UCLouvain

FACULTE DES SCIENCES

Faculté des sciences

Modélisation de thématiques appliquée à la sensibilisation aux risques de catastrophes (Topic modeling on disaster risk awareness)

Auteur : Belanov Bogning Tatchinda
Promoteur : Pr Eugen Pircalabelu
Co-promoteur : Dr Damien Delforge (IRSS, UCLouvain)
Lecteur : Pr Olivier Caelen
Année académique : 2025 - 2026
LDATS2M : Master [120] en sciences de données orientation statistique

Résumé

Ce mémoire se situe à la croisée du traitement automatique du langage (TAL) et des sciences de l'environnement. Il porte sur la modélisation de thématiques (*topic modelling*) appliquée à l'étude de l'évolution du discours médiatique sur les risques de catastrophes naturelles en Belgique francophone. S'appuyant sur le corpus RTBF (Escouflaire et al., 2023) — 767 204 articles publiés entre 2008 et 2021 par la Radio-Télévision Belge Francophone. Nous isolons, après un tri lexical fondé sur un dictionnaire d'aléas climatiques et l'algorithme d'Aho-Corasick (1975), 8 958 articles pertinents couvrant cinq catégories d'aléas : inondations, tempêtes, intempéries, sécheresses et vagues de chaleur.

Quatre modèles sont implémentés, optimisés et comparés : la *Latent Dirichlet Allocation* (LDA), le *Dynamic Topic Model* (DTM), BERTopic et le BERTopic dynamique. L'évaluation repose sur la cohérence sémantique (C_v, $C_{U_{\text{mass}}}$), la stabilité lexicale (indice de Jaccard) et la capacité à capturer des évolutions diachroniques, mesurée par la divergence de Jensen-Shannon et le taux de renouvellement lexical (T_r).

Les résultats révèlent que le discours médiatique de la RTBF sur les catastrophes naturelles présente une **forte inertie thématique** sur l'ensemble de la période étudiée. On observe néanmoins une transition progressive, amorcée autour de 2016–2018 : la couverture, longtemps centrée sur la *réaction immédiate* aux événements — secours, évacuations, dégâts — intègre désormais de manière croissante les *causes structurelles* que sont le changement climatique, la perte de biodiversité et la pression sur les ressources hydriques. Cette inflexion coïncide avec la montée en puissance des épisodes de sécheresse et de vagues de chaleur dans le corpus, et se manifeste plus nettement dans les modèles dynamiques (DTM, BERTopic dynamique) que dans leurs homologues statiques.

Sur le plan méthodologique, les modèles probabilistes (LDA, DTM) offrent une meilleure interprétabilité et un périmètre thématique mieux contrôlé, tandis que BERTopic produit une cohérence sémantique supérieure mais génère une contamination thématique plus importante sur ce corpus généraliste. Pour l'analyse de corpus journalistiques traitant de thématiques environnementales, les approches probabilistes conservent un avantage décisif lorsque la transparence des résultats et le dialogue avec des experts non informaticiens constituent des exigences premières (Tanguy et al., 2025).

Mots-clés : modélisation de thématiques, *topic modelling*, LDA, DTM, BERTopic, catastrophes naturelles, discours médiatique, RTBF, analyse diachronique, traitement automatique du langage.

Remerciements

Je tiens en premier lieu à exprimer ma profonde gratitude à mes encadrants. Au **Professeur Eugène Pircalabelu**, promoteur de ce mémoire et professeur de statistique à la LSBA, pour sa disponibilité, ses critiques constructives, son sens de la rigueur qui m'ont régulièrement conduit à remettre en question mes choix, et son accompagnement constant tout au long de la rédaction de ce travail. Au **Docteur Damien Delforge**, co-promoteur de ce mémoire et chercheur à l'Institut de la Recherche Santé et Société de l'UCLouvain, pour son aide précieuse, son regard rigoureux, son soutien dans les moments décisifs de cette recherche et son souci du détail.

Je remercie également le **Professeur Olivier Caelen**, professeur à l'UCLouvain (LSBA), d'avoir accepté d'être le lecteur de ce mémoire et de lui avoir consacré son temps et son attention.

Mes remerciements vont aussi au **Docteur Louis Escouflaire**, Postdoctoral Researcher à l'UNamur, pour sa participation à la labellisation des thématiques et pour la générosité avec laquelle il a mis son expertise en traitement automatique du langage au service de ce travail.

Je remercie ensuite chaleureusement ma famille , mes proches et toutes les personnes qui m'ont entouré et encouragé jusqu'au bout de ce travail. Votre présence, votre confiance et votre soutien indéfectible ont été une source de motivation constante.

Ma reconnaissance va également à l'ensemble du jury ainsi qu'aux membres de l'Institut de Statistique de l'Université catholique de Louvain, qui m'ont permis de poursuivre ce master et d'acquérir les compétences nécessaires à la réalisation de cette recherche.

