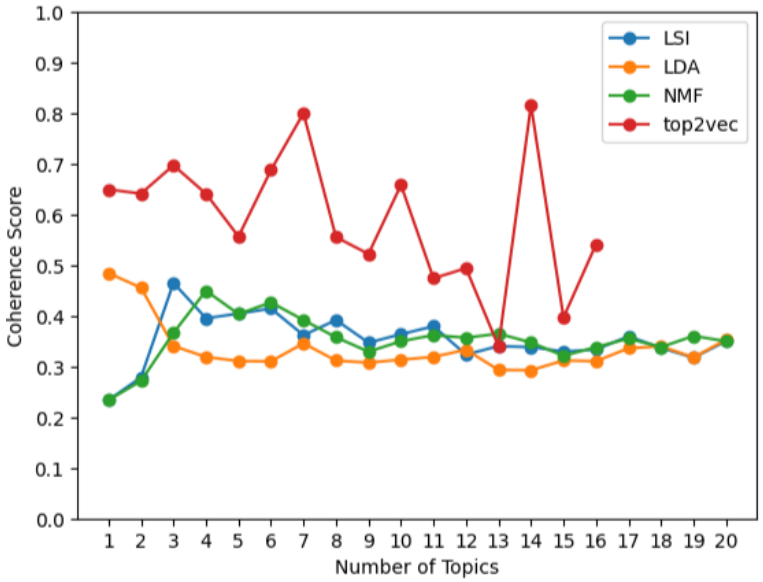


Figure 13. Le nombre optimal de thématique après avoir appliqué la ARLStemmer & lemmatisation.

→ **La performance des modèles thématiques en fonction du nombre de sujets après avoir appliqué le filtrage**

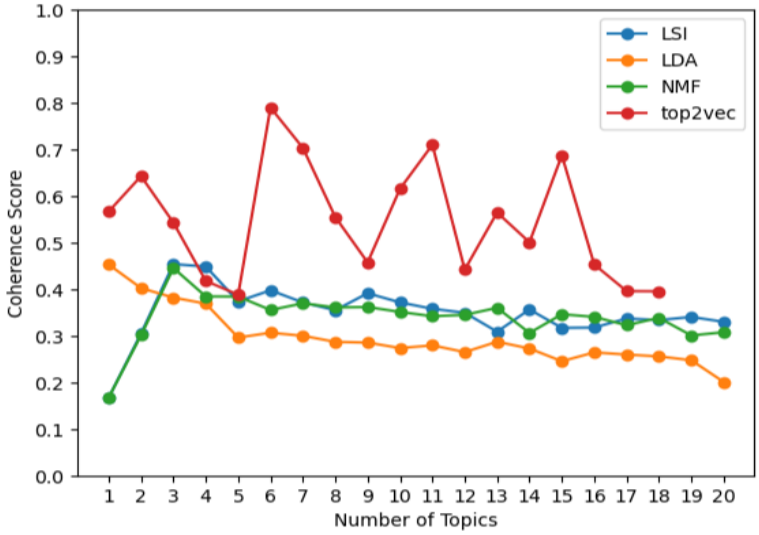


Figure 14. Le nombre optimal de thématique après avoir appliqué le filtrage.

4.6. Discussion générale des résultats

4.6.1. Description générale des performances des méthodes

Après avoir appliqué les algorithmes de modélisation thématiques à notre corpus en langue arabe, nous avons obtenu des résultats importants qui méritent attention. D'un point de vue général, les résultats ont montré que les méthodes avancées (BerTopic, Top2Vec) sont supérieures aux méthodes de base (LDA, LSI, NMF) dans la diversité des sujets extraits. Nous constatons que cette performance est dû principalement à l'utilisation de techniques d'apprentissage en profondeur dans ces modèles.

D'un point de vue spécifique, nous pouvons tirer de ces résultats deux remarques importantes. D'une part, en comparant les modèles de base entre eux, nous notons que LSI et NMF offrent des résultats plus variés et différenciés en termes de sujets extraits par rapport à LDA. Nous estimons que ces deux algorithmes utilisent des méthodes variées pour traiter les matrices textuelles et extraire les sujets correspondants. D'autre part, concernant les méthodes avancées, nous notons que, malgré la performance quantitative de BerTopic, le modèle Top2Vec est montré une performance quantitative et qualitative exceptionnelle. Ce modèle surpasse les autres en ce qui concerne la représentation des sujets et la fourniture d'une vue complète des données textuelles. Cela est dû à son utilisation d'un style probabiliste dans la définition des sujets, ce qui signifie qu'il y a plus d'un sujet possible dans un seul texte, et les sujets peuvent être classé en fonction de leur importance [48].

4.6.2. Impact des fonctions de pré-traitement sur les performances des modèles thématique

- **Discussion des résultats de base**

Après avoir calculé le taux de cohérence de base, nous avons constaté que LSI pour les méthodes de base et BerTopic pour les méthodes avancées, en s'appuyant sur le modèle linguistique arabe « AraBERT », avaient atteint le taux de cohérence le plus élevé.

L'augmentation de la cohérence dans LSI est due à sa méthode de réduction de dimension dans la matrice textuelle, ce qui améliore la représentation des sujets et conduit à une augmentation de la cohérence entre les résultats [49].