Table des matières

Liste des Figures

Liste des Tables

0.1 Notations statistiques

Table 1: Tableau des notations statistiques

Symbole	Définition
Indices	
w	indice d'un mot
d	Indice d'un document ($d = 1, \dots, M$)
k	Indice d'une thématique ($k = 1, \dots, K$)
n	Position lexicale d'un mot dans un document ($n = 1, \dots, N_d$)
t	Indice d'une période temporelle ($t = 1, \dots, T$)
Dimensions globales	
K	Nombre total de thématiques
M	Nombre total de documents dans le corpus
V	Taille du vocabulaire (nombre de mots distincts)
N_d	Nombre de mots du document d
T	Nombre de périodes temporelles (modèles dynamiques)
Matrices et vecteurs	
X	Matrice document-terme, de dimension $M \times V$
$U, \Sigma, V^\top$	Matrices issues de la SVD de X (LSA)
W	Matrice document-thème de la NMF, de dimension $M \times K$
H	Matrice thème-terme de la NMF, de dimension $K \times V$
$\mathbf{e}_d$	Vecteur d'embedding du document d
E	Matrice des embeddings du corpus, de dimension $M \times p$
$\mathbf{C}$	Matrice de covariance entre thèmes du CTM (dimension $K \times K$)
Paramètres probabilistes	
α	Hyperparamètre de Dirichlet contrôlant la concentration des thématiques par document
η	Hyperparamètre de Dirichlet contrôlant la concentration des mots par thématique
θ_d	Distribution des thématiques du document d (vecteur de taille K)
β_k	Distribution lexicale de la thématique k (vecteur de taille V)
$z_{d,n}$	Thématique latente associée à la position n du document d
$w_{d,n}$	Mot observé à la position n du document d
$P(z \mid d)$	Probabilité que la thématique z soit active dans le document d (pLSA)
$\mathbf{u}_d$	Vecteur latent du CTM pour le document d (dimension K)
u	vecteur de moyennes de la loi normale du CTM (dimension K)

Symbole	Définition
$P(w \mid z)$	Probabilité que le mot w soit généré par la thématique z (pLSA)
Métriques	
C_v	Cohérence sémantique (mesure extrinsèque, fenêtre glissante)
$C_{U_{\text{mass}}}$	Cohérence sémantique intrinsèque (co-occurrence intra-document)
$J(A, B)$	Indice de Jaccard entre deux ensembles A et B
$\bar{J}$	Indice de Jaccard moyen sur l'ensemble des paires de graines aléatoires
$JS(P \parallel Q)$	Divergence de Jensen–Shannon entre deux distributions P et Q
$T_r(k,\, t \to t+1)$	Taux de renouvellement lexical de la thématique k entre t et $t+1$
$\mathcal{S}$	Score composite d'optimisation du modèle BERTopic
$\cos(\mathbf{x}, \mathbf{y})$	Similarité cosinus entre deux vecteurs $\mathbf{x}$ et $\mathbf{y}$
κ	Coefficient de Cohen (accord inter-annotateurs binaire)
κ_{Fleiss}	Coefficient de Fleiss (accord inter-annotateurs multi-annotateurs)
Abréviations	
LSA	Latent Semantic Analysis
SVD	Singular Value Decomposition
NMF	Non-negative Matrix Factorization
pLSA	Probabilistic Latent Semantic Analysis
LDA	Latent Dirichlet Allocation (*modèle probabiliste de thématiques*)
DTM	Dynamic Topic Model (*extension temporelle du LDA*)
BERTopic	Modèle neuronal de modélisation de thématiques fondé sur les Transformers
SBERT	Sentence-BERT (*modèle d'embeddings de phrases*)
UMAP	Uniform Manifold Approximation and Projection (*réduction de dimensionnalité*)
HDBSCAN	Hierarchical Density-Based Spatial Clustering (*algorithme de clustering*)
c-TF-IDF	Class-based TF-IDF (*pondération lexicale par thématique (BERTopic)*)
STM	Structural Topics Model
CTM	Correlated Topic model
ATM	Author Topic Model
HDP	Hierarchical Dirichlet Process
LSTM	Long Short-Term Memory
RNN	*Recurrent Neural Network*

Convention de notation. Dans l'ensemble du mémoire, les notations suivent une convention unifiée. Les **majuscules romaines** (K, M, V, T, $\mathbf{C}$) désignent les dimensions globales du corpus. Les **minuscules** (d, k, n, t) sont réservées aux indices. Les **lettres grecques** (α, η, θ_d,

β_k) représentent les paramètres probabilistes. Les **vecteurs** sont en gras ($\mathbf{e}_d$). La notation V désigne exclusivement la taille du vocabulaire ; la matrice document-terme est notée X.

1 Introduction

Les bouleversements climatiques récents nous imposent une reconsidération profonde de notre rapport aux risques de catastrophes naturelles. Depuis le début du XXI^e siècle, la multiplication et l'intensification des événements extrêmes,— inondations, vagues de chaleur, sécheresse prolongée, tempêtes tropicales — et bien d'autres; ne sont plus des phénomènes exceptionnels mais constituent une réalité croissante dans notre quotidien à laquelle toutes nos sociétés modernes font face et devront continuer à le faire. Le Cinquième rapport intergouvernemental sur l'évolution du climat (GIEC)[1] établit clairement que ces transformations résultent en grande partie de l'activité anthropique, tandis que le sixième rapport (IPCC 2021) confirmait que les événements climatiques extrêmes deviendront encore plus fréquents dans les décennies à venir. Face à cette réalité, la question de la sensibilisation aux risques de catastrophes naturels et ses déterminants devient un enjeu scientifique et sociétal majeur.