En utilisant AraBERT avec BerTopic, le taux de cohérence augmente car AraBERT est entraîné sur la langue arabe et fonctionne efficacement dans l'analyse des textes arabes. BerTopic utilise un modèle probabiliste pour identifier les sujets, ce qui lui permet de gérer la présence de sujets multiples dans les textes. Cela contribue à améliorer la représentation des sujets et à augmenter la cohérence des résultats [50].

D'un point de vue qualitatif, les résultats de Top2Vec peuvent être pris en compte car ce modèle peut gérer plusieurs sujets dans un même texte, offrant ainsi une vision complète des données textuelles [44].

- **Discussion de l'impact de la lemmatisation**

Après avoir effectué une lemmatisation sur les textes avec le lemmatiseur arabe Qalsadi, nous avons constaté une amélioration des résultats pour LSI, BerTopic_Base, NMF, LDA et Top2Vec. En fait, la lemmatisation améliore la représentation et l'unification des mots, ce qui renforce la capacité de ces méthodes à extraire les significations essentielles et à classer les sujets de manière plus précise [51].

En revanche, l'efficacité de BerTopic avec AraBERT a diminué, ce qui peut être attribué à une distorsion de la représentation linguistique des textes et à un impact négatif sur l'extraction des sujets en raison de l'incompatibilité entre la technique de lemmatisation du lemmatiseur Qalsadi et du modèle de Language pré-entraîné AraBERT [50].

- **Discussion de l'impact de la racinisation**

Après avoir appliqué les quatre stemmers aux modèles, nous avons noté ce qui suit :

Les résultats de LDA, LSI, BerTopic_Base et Top2Vec ont tous été améliorés lors de l'utilisation des stemmers Tachaphyne, ARLStemmer et Khoja. Cela est dû à leur capacité à réduire les mots à leurs racines de base, ce qui réduit la variation dans la représentation et facilite

l'agrégation des mots liés sous un même sujet. De plus, l'amélioration de la représentation linguistique des mots aide les méthodes à extraire des sujets plus précis et distincts des textes traités.

En revanche, les méthodes NMF et BerTopic_AraBERT ne se sont pas bénéficié de ces stemmers. Cela peut être dû à une incompatibilité entre la technique de racinisation et la représentation linguistique, ce qui entraîne une distorsion des mots et a un impact négatif sur l'extraction des sujets.

Reste à noter que ISRIStemmer n'a rien amélioré dans les résultats de performance de nos algorithmes thématiques. La raison pourrait être que cette stemmer est lourd et que l'efficacité du processus de réduction des mots n'était pas suffisante pour améliorer la représentation linguistique et l'extraction précise des sujets [52].

- **Discussion de l'impact de l'application de la lemmatisation et la racinisation ensemble**

En appliquant les deux fonctions de lemmatisation avec le lemmatiseur Qalsadi et de racinisation avec le stemmer ARLStemmer, nous avons observé une amélioration des résultats pour LDA, LSI, Top2Vec et BerTopic_Base. En effet, appliquer ces deux fonctions en même temps durant l'étape de pré-traitement de données aide à unir et à regrouper des mots similaires sous le même sujet et à réduire les variations de représentation linguistique.

Malheureusement, cette observation n'est pas valide pour le modèle de base NMF et le modèle avancé BerTopic_AraBERT, ce qui peut être dû à l'incompatibilité entre la technique de transformation racine utilisée par Qalsadi et ARLStemmer et la syntaxe utilisée dans ces deux modèles thématiques [50] [51] [52].

- **Discussion de l'impact de filtrage des mots moins fréquents**

Lorsque nous avons filtré les mots les moins fréquents de notre corpus textuel, nous avons constaté une diminution de la qualité des résultats pour toutes les modèles utilisés. Cela est dû à la suppression de certains mots importants et hautement significatifs, ce qui affecte la représentation précise des sujets et leur extraction correcte.

4.6.3. Impact de nombre de thématiques pré-determiné sur la performance des modèles thématique

Exécutant les algorithmes thématiques en variant, à chaque fois, la valeur du nombre de sujets à extraire, nous avons obtenu les résultats suivants :

- Pour LDA, le rapport de cohérence le plus élevé était de 0,51 lorsque le nombre de sujets était de **18**.
- Pour LSI, le rapport de cohérence le plus élevé était de 0,42 lorsque le nombre de sujets était de **9**.
- Pour NMF, le rapport de cohérence le plus élevé était de 0,54 lorsque le nombre de sujets était de **3**.
- Pour Top2Vec, le rapport de cohérence le plus élevé était de 0,78 lorsque le nombre de sujets était de **8**.
- Pour BerTopic, le rapport de cohérence le plus élevé était de 0,85 lorsque le nombre de sujets était de **2**.

La divergence des résultats de cohérence peut survenir en raison des hypothèses sous-jacentes, des algorithmes, des objectifs d'optimisation de chaque méthode et des caractéristiques de l'ensemble de données, telles que sa complexité, sa diversité et le niveau d'hétérogénéité des sujets. Certaines méthodes comme LDA, LSI et Top2Vec peuvent donner la priorité à la capture de détails plus fins et de subtiles variations de sujet, ce qui conduit à une préférence pour un plus grand nombre de sujets. À l'inverse, d'autres méthodes comme NMF et BerTopic peuvent mettre l'accent sur la cohérence et l'interopérabilité des sujets, favorisant un plus petit nombre de sujets qui capturent les thèmes les plus saillants et les plus distincts.