La conscience du risque climatique n'est pas un état figé. Elle se construit, évolue et se diffuse au sein de la société par l'intermédiaire des canaux médiatiques — presse écrite, réseaux sociaux, podcasts — qui jouent un rôle central dans la manière dont le public perçoit les catastrophes. Les médias de masse (RTBF, RTL, TF1, etc.) ne se contentent pas de relayer l'information factuelle : ils contribuent à construire et à structurer les représentations collectives, à hiérarchiser les priorités et à cadrer les débats autour de la gestion et de la prévention des risques. Pourtant, comme le soulignent Rouquette et Bihay (2022), certains aléas naturels demeurent « médiatiquement invisibles », faiblement couverts en raison de leur temporalité longue, de leur complexité scientifique ou de l'absence d'événements déclencheurs spectaculaires. Ce contraste entre la réalité du risque et sa représentation médiatique constitue un obstacle majeur à la sensibilisation du public et, par conséquent, à l'élaboration de politiques efficaces.

L'analyse à grande échelle de la couverture médiatique des catastrophes naturelles soulève cependant d'importants défis méthodologiques. Les approches qualitatives traditionnelles, principalement fondées sur l'analyse de contenu (Berelson (1952) ; Krippendorff (2004)), offrent une richesse interprétative considérable, mais se heurtent aux limites imposées par l'augmentation massive du volume de données textuelles produites quotidiennement. Face à ces contraintes, les méthodes computationnelles issues du traitement automatique du langage (TAL) apparaissent comme des outils particulièrement adaptés à l'exploration de grands corpus documentaires. Parmi elles, la modélisation de thématiques (*topic modeling*) occupe aujourd'hui une place centrale dans l'analyse automatisée des discours.

C'est dans cette perspective que s'inscrit le présent mémoire, à l'intersection des sciences des données, du traitement automatique du langage et des sciences de l'environnement.

Le sujet de ce travail, défini conjointement avec le Docteur Damien Delforge (IRSS, UCLouvain), repose sur l'hypothèse qu'un glissement thématique progressif s'est opéré dans le discours médiatique relatif aux risques naturels : d'une couverture principalement centrée sur la réaction immédiate aux catastrophes vers une compréhension plus large des causes structurelles, de la prévention et de l'anticipation des risques. Cette hypothèse conduit à une interrogation méthodologique et comparative centrale :

[1] Groupe d'experts intergouvernemental sur l'évolution du climat (2014)

Comment choisir et mettre en œuvre la méthode de modélisation de thématiques la plus adaptée à l'étude de la sensibilisation aux risques de catastrophes naturelles dans un corpus de grande taille, tout en conciliant l'interprétabilité statistique des modèles probabilistes et des performances récentes des approches neuronales ?

Afin de répondre à cette question, ce mémoire poursuit quatre objectifs complémentaires :

1. Comprendre les défis et spécificités de la modélisation dynamique de thématiques probabilistes : à travers l'implémentation et l'optimisation du LDA et du DTM sur un corpus journalistique réel ;
2. Proposer une synthèse des principales méthodologies de l'état de l'art, en couvrant aussi bien les approches classiques (LSA, NMF, pLSA, LDA, DTM) que les approches neuronales récentes (BERTopic, Top2Vec, Dynamic BERTopic) ;
3. Analyser l'évolution diachronique des discours médiatiques liés aux catastrophes naturelles dans les archives de la RTBF entre 2008 et 2021 ;
4. Comparer les performances et les propriétés des principaux algorithmes de topic modeling, à travers une évaluation croisée mobilisant la cohérence sémantique (C_v, $C_{U_{\text{mass}}}$), la stabilité thématique (indice de Jaccard), l'interprétabilité et la capacité à modéliser les évolutions temporelles.

Ce mémoire s'appuie sur le corpus RTBF (Escouflaire et al. (2023)), constitué de 767,204 articles journalistiques publiés entre 2008 et 2021 par la Radio-Télévision Belge Francophone. Après un filtrage lexical fondé sur un dictionnaire spécialisé des aléas climatiques et l'utilisation de l'algorithme d'Aho et Corasick (1975), 8,958 articles ont été retenus pour l'analyse finale. La démarche adoptée est progressive et comparative : elle part des approches probabilistes classiques pour évoluer vers les approches neuronales contemporaines. Dans cette perspective, ce travail apporte trois contributions principales :

- une **évaluation comparée** de plusieurs modèles statiques et dynamiques de topic modeling appliqués à un corpus journalistique francophone consacré aux catastrophes naturelles ;
- une **analyse diachronique** du discours médiatique de la RTBF entre 2008 et 2021, mettant en évidence une transition progressive d'un discours centré sur la réaction immédiate vers une prise en compte croissante des causes structurelles du changement climatique, particulièrement marquée à partir des années 2016–2018 ;
- un **cadre méthodologique reproductible**, combinant filtrage lexical, modélisation thématique, analyse de stabilité et labellisation automatique, transférable à d'autres corpus médiatiques ou à d'autres langues.