En fin de compte, la variabilité des résultats de cohérence optimale avec différents nombres préfixés de sujets dans diverses méthodes de modélisation de sujets souligne la nécessité d'une approche réfléchie et itérative. On doit soigneusement évaluer et sélectionner la méthode et le nombre de sujets qui correspondent le mieux à l'ensemble de données spécifique, aux objectifs de recherche et au niveau souhaité de granularité et d'interopérabilité du sujet.

4.6.4. Evaluation qualitative des résultats

L’évaluation quantitative sur la base des résultats de cohérence montre que BerTopic_AraBERT et Top2Vec sont plus efficaces que tous les autres modèles de base. Cela s’accorde avec les études dans le domaine de la modélisation thématique qui indiquent généralement que les modèles thématiques basés sur l’apprentissage en profondeur ont des performances supérieures à celle des modèles thématiques de base [46].

L’évaluation quantitative suggère aussi que BerTopic est plus performant que Top2Vec. Cependant, l’évaluation qualitative des thématiques générées par les deux modèles montre que BerTopic, qui a généré uniquement deux sujets, ne représente pas le meilleur choix pour notre jeu de données qui reflète une grande variété de sujets, y compris la politique, l'économie, la science, le féminisme, les organisations secrètes, la santé, la guerre et la religion.

Il est possible qu'un algorithme de modélisation thématique à l’exemple de BerTopic, avec un score de cohérence théoriquement élevé ne convainque pas un expert humain. Cet écart peut se produire pour plusieurs raisons. Les scores de cohérence sont calculés sur la base de mesures statistiques des modèles de cooccurrence des mots et de la similarité sémantique. Bien que des scores de cohérence élevés indiquent une représentation statistiquement cohérente des sujets, ils peuvent ne pas toujours correspondre à la compréhension humaine ou aux connaissances spécifiques à un domaine.

L’évaluation qualitative sur la base du jugement humain prend en compte des facteurs supplémentaires tels que les connaissances de base, la pertinence contextuelle et l'interopérabilité. Un algorithme de modélisation thématique peut générer des sujets qui obtiennent des scores de cohérence élevés sur la base de modèles statistiques, mais ces sujets peuvent manquer d'interprétations significatives ou ne pas saisir les nuances des concepts sous-jacents. Les évaluateurs humains sont capables d'identifier de telles lacunes et peuvent fournir des informations critiques sur l'utilité pratique et la validité des sujets générés.

Concernant notre cas d’étude, la bonne performance quantitative et qualitative de Top2Vec sur les données textuelles en langue arabe peut s'expliquer par plusieurs raisons. Tout d'abord, Top2Vec se distingue par sa capacité à extraire des thèmes qui correspondent mieux aux sujets. Il

peut regrouper de manière plus efficace les mots pertinents sous des concepts ou des sous-groupes au sein des thèmes principaux par rapport à BerTopic. Deuxièmement, Top2Vec est capable de traiter efficacement les difficultés propres à la langue arabe, telles que la pluralité linguistique, les structures de phrases complexes et les variations linguistiques. Cela contribue à obtenir une précision élevée dans l'extraction et la compréhension des thèmes. De plus, il est noté que Top2Vec surpasse BerTopic en termes de temps d'exécution. Il nécessite beaucoup moins de temps pour extraire les thèmes, ce qui se traduit par une augmentation de l'efficacité et des économies de temps [53] [54] [55].

Le Tableau 18 présente les thématiques finales générées par Top2Vec en considérant ses paramètres optimaux en termes de racinisation (stemmer = **ARLStemmer**) et de nombre optimal des sujets (nombre_sujets = **8**).

A partir des mots clés de chaque thématique, il est évident que Top2Vec était capable de détecter les huit principaux sujets des théories de complot suivants : **Féminisme** et **Anti-féminisme** (Topic 0), **Astronomie** (Topic 1), **Politique** du **Moyen-Orient** (Topic 2), **Religion** (Topic 3), **Phénomènes mystérieux** (Topic 4), **Organisations secrètes** (Topic 5), **Culture** (Topic 6), et **Politique** (Topic 7).

4.7. Conclusion

Ce chapitre contient l'environnement de travail et des informations générales sur le jeu de données, processus de modélisation thématique, les résultats obtenus, et enfin la discussion générale des résultats.

Tableau 18. Les thématiques finales générées par Topic2Vec.