Le mémoire est structuré comme suit. Le chapitre 2 présente une revue de littérature consacrée aux fondements théoriques du traitement automatique du langage et de la modélisation de thématiques, tant classique que neuronale. Le chapitre 3 décrit la méthodologie adoptée, depuis la constitution du corpus jusqu'à l'optimisation et à l'évaluation des modèles. Les résultats expérimentaux sont ensuite présentés et discutés dans le chapitre 4, avant qu'une conclusion générale ne synthétise les principaux apports du travail ainsi que les perspectives de recherche futures.

2 Revue de littérature.

Cette section présente les principaux travaux théoriques et méthodologiques mobilisés dans ce mémoire autour de la modélisation de thématiques et, plus largement, du traitement automatique du langage (TAL). Nous revenons d'abord sur les modèles probabilistes génératifs[1] et leurs extensions temporelles, qui ont structuré le développement du *topic modeling* classique entre les années 1990 et 2015. Nous présentons ensuite les avancées introduites par les représentations distribuées et les architectures *Transformer*, à l'origine d'un renouvellement profond des méthodes modernes de TAL. Enfin, nous examinons les approches neuronales récentes de modélisation de thématiques, telles que Top2Vec et BERTopic, qui cherchent à concilier richesse sémantique et interprétabilité des thèmes. Cette revue permet de situer le positionnement méthodologique du mémoire ainsi que les principales limites auxquelles notre travail tente d'apporter une réponse.

2.1 Fondements: catégorisation automatique et recherche des motifs

2.1.1 Catégorisation automatique de textes

La catégorisation automatique de textes (Text Categorisation, TC) constitue l'une des bases historiques du traitement automatique des langues. Dans son article de référence, Sebastiani (2002) propose une synthèse approfondie des principales méthodes d'apprentissage automatique appliquées à la classification de documents textuels. Le problème y est formalisé comme une tâche de classification supervisée, généralement mono-étiquette, dans laquelle chaque document est représenté par un vecteur de caractéristiques de grande dimension. Les représentations traditionnelles reposent principalement sur le modèle sac-de-mots (*bag-of-words*), souvent pondéré à l'aide d'un schéma TF–IDF, permettant de mesurer l'importance relative des termes dans le corpus. Sebastiani compare plusieurs familles d'algorithmes, notamment Naive Bayes, k-plus proches voisins (*K-nearest neighbors*), les arbres de décision, les méthodes de type Rocchio ainsi que les machines à vecteurs de support (SVM). Les résultats montrent que les SVM obtiennent généralement les meilleures performances dans les espaces textuels de grande dimension, bien que cette efficacité se fasse souvent au détriment d'une interprétabilité limitée.

D'un point de vue statistique, ces approches reposent sur des modèles discriminants nécessitant des données annotées. Elles permettent de prédire des catégories prédéfinies, mais ne cherchent pas à découvrir la structure thématique latente d'un corpus. Cette limitation réduit leur intérêt dans les contextes exploratoires, tels que l'analyse de grands corpus médiatiques, où les thématiques émergent souvent de manière non supervisée. Néanmoins, ces travaux ont joué un rôle fondamental dans l'évolution des méthodes modernes d'analyse textuelle et ont ouvert la voie aux approches non supervisées, notamment à la modélisation de thématiques (*topic modeling*), dont l'objectif est précisément d'identifier automatiquement des structures sémantiques latentes au sein de vastes collections documentaires.

[1] Le terme « génératif » renvoie ici au processus mathématique qui préside à la création des thématiques, et non à la notion de modèle génératif au sens des grands modèles de langage récents.

2.1.2 Algorithmes de recherche de motifs et efficacité computationnelle

L'analyse statistique de grands corpus textuels nécessite également des algorithmes capables de traiter efficacement des volumes importants de données lexicales. Dans cette perspective, l'algorithme proposé par Aho et Corasick (1975) constitue une contribution fondatrice. Cet algorithme permet de rechercher simultanément un ensemble de motifs lexicaux dans un texte en un seul parcours linéaire, avec une complexité temporelle :

$$O(\ell + m + z)$$

où :

- ℓ est la longueur du texte ;
- m la somme des longueurs des motifs ;
- z le nombre d'occurrences trouvées.

L'intérêt majeur de cette approche réside dans sa capacité à effectuer des recherches multiples de manière extrêmement efficace, même sur des corpus de grande taille. Plusieurs travaux ultérieurs ont cherché à améliorer les performances et la scalabilité de cette méthode. Dans une optique de réduction de la complexité mémoire, Belazzougui (2010) proposent notamment une version succincte de l'automate Aho–Corasick, capable d'atteindre une empreinte mémoire proche de l'entropie minimale sans dégradation significative des performances. Par ailleurs, Tumeo, Villa, et Chavarria-Miranda (2012) étudient l'adaptation de l'algorithme à des environnements parallèles et distribués, afin de répondre aux contraintes computationnelles imposées par le traitement massif de données textuelles.

Dans le cadre de ce mémoire, l'algorithme d'Aho–Corasick est mobilisé pour appliquer un dictionnaire spécialisé de termes associés aux catastrophes naturelles (voir Section 3.3). Cette approche lexicale permet de réaliser un premier filtrage rapide, reproductible et computationnellement efficace du corpus journalistique avant l'étape de modélisation thématique.

2.2 Modèles probabilistes génératifs et *topic modelling* classique (1990 à 2015)

L'ensemble des modèles présentés dans cette section partagent une hypothèse fondamentale commune sur la nature des documents textuels : un document est représenté comme un **sac de mots** (*bag of words*), c'est-à-dire que l'ordre des mots est ignoré au profit de la fréquence d'apparition. Dans ce cadre, les documents sont interprétés comme des combinaisons de thèmes latents (*topics*) non directement observables, mais inférables à partir des cooccurrences lexicales présentes dans le corpus.

Sur le plan technique, ces approches s'appuient sur la construction d'une matrice document-terme (Document-Term Matrix, DTM), notée X et de dimension $M \times V$, où :

- chaque ligne représente un document ;
- chaque colonne correspond à un terme distinct du vocabulaire ;
- chaque cellule (d, w) indique le nombre d'occurrences du mot w dans le document d.

Le principe général du topic modeling consiste alors à décomposer cette matrice en composantes latentes révélant la structure thématique du corpus. Les différentes méthodes présentées ci-dessous se distinguent principalement par :

- la nature mathématique de cette décomposition ;

- les hypothèses statistiques sous-jacentes ;
- le degré d'interprétabilité des représentations produites.

L'évolution historique de ces modèles montre une transition progressive : - d'approches purement algébriques et géométriques, - vers des modèles probabilistes génératifs plus rigoureux sur le plan statistique.

2.2.1 Mini-corpus d'illustration : six articles journalistiques

! Note sur les valeurs numériques de ces illustrations

Les valeurs numériques de ces tableaux (LSA, NMF, pLSA et LDA) sont simulées pour illustrer le fonctionnement théorique des modèles. Elles ne proviennent pas d'une exécution réelle sur les six articles. Une vraie implémentation donnerait des chiffres différents, mais révélerait les mêmes thèmes (*Hydrologie, Gestion institutionnelle et Médias*). L'objectif est purement didactique : rendre l'explication plus intuitive.

Avant de présenter chaque méthode, nous introduisons un mini-corpus fictif de six articles journalistiques consacrés aux inondations, généré à des fins illustratives. Ce corpus servira de support commun à l'ensemble des démonstrations afin de faciliter la comparaison directe entre les différents modèles étudiés.

Les articles couvrent volontairement trois registres distincts :

- les faits hydrologiques (A1, A4) ;
- la gestion institutionnelle et politique (A2, A6) ;
- la couverture médiatique et le débat public (A3, A5).

Cette diversité thématique permet de mettre en évidence la manière dont chaque modèle identifie, structure et différencie les thèmes latents présents dans le corpus.

A1 — La Meuse déborde : des villages entiers sous les eaux

L'inondation massive de la Meuse a submergé plusieurs villages. Les digues ont cédé sous la pression des eaux. Les habitants ont dû être évacués en urgence. Les secours s'organisent pour pomper l'eau et réparer les digues endommagées.

Mots-clés : `inondation, Meuse, villages, digues, eaux, évacuation, secours, pomper, réparer`

A2 — Réunion au ministère : validation du plan de prévention des crues

Le ministre de l'environnement a présidé une réunion de travail pour finaliser le plan de prévention des crues. Le gouvernement annonce un budget exceptionnel pour renforcer les infrastructures hydrauliques et les systèmes d'alerte précoce.

Mots-clés : `gouvernement, prévention, crues, ministre, budget, infrastructures, alerte, politique`

A3 — Les médias couvrent en direct les inondations dans le sud

Les chaînes de télévision et les journaux en ligne diffusent en continu des images des inondations. Les journalistes sont déployés sur le terrain. La couverture médiatique intense sensibilise l'opinion publique au risque d'inondation.

Mots-clés : `médias, télévision, journaux, journalistes, couverture, images, opinion,`

`inondations`

A4 — Bilan des inondations : infrastructures dévastées, routes coupées

Les inondations ont causé d'importants dégâts aux infrastructures. Des routes sont coupées, des ponts endommagés, des réseaux électriques hors service. Les autorités évaluent les coûts de reconstruction et demandent une aide de l'État.

Mots-clés : `inondations, infrastructures, routes, ponts, dégâts, reconstruction, État, coûts`

A5 — Journalistes et élus : un dialogue de sourds sur les crues ?

Un journaliste d'investigation publie un dossier accusant les élus de minimiser les risques d'inondation. Les responsables politiques répondent dans la presse. Le débat public s'intensifie sur la responsabilité gouvernementale face aux catastrophes.

Mots-clés : `journaliste, élus, crues, presse, débat, politique, responsabilité, risques`

A6 — Alerte aux crues : le système d'alerte précoce a-t-il fonctionné ?