Topic00		Topic01		Topic02		Topic03		Topic04		Topic05		Topic06		Topic07	
Mot clé	**Scroll**	**Mot clé**	**Scroll**	**Mot clé**	**Scroll**	**Mot clé**	**Scroll**	**Mot clé**	**Scroll**	**Mot clé**	**Scroll**	**Mot clé**	**Scroll**	**Mot clé**	**Scroll**
النساء	0.7481	المسطح	0.6970	عسكر	0.6152	رسول	0.7118	الفضاء	0.5741	رموز	0.5798	اخلاق	0.5559	الاممي	0.6210
الرجال	0.6660	النهار	0.4633	الايران	0.6056	الأحاديث	0.6025	مقاطع	0.5205	شعار	0.5495	الاجتماع	0.5275	البروتوكول	0.5996
النسو	0.6125	الاقمار	0.3674	الخليج	0.5467	هرير	0.5858	الغريب	0.5025	الماسون	0.5192	الليبرال	0.5192	السياس	0.5083
الرجل	0.5898	مسطح	0.6054	الصواريخ	0.5286	المهد	0.5607	مشاهد	0.4755	غامض	0.3268	المرء	0.5070	القوان	0.5068
الزواج	0.5709	الكواكب	0.6030	لحرب	0.5074	البخار	0.5479	اختف	0.4686	الشيطان	0.4219	الثقاف	0.4950	السلط	0.4652
الذكور	0.5623	الارض	0.5876	افغانسا	0.4828	يقتل	0.4934	الطائر	0.4685	طقوس	0.4202	المجتمع	0.4927	الصحاف	0.4487
الحب	0.5073	الشمس	0.5684	الاسرائيل	0.4826	صحيح	0.4725	الاطباق	0.4299	الرق	0.4011	الافكار	0.4587	الشعوب	0.4374
المجتمع	0.4637	السماء	0.5617	امريك	0.4489	رب	0.4677	الغامض	0.4296	الوثن	0.3788	شعور	0.4425	حكماء	0.4360
الزوج	0.4243	الفضاء	0.3979	الاستخبا	0.4644	سبحان	0.4104	تجرب	0.4288	المتنور	0.3689	ديمقراط	0.4230	الحروب	0.4225
اخلاق	0.4128	القمر	0.5426	الروس	0.4327	ذكر	0.4137	تكنولوج	0.4272	السر	0.3620	الحر	0.4118	السيطر	0.4175
عقل	0.3668	سطح	0.5311	قنبل	0.4270	يفتح	0.4147	اشخاص	0.4046	محفل	0.3581	الافراد	0.4073	الاجتماع	0.4164
اطفال	0.3682	الليل	0.4720	الدفاع	0.4216	قول	0.4214	مجهول	0.3846	الهيكل	0.3563	مبادئ	0.4072	الجماهير	0.3784
القيم	0.3847	كوكب	0.4648	القذاف	0.4100	المؤمن	0.4226	انتاركتيك	0.3401	شرك	0.3421	الانسان	0.4050	افساد	0.3659
ضعيف	0.3748	تدور	0.4086	استراتيج	0.4082	نزل	0.4260	ابحاث	0.3386	المسيح	0.3190	الحزب	0.3599	خداع	0.3587
الاجتماع	0.3649	ثابت	0.3991	الجيش	0.4041	بعث	0.4401	تصوير	0.3406	بابل	0.3081	الشيوع	0.3504	استغلال	0.3632

Conclusion générale

Les plates-formes de médias sociaux assistent à une large diffusion des théories du complot, et Facebook joue un rôle essentiel dans leur diffusion. Cette diffusion croissante a contribué à la multiplicité des sujets et des problèmes qu'elles couvrent pour inclure des conspirations politiques, économiques, sanitaires et sociales et bien d'autres. Malgré les efforts de Facebook pour lutter contre la propagation de la désinformation, le défi demeure dans leur impact sur l'opinion publique et la formation des croyances chez les utilisateurs.

Cette étude visait à découvrir des sujets latents liés à de fausses informations sur les théories du complot dans une collection de textes arabes extraits de Facebook, en utilisant diverses méthodes de modélisation thématiques telles que LDA, LSA, NMF, Top2Vec et BerTopic. Les sorties de cette étude ont mis en lumière les performances et la fiabilité de ces méthodes dans le contexte de l'analyse de la langue arabe.

Les résultats ont révélé que BerTopic présentait les performances quantitatives les plus élevées en termes de scores de cohérence. Cependant, les résultats obtenus avec la méthode Top2Vec étaient plus fiables et convaincants d'un point de vue qualitatif. Cela souligne l'importance de considérer à la fois les mesures d'évaluation quantitatives et qualitatives dans l'évaluation de l'efficacité des méthodes de modélisation thématique.

L'importance de cette étude réside dans sa contribution à la compréhension des complexités et des défis posés par la langue arabe dans le domaine de la modélisation thématique pour les théories du complot. L'arabe, avec ses caractéristiques linguistiques uniques et ses nuances culturelles, sa grammaire complexe, sa diversité dialectale, son écriture non standard et les significations multiples de certains de ces mots, présente des défis distincts qui nécessitent des approches sur mesure pour détecter et analyser les fausses informations.

Il est important de reconnaître les limites de cette étude. L'évaluation et la comparaison des méthodes de modélisation thématique étaient basées sur un jeu de données particulier et un ensemble prédéfini de critères d'évaluation. La possibilité de généraliser les résultats à d'autres ensembles de données ou contextes peut varier. De plus, le choix des méthodes de modélisation

thématique incluses dans cette étude n'est pas exhaustif, et d'autres méthodes émergentes pourraient justifier une exploration dans de futures recherches.

Les travaux futurs devraient se concentrer sur la résolution de ces limites et l'exploration de voies supplémentaires pour améliorer les performances des méthodes de modélisation thématique dans l'analyse de la langue arabe. Une enquête plus approfondie sur l'optimisation des paramètres, l'affinement des techniques de prétraitement et l'intégration de connaissances spécifiques au domaine peuvent potentiellement améliorer la précision et l'interopérabilité des sujets générés. De plus, l'exploration de méthodes d'apprentissage par ensemble ou d'approches hybrides qui combinent les puissances de plusieurs modèles pourrait être une direction prometteuse pour les recherches futures.