Les experts hydrologiques analysent le fonctionnement du système d'alerte aux crues. Les données météorologiques n'ont pas été correctement transmises aux autorités locales. Une enquête parlementaire est ouverte pour évaluer les défaillances du dispositif de prévention.

Mots-clés : `alerte, crues, système, experts, données, prévention, enquête, autorités, défaillances`

Avant l'application de tout modèle de topic modeling, chaque document est transformé en un vecteur de fréquences lexicales. L'ensemble de ces vecteurs constitue la matrice document-terme (Document-Term Matrix, DTM), notée :

$$X \in \mathbb{R}^{M \times V}$$

où :

- M représente le nombre de documents ;
- V correspond à la taille du vocabulaire lexical (nombre de mots distincts conservés dans le corpus après le prétraitement).

Dans notre exemple illustratif, la matrice est de dimension 6×10.

Cette matrice constitue l'unique représentation textuelle exploitée par les modèles présentés dans cette section. Les algorithmes n'analysent ni la syntaxe ni la structure grammaticale des phrases ; ils manipulent uniquement des fréquences de mots. Toute la difficulté du *topic modeling* consiste alors à identifier les structures thématiques latentes susceptibles d'expliquer les cooccurrences observées dans cette représentation matricielle.

La table suivante présente la DTM construite à partir des six articles fictifs introduits précédemment.

2.2.2 Description : Analyse Sémantique Latente (LSA/LSI)

L'Analyse Sémantique Latente (*Latent Semantic Analysis*, LSA), introduite par Deerwester et al. (1990), constitue l'une des premières méthodes formalisées d'extraction automatique de structures sémantiques

Table 2.1: Matrice document-terme (DTM) — fréquences de 10 mots-clés dans les 6 articles

[H]

	inond.	digue	crue	médias	journ.	gouv.	alerte	infra.	prév.	évac.
A1	**4**	2	·	1	·	·	·	1	1	2
A2	1	·	**3**	·	·	**3**	2	2	**3**	·
A3	2	·	·	**4**	**3**	·	·	·	·	·
A4	**3**	1	1	·	·	1	·	**4**	1	·
A5	1	·	2	2	**3**	2	·	·	1	·
A6	·	·	**3**	·	·	2	**4**	·	**3**	·

latentes à partir de corpus textuels. Cette approche repose sur la **décomposition en valeurs singulières** (*Singular Value Decomposition*, SVD) de la matrice document-terme X :

$$X = U\Sigma V_{\text{svd}}^{\top} \tag{2.1}$$

où :

- U est une matrice orthogonale de dimension $M \times M$ contenant les vecteurs singuliers gauches, associés aux représentations latentes des documents ;
- V_{svd} est une matrice orthogonale de dimension $V \times V$ contenant les vecteurs singuliers droits, associés aux représentations latentes des termes ;
- Σ est une matrice diagonale de dimension $M \times V$ contenant les valeurs singulières σ_i, ordonnées par importance décroissante.

Afin d'extraire les K dimensions latentes les plus significatives, la LSA réalise une **SVD tronquée** en conservant uniquement les K plus grandes valeurs singulières de Σ, ainsi que les colonnes correspondantes de U et V_{svd}. On obtient alors une approximation de rang réduit de la matrice initiale :

$$X \approx X_K = U_K \Sigma_K V_K^{\top} \tag{2.2}$$

D'après le théorème d'Eckart–Young[2] (Eckart et Young (1936)), cette approximation constitue la meilleure approximation de rang K de la matrice X au sens de la norme de Frobenius.

La LSA introduit ainsi l'idée selon laquelle les relations sémantiques entre documents et termes peuvent être représentées dans un espace latent de dimension réduite. Deux mots peuvent alors être considérés comme sémantiquement proches même s'ils n'apparaissent jamais explicitement ensemble, dès lors qu'ils surviennent dans des contextes similaires.

Malgré son caractère pionnier, la LSA présente plusieurs limites importantes. La SVD demeure une méthode purement algébrique ne reposant sur aucun modèle probabiliste explicite. Les dimensions latentes extraites sont donc souvent difficiles à interpréter directement comme des thèmes sémantiques clairement identifiables. Par ailleurs, les matrices produites peuvent contenir des valeurs négatives, ce qui complique leur interprétation dans le cadre du *topic modeling*, où les fréquences et proportions thématiques sont généralement supposées positives. Enfin, la LSA ne fournit pas de véritable mécanisme génératif permettant d'estimer naturellement la probabilité d'apparition de nouveaux documents en dehors du corpus d'apprentissage.

[2] le théorème d'Eckart–Young garantit que cette méthode de compression conserve le maximum d'informations du texte d'origine. Si l'on réduit le corpus à ses K thèmes majeurs, cette formule mathématique offre le « résumé » le plus fidèle possible, en réduisant la perte de sens au strict minimum.

2.2.2.1 Illustration : Analyse Sémantique Latente (LSA)

Intuition

La LSA peut être vue comme une méthode de compression de l'information textuelle. À partir de la matrice document-terme X, elle cherche à extraire les principales directions de variation du corpus tout en réduisant le bruit et les redondances lexicales.