Les références

[1] C. Taguia and Z. Sid, “Détection de fausses informations sur le web et les réseaux sociaux,” *Mémoire de Master*, Université MOHAMED EL BACHIR EL IBRAHIMI-Bordj Bou Arreridj Faculté des Mathématiques et d’Informatique, 2020.

[2] M. Mancosu and S. Vassallo, “The life cycle of conspiracy theories: evidence from a long-term panel survey on conspiracy beliefs in Italy,” *Italian Political Science Review/Rivista Italiana di Scienza Politica*, vol. 52, no. 1, pp. 1-17, 2022, doi: https://doi.org/10.1017/ipo.2021.57.

[3] G. Koehler-Derrick, R. Nielsen, and D. Romney, “Conspiracy Theories in the Egyptian State-Controlled Press,” 2017. Available: https://aalims.org/uploads/conspiracy_10april2017_AALIMS.pdf (accessed Jun. 2, 2023).

[4] A. Heft and K. Buehling, “Measuring the diffusion of conspiracy theories in digital information ecologies,” *Convergence: The International Journal of Research into New Media Technologies*, vol. 28, no. 4, pp. 940–961, Apr. 2022, doi: https://doi.org/10.1177/13548565221091809.

[5] A. Fong, J. Roozenbeek, D. Goldwert, S. Rathje, and S. van der Linden, “The language of conspiracy: A psychological analysis of speech used by conspiracy theorists and their followers on Twitter,” *Group Processes & Intergroup Relations*, vol. 24, no. 4, pp. 606–623, May 2021, doi: https://doi.org/10.1177/1368430220987596.

[6] Y. Kim, “How conspiracy theories can stimulate political engagement,” *Journal of Elections, Public Opinion and Parties*, vol. 32, no. 1, pp. 1–21, Aug. 2019, doi: https://doi.org/10.1080/17457289.2019.1651321.

[7] K. M. Douglas et al., “Understanding Conspiracy Theories,” *Political Psychology*, vol. 40, no. S1, pp. 3–35, Feb. 2019, doi: https://doi.org/10.1111/pops.12568.

[8] J. P. Hughes, A. Efstratiou, S. R. Komer, L. A. Baxter, M. Vasiljevic, and A. C. Leite, “The impact of risk perceptions and belief in conspiracy theories on COVID-19 pandemic-related

behaviours," *PLOS ONE*, vol. 17, no. 2, p. e0263716, Feb. 2022, doi: https://doi.org/10.1371/journal.pone.0263716.

[9] Z. Zoccatelli, "الإشاعات ونظريات المؤامرة.. أسبابها ومخاطرها وجاذبيتها," *SWI swissinfo.ch*, https://www.swissinfo.ch/ara/society/%D8%B9%D9%84%D9%85-%D8%A7%D9%84%D9%86%D9%81%D8%B3%D8%A7%D9%84%D8%A7%D8%AC%D8%AA%D9%85%D8%A7%D8%B9%D9%8A-%D8%A7%D9%84%D8%A5%D8%B4%D8%A7%D8%B9%D8%A7%D8%AA-%D9%88%D9%86%D8%B8%D8%B1%D9%8A%D8%A7%D8%AA-%D8%A7%D9%84%D9%85%D8%A4%D8%A7%D9%85%D8%B1%D8%A9---%D8%A3%D8%B3%D8%A8%D8%A7%D8%A8%D9%87%D8%A7-%D9%88%D9%85%D8%AE%D8%A7%D8%B7%D8%B1%D9%87%D8%A7-%D9%88%D8%AC%D8%A7%D8%B0%D8%A8%D9%8A%D8%AA%D9%87%D8%A7/46568786 (accessed Jun. 2, 2023).

[10] J. B. Renard, "Les causes de l'adhésion aux théories du complot," *Diogène*, vol. 249–250, no. 1, p. 107, 2015, doi: https://doi.org/10.3917/dio.249.0107.

[11] J. W. van Prooijen and K. M. Douglas, "Belief in conspiracy theories: Basic principles of an emerging research domain," *European Journal of Social Psychology*, vol. 48, no. 7, pp. 897–908, Aug. 2018, doi: https://doi.org/10.1002/ejsp.2530.

[12] D. Jolley, S. Mari, and K. M. Douglas, "Consequences of Conspiracy Theories," *Northumbria University Research Portal*, Feb. 26, 2020. https://researchportal.northumbria.ac.uk/en/publications/consequences-of-conspiracy-theories (accessed Jun. 02, 2023).

[13] C. R. Sunstein and A. Vermeule, "Conspiracy Theories," *papers.ssrn.com*, Jan. 15, 2008. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1084585. (accessed Dec. 18, 2022).

[14] S. N. Mohammed, "Conspiracy theories and flat-earth videos on YouTube". *The Journal of Social Media in Society*, vol. 8, no. 8, pp. 84–102, 2019.

[15] I. Ullah, K. S. Khan, M. J. Tahir, A. Ahmed, and H. Harapan, “Myths and conspiracy theories on vaccines and COVID-19: Potential effect on global vaccine refusals,” *Vacunas*, vol. 22, no. 2, Mar. 2021, doi: https://doi.org/10.1016/j.vacun.2021.01.001.

[16] وليم كار, اليهود وراء كل جريمة. ktab INC., 2018.

[17] T. Ma, Raeed Al-Sabri, L. Zhang, Bockarie Daniel Marah, and Najla Al-Nabhan, “The Impact of Weighting Schemes and Stemming Process on Topic Modeling of Arabic Long and Short Texts,” *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 19, no. 6, pp. 1–23, Nov. 2020, doi: https://doi.org/10.1145/3405843.

[18] R. Albalawi, T. H. Yeap, and M. Benyoucef, “Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis,” *Frontiers in Artificial Intelligence*, vol. 3, Jul. 2020, doi: https://doi.org/10.3389/frai.2020.00042.

[19] Y. Hu, J. Boyd-Graber, B. Satinoff, and A. Smith, “Interactive topic modeling,” *Machine Learning*, vol. 95, no. 3, pp. 423–469, Oct. 2013, doi: https://doi.org/10.1007/s10994-013-5413-0.

[20] B. V. Barde and A. M. Bainwad, “An overview of topic modeling methods and tools,” *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2017, pp. 745–750, doi: 10.1109/ICCONS.2017.8250563.

[21] T. Ma, Raeed Al-Sabri, L. Zhang, Bockarie Daniel Marah, and Najla Al-Nabhan, “The Impact of Weighting Schemes and Stemming Process on Topic Modeling of Arabic Long and Short Texts,” *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 19, no. 6, pp. 1–23, Nov. 2020, doi: https://doi.org/10.1145/3405843.

[22] C. M. Amine and C. H. Djilali, “Conception et Réalisation d’un lemmatiseur hybride de texte arabe - PDF Téléchargement Gratuit,” *docplayer.fr*. https://docplayer.fr/62318325-Conception-et-realisation-d-un-lemmatiseur-hybride-de-texte-arabe.html (accessed Jun. 02, 2023).

[23] S. George and S. Vasudevan, “Comparison of LDA and NMF topic modeling techniques for restaurant reviews,” *Indian J. Nat. Sci*, vol. 10, no. 62, pp. 28210-28216, 2020.

[24] W. Bouhali and B. Ammara, “La modélisation thématique pour le texte arabe,” *Mémoire de Master*, Université MOHAMED EL BACHIR EL IBRAHIMI-Bordj Bou Arreridj Faculté des Mathématiques et d’Informatique, 2022.

[25] N. Seth, “Part 2: Topic modeling and Latent Dirichlet allocation (LDA) using Gensim and Sklearn,” *Analytics Vidhya*, Jun. 28, 2021. https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/ (accessed Jun. 2, 2023).

[26] V. Choubey, “Topic Modelling Using NMF,” *Voice Tech Podcast*, Jul. 21, 2020. Jul. 21, 2020. https://medium.com/voice-tech-podcast/topic-modelling-using-nmf-2f510d962b6e (accessed Jun. 02, 2023).

[27] A. Mavuduru, “How to perform topic modeling with top2vec,” *Medium*, Nov. 17, 2021. https://towardsdatascience.com/how-to-perform-topic-modeling-with-top2vec-1ae9bb4e89dc (accessed Jun. 2, 2023).

[28] J. Xu, “Topic modeling with LSA, PSLA, Lda & lda2Vec,” *Medium*, Dec. 20, 2018. https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05 (accessed May. 25, 2023).

[29] D. David, “NLP tutorial: Topic modeling in python with bertopic,” *HackerNoon.com*, https://hackernoon.com/nlp-tutorial-topic-modeling-in-python-with-bertopic-372w35l9 (accessed Apr. 10, 2023).

[30] K. Arellano, “What are Facebook scraper tools and why are they used?,” *The Last Call - RapidAPI Blog*, Dec. 14, 2020. https://rapidapi.com/blog/what-are-facebook-scraper-tools-and-why-are-they-used/ (accessed Mar. 10, 2023).

[31] T. Z. (طه زروقي), “Qalsadi Arabic Morphological Analyzer and Lemmatizer for Python,” *GitHub*, Jun. 02, 2023. https://github.com/linuxscout/qalsadi (accessed Apr. 18, 2023).

[32] “What is bag-of-words model?: AI terms explained - AI For Anyone,” *www.aiforanyone.org*. https://www.aiforanyone.org/glossary/bag-of-words-model (accessed Mar. 03, 2023).

[33] M. M. Bellaouar and I. E. Ghada, “Modélisation thématique: cas des publications scientifiques,” *Mémoire de Master*, Université de Ghardaïa Faculté des Sciences et de la Technologie, 2020.

[34] Y. Derfoufi, “Programmation en langage Python,” *hal.science*, May 12, 2019. https://hal.science/hal-02126596v1 (accessed Jun. 02, 2023).

[35] Jupyter, “Project Jupyter,” *Jupyter.org*, 2019. https://jupyter.org/ (accessed Apr. 19, 2023).

[36] Google, “Colaboratory – Google,” *research.google.com*. https://research.google.com/colaboratory/faq.html (accessed Mar. 21, 2023).

[37] “Gensim: topic modelling for humans,” *radimrehurek.com*. https://radimrehurek.com/gensim/intro.html (accessed Apr. 09, 2023).

[38] NLTK, “Natural Language Toolkit — NLTK 3.4.4 documentation,” *Nltk.org*, 2009. https://www.nltk.org/ (accessed Apr. 19, 2023).