Analogie. Imaginons les six articles projetés dans un espace à dix dimensions, chaque dimension correspondant à un mot du vocabulaire. La LSA cherche alors les deux ou trois axes latents qui résument le mieux la structure globale du corpus, puis projette les documents sur ces nouveaux axes.

Correspondance mathématique. La SVD tronquée $X \approx U_K \Sigma_K V_K^\top$ décompose la matrice initiale en :

- une matrice U_K de dimension $M \times K$ décrivant les documents dans l'espace latent ;
- une matrice $V_K^\top$ de dimension $K \times V$ décrivant les termes ;
- une matrice diagonale Σ_K contenant les valeurs singulières qui pondèrent l'importance de chaque dimension latente.

1. **projection des documents sur 2 dimensions latentes**

En fixant $K = 2$, la SVD tronquée génère deux matrices principales qui projettent le corpus dans un espace géométrique réduit :

- U_K, qui contient les coordonnées des documents dans l'espace latent ;
- $V_K^\top$, qui contient les coordonnées des termes.

Ces nouvelles coordonnées sont pondérées par les deux plus grandes valeurs singulières $\sigma_1 > \sigma_2$, qui quantifient l'importance et l'apport d'information de chaque dimension latente.

- Matrice U_K — Représentation des documents dans l'espace latent.

Doc	Concept 1	Concept 2	Interprétation
A1	**+0.61**	+0.22	Fortement associé à la dimension hydrologique
A2	+0.18	**+0.67**	Fortement associé à la gestion institutionnelle
A3	+0.15	**−0.41**	Position opposée au Concept 2 : registre principalement médiatique
A4	**+0.55**	+0.28	Très proche de A1 (inondations et infrastructures)
A5	+0.28	−0.30	Position intermédiaire entre médias et gestion politique
A6	+0.22	**+0.64**	Très proche de A2 (gestion et prévention)

- Matrice $V_K^\top$ — Contribution des termes aux concepts latents.

Concept	inond.	digue	crue	médias	journ.	gouv.	alerte	infra.	prév.	évac.
C1	**.62**	**.41**	.18	.11	.09	.12	.10	.45	.14	.38
C2	.10	−.08	**.55**	.05	.08	**.58**	**.64**	.10	**.58**	−.05

▶ **Limite des valeurs négatives dans la LSA**
On observe par exemple la valeur −0.41 pour A3 sur le Concept 2, ainsi que la valeur −0.08 pour le terme « digue » dans ce même concept. Ces coefficients négatifs signifient mathématiquement que certains documents ou termes se situent dans une direction opposée à celle du concept latent considéré.
Cependant, dans une interprétation thématique, une proportion ou une intensité thématique négative ne possède pas de signification intuitive. Cette limite provient directement de la SVD, qui opère sur l'espace vectoriel réel $\mathbb{R}$ sans imposer de contrainte de positivité sur les coefficients.
C'est précisément cette difficulté interprétative qui a motivé le développement de méthodes alternatives comme la NMF, fondées sur des représentations strictement non négatives.

2.2.3 Description : Factorisation de Matrices Non-Négatives (NMF)

La Factorisation de Matrices Non-Négatives (*Non-Negative Matrix Factorization*, NMF), proposée par Lee et Seung (1999), constitue une alternative à la LSA visant à améliorer l'interprétabilité des représentations latentes. Contrairement à la SVD utilisée en LSA, la NMF impose une contrainte stricte de non-négativité sur les matrices produites. À partir de la matrice document-terme X de dimension $M \times V$, la décomposition s'écrit :

$$X \approx W \cdot H, \qquad \text{avec } W \geq 0 \text{ et } H \geq 0 \tag{2.3}$$

où :

- W est une matrice de dimension $M \times K$ représentant l'intensité d'association des documents aux K thèmes latents ;
- H est une matrice de dimension $K \times V$ représentant la contribution des termes à chacun des thèmes.

La contrainte de non-négativité introduit une propriété importante d'**additivité des composantes**. Contrairement à la SVD, qui autorise des coefficients négatifs et des combinaisons soustractives difficiles à interpréter, la NMF reconstruit les documents comme des combinaisons additives de composantes sémantiques positives. Cette propriété améliore fortement l'interprétabilité des thèmes extraits.

La NMF produit ainsi des représentations généralement plus parcimonieuses (*sparse*) et plus facilement interprétables dans le contexte du *topic modeling*.

2.2.3.1 Fonctions de perte et optimisation

L'estimation des matrices W et H repose sur la minimisation d'une fonction de coût mesurant l'écart entre la matrice originale X et sa reconstruction WH. Deux fonctions de perte sont principalement utilisées dans la littérature.