[39] M. P. Grootendorst, “Parameter tuning - BERTopic,” *Github.io*, 2021. https://maartengr.github.io/BERTopic/getting_started/parameter%20tuning/parametertuning.html (accessed Mar. 31, 2023).

[40] M. P. Grootendorst, “Embeddings - BERTopic,” *maartengr.github.io*. https://maartengr.github.io/BERTopic/getting_started/embeddings/embeddings.html (accessed Apr. 15, 2023).

[41] E. Zvornicanin, “When Coherence Score is Good or Bad in Topic Modeling? | Baeldung on Computer Science,” *www.baeldung.com*, Dec. 07, 2021. https://www.baeldung.com/cs/topic-modeling-coherence-score (accessed May. 05, 2023).

[42] Y. A. Alhaj et al., “A study of the effects of stemming strategies on Arabic Document Classification,” *IEEE Access*, vol. 7, pp. 32664–32671, Mar. 2019. doi:10.1109/access.2019.2903331.

[43] A. Andrey, “What Is Text Vectorization? Everything You Need to Know,” *Deepset*, Dec. 03, 2021. https://www.deepset.ai/blog/what-is-text-vectorization-in-nlp (accessed May. 15, 2023).

[44] “Top2Vec — Top2Vec 1.0.29 documentation,” *top2vec.readthedocs.io*. https://top2vec.readthedocs.io/en/stable/Top2Vec.html (accessed Apr. 10, 2023).

[45] G. Claude, “Etude qualitative et quantitative - définitions et différences,” *Scribbr*, Oct. 14, 2019. https://www.scribbr.fr/methodologie/etude-qualitative-et-quantitative/ (accessed May. 21, 2023).

[46] H. Zhao et al., “Topic modelling meets Deep Neural Networks: A survey,” *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, Feb. 2021, doi:10.24963/ijcai.2021/638.

[47] A. Otmani and A. Bouras, “Architecture de transformateur pour la modélisation de la langue arabe,” *Mémoire de Master*, UNIVERSITE ECHAHID HAMMA LAKHDAR - EL OUED FACULTÉ DES SCIENCES EXACTES, 2022.

[48] R. Egger and J. Yu, “A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts,” *Frontiers in Sociology*, vol. 7, May 2022, doi: https://doi.org/10.3389/fsoc.2022.886498.

[49] “LSI / LSA المبرمج العربي - خوارزمية وممارسة مبدأ,” *arabicprogrammer.com*. https://arabicprogrammer.com/article/2391550817/ (accessed Jun. 04, 2023).

[50] A. Abuzayed and H. Al-Khalifa, “BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique,” *Procedia Computer Science*, vol. 189, pp. 191–194, 2021, doi: https://doi.org/10.1016/j.procs.2021.05.096.

[51] admin, “Benchmark on Arabic Embeddings for Topic Modeling,” *Bitext. We help AI understand humans. - chatbots that work*, Feb. 16, 2023. https://www.bitext.com/blog/benchmark-on-arabic-embeddings-for-topic-modeling 1/#:~:text=This%20benchmark%20focuses%20on%20how%20linguistic%20data%20affects (accessed Jun. 04, 2023).

[52] M. Naili, A. H. Chaibi, and H. H. Ben Ghezala, “Comparative Study of Arabic Stemming Algorithms for Topic Identification,” *Procedia Computer Science*, vol. 159, pp. 794–802, 2019, doi: https://doi.org/10.1016/j.procs.2019.09.238.

[53] B. Palese and G. Piccoli, “Evaluating topic modeling interpretability using topic labeled gold standard sets,” *Communications of the Association for Information Systems*, vol. 47, pp. 433–451, 2020. doi:10.17705/1cais.04720.

[54] Liqiang Niu, Xinyu Dai, Jianbing Zhang, and Jiajun Chen, “Topic2vec: Learning distributed representations of topics,” *2015 International Conference on Asian Language Processing (IALP)*, 2015. doi:10.1109/ialp.2015.7451564.

[55] N. C. Albanese, “Topic Modeling with LSA, pLSA, LDA, NMF, BERTopic, Top2Vec: a Comparison,” *Medium*, Sep. 22, 2022. https://towardsdatascience.com/topic-modeling-with-lsa-plsa-lda-nmf-bertopic-top2vec-a-comparison-5e6ce4b1e4a5 (accessed Jun. 04, 2023).

Annexe A

Sources de données

La section suivante présente les liens des pages et groupes Facebook duquel le jeu de données de cette étude a été extrait :

A.1. Pages

د.محمد بن الحسن

https://www.facebook.com/profile.php?id=100011969508210

كشف الاسرار

https://www.facebook.com/profile.php?id=100057573184709

Charlie Hebdo شارلي إيبدو

https://www.facebook.com/profile.php?id=100063499931507

الحبة الحمراءThe red pill

https://www.facebook.com/profile.php?id=100064839731893

□!!علوم آخر الزمان!!□
https://www.facebook.com/profile.php?id=100065620747769&paipv=0&eav=AfYPg6-xKLOMcKlOW2TNTqjCd8AnwSEGXWurziPVSs6rwwRbMU3YYDed0Z75Ay4Oh9Q

Meninism

https://www.facebook.com/menenism.dz?locale=fr_FR

ثوار ضد الرأسمالية المتوحشة و المستغلة
https://www.facebook.com/profile.php?id=100069157721350&paipv=0&eav=AfZZ6ASutXbjyMS5l_hUsLyRh8dSmVHhHXapqJ6v3PtLTg2ywZXfCtxC2uVdmeelMbI

الارض المسطحة

https://www.facebook.com/people/%D8%A7%D9%84%D8%A7%D8%B1%D8%B6-%D8%A7%D9%84%D9%85%D8%B3%D8%B7%D8%AD%D8%A9/100069344910238/?paipv=0&eav=AfbOrX5fMU8LPGkyUZ5dIuAtCXseDIpTIXgDue4pHlnDMYAhFMPsVoDJj-LMLF4JMSI&_rdr

الانوناكي والجن والحضارات القديمة والنظام العالمي الجديد
https://www.facebook.com/people/%D8%A7%D9%84%D8%A7%D9%86%D9%88%D9%8

6%D8%A7%D9%83%D9%8A-%D9%88%D8%A7%D9%84%D8%AC%D9%86-%D9%88%D8%A7%D9%84%D8%AD%D8%B6%D8%A7%D8%B1%D8%A7%D8%AA-%D8%A7%D9%84%D9%82%D8%AF%D9%8A%D9%85%D8%A9-%D9%88%D8%A7%D9%84%D9%86%D8%B8%D8%A7%D9%85-%D8%A7%D9%84%D8%B9%D8%A7%D9%84%D9%85%D9%8A-%D8%A7%D9%84%D8%AC%D8%AF%D9%8A%D8%AF/100070780278950/?paipv=0&eav=AfYMXIZdj6SAgdaTWL6liol9VAHwIbG7vCkWYcRSXD59PfN40tt0qa5x_vIJ4a4g7fI&_rdr

هدم خرافات علوم الفضاء

https://www.facebook.com/profile.php?id=100075926317822

آخر أيام الأرض. نهاية العالم

https://www.facebook.com/%D8%A2%D8%AE%D8%B1-%D8%A3%D9%8A%D8%A7%D9%85-%D8%A7%D9%84%D8%A3%D8%B1%D8%B6-%D9%86%D9%87%D8%A7%D9%8A%D8%A9-%D8%A7%D9%84%D8%B9%D8%A7%D9%84%D9%85-104687321126853/

افكار تنويرية

https://www.facebook.com/Afkar.Tanwiriya/videos/

Achraf Assal

https://www.facebook.com/simo.achraf.509/?locale=ar_AR

Dr. Khaled M Mostafa - Medical Page

https://www.facebook.com/MOSTAFA.Khaled.DR/

قناه ما خفي اعظم حرر عقلك

https://www.facebook.com/ma5afiyaa3them7areraklak/videos/?paipv=0&eav=AfbAl_Xtyd724zYn1y60bxe79mTQZyIDoDmSeHl7trhYjeED2DbYN84DeXfrZTQynsI&_rdr

Mazen Elshaal

https://www.facebook.com/mazenelshaal/

saliverse -عالم مقلوب

https://www.facebook.com/rachidsaleel/

Red Pill Arabic

https://www.facebook.com/RedPill8/

الجزايري اللغز

https://www.facebook.com/ridhatebessa1012/?paipv=0&eav=AfY6b-vBlj2hkzxayTncOBDXRRds_K07vaAyV9gAi-N0doyLNHXN9FzF4u7-RsWsjXM&_rdr

معركة وعي

https://www.facebook.com/slalibanihashim/?paipv=0&eav=AfYX6H2WcUSIceoEk3vQG4D5dA7cYb6zWO6m4yqVN9luPLSzZRAPU0Lz-aa-kaQmui8&_rdr

التفهيم

https://www.facebook.com/Tafhim.org/

Hasan Mostafa

https://www.facebook.com/toopardyy/

شفرة الحقيقة- Eslam Elbahrawy

https://www.facebook.com/TruthCode/

الحرب العالمية الثالثة

https://www.facebook.com/WorldWar3arab/?locale=ko_KR

A.2. Groupes

علوم الطاقة الكونية: ◈ Eternal Awareness الوعي السرمدي◈

https://www.facebook.com/groups/139019429778/

الأرض المسطحة

https://www.facebook.com/groups/119238771520384/

بروتوكولات حكماء صهيون

https://www.facebook.com/groups/220991834746730/*

الأرض العظيمة

https://www.facebook.com/groups/254174056814614/

قل وأنصت

https://www.facebook.com/groups/291984432006929/

● الحقيقة__THE TRUTH

https://www.facebook.com/groups/354889688856056/

أسرار وحقائق الارض المسطحة

https://www.facebook.com/groups/604614473077048/

نظرية التطور

https://www.facebook.com/groups/627621665323906/

الوعد علوم اخر الزمان

https://www.facebook.com/groups/751795644920910/?locale=fr_FR

♟ العالم في رقعة الشطرنج

https://www.facebook.com/groups/827243358727397/

السياسة الجزائرية

https://www.facebook.com/groups/852998851515960/?locale=ms_MY

⏳الوعـد

https://www.facebook.com/groups/1076261046145658/?locale=fi_FI

حقيقة الأرض المسطحة والسماوات

https://www.facebook.com/groups/1124911471360773/

إتباع أثر السلف في خلق السماوات والأرض

https://www.facebook.com/groups/2235437543277488/