1. **Erreur quadratique**

Généralement appliquée lorsque le bruit des données est supposé gaussien, cette fonction s'appuie sur la **norme de Frobenius au carré [1, 2]** pour mesurer la distance entre les matrices [1, 2] :

$$\mathcal{L}_{\text{Frobenius}} = \|X - WH\|_F^2 = \sum_{d=1}^{M} \sum_{v=1}^{V} \left(X_{dv} - (WH)_{dv}\right)^2$$

2. **Divergence généralisée de Kullback–Leibler**

Cette divergence est souvent mieux adaptée aux données textuelles de comptage de mots, fréquemment assimilées à des variables de type Poisson :

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}(X \parallel WH) = \sum_{d=1}^{M} \sum_{v=1}^{V} \left(X_{dv} \log \frac{X_{dv}}{(WH)_{dv}} - X_{dv} + (WH)_{dv} \right)$$

L'optimisation de ces fonctions, non convexes lorsqu'elles sont considérées conjointement par rapport à W et H, est réalisée de manière alternée à l'aide des **règles de mise à jour multiplicatives** proposées par Lee et Seung (2001). À chaque itération, l'une des matrices est maintenue fixe pendant que l'autre est mise à jour, ce qui garantit : - une décroissance monotone de la fonction de perte ; - le maintien de la non-négativité des coefficients.

Plusieurs travaux, notamment Greene, O'Callaghan, et Cunningham (2014), montrent que la NMF peut produire des thèmes à la fois cohérents et relativement stables sur des corpus textuels de taille moyenne.

Malgré ses qualités interprétatives, la NMF conserve toutefois certaines limites héritées des approches géométriques. Elle ne repose pas sur un modèle génératif probabiliste explicite, ce qui limite le recours à l'inférence bayésienne formelle ainsi que la modélisation probabiliste complète du processus de génération des documents.

2.2.3.2 Illustration : Factorisation de Matrices Non-Négatives (NMF)

Intuition

La NMF peut être interprétée comme une méthode d'assemblage additif de composantes thématiques positives. Chaque document est reconstruit à partir d'une combinaison de thèmes, sans jamais recourir à des coefficients négatifs.

Analogie. Chaque article peut être vu comme une « recette » combinant plusieurs composantes thématiques (Hydrologie, Gestion, Médias) dans des proportions positives. Par exemple, A5 correspond à un mélange réparti entre Hydrologie (18 %), Gestion (38 %) et Médias (44 %), tandis que A1 est principalement dominé par le thème Hydrologie (72 %).

Correspondance mathématique. La contrainte $X \approx W \cdot H,\ W \geq 0,\ H \geq 0$ garantit une décomposition strictement additive. La matrice W (de dimension $M \times K$) décrit la composition thématique des documents, tandis que la matrice H (de dimension $K \times V$) décrit l'importance relative des termes au sein de chaque thème. L'optimisation cherche à minimiser une fonction de coût, souvent fondée sur la divergence de Kullback–Leibler $D_{\text{KL}}(X \,\|\, WH)$, particulièrement adaptée aux données textuelles de comptage.

Résultats : décomposition en 3 thèmes non-négatifs ($K = 3$) :

- Matrice H — Profil thématique des termes.

Thème	inond.	digue	crue	médias	journ.	gouv.	alerte	infra.	prév.	évac.
T1 Hydro.	**.55**	**.38**	.12	.05	.04	.06	.08	**.42**	.10	**.35**
T2 Gestion	.08	.04	**.48**	.06	.05	**.52**	**.58**	.12	**.55**	.03
T3 Médias	.14	.02	.10	**.62**	**.58**	.10	.04	.05	.08	.04

- Matrice W — Appartenance thématique des documents.

Doc	T1 Hydro.	T2 Gestion	T3 Médias	Thème dominant
A1	**0.72**	0.18	0.10	*Hydrologie & Urgence*
A2	0.08	**0.75**	0.17	*Gestion & Politique*
A3	0.12	0.05	**0.83**	*Médias & Opinion*
A4	**0.58**	0.30	0.12	*Hydrologie & Urgence*
A5	0.18	0.38	0.44	*Mixte Médias/Gestion*
A6	0.10	**0.72**	0.18	*Gestion & Politique*

✓ **Interprétation de la matrice W.**

Chaque valeur de la matrice W correspond à une proportion comprise entre 0 et 1 indiquant la part relative de chaque thème dans un document. La somme des coefficients d'une ligne est égale à 1, ce qui signifie que l'ensemble des thèmes reconstitue intégralement le contenu du document.

Exemple pour A5 (*Journalistes et élus*) :

$$\underbrace{0.18}_{\text{Hydrologie}} + \underbrace{0.38}_{\text{Gestion}} + \underbrace{0.44}_{\text{Médias}} = 1.00$$

Contrairement à la matrice W, les lignes de la matrice H ne représentent pas des probabilités et ne somment donc pas à 1. Les coefficients de H doivent être interprétés comme des poids relatifs indiquant l'importance des mots dans la définition d'un thème.
Par exemple, la valeur élevée associée au terme « inondation » dans le thème T1 indique que ce mot joue un rôle particulièrement discriminant dans la caractérisation de ce thème.
L'absence totale de coefficients négatifs facilite fortement l'interprétation des résultats, propriété directement garantie par les contraintes $W \geq 0$ et $H \geq 0$.

2.2.4 Description : Analyse Sémantique Latente Probabiliste (pLSA / pLSI)

Afin de dépasser les limites de la LSA, Hofmann (1999a) introduit l'**Analyse Sémantique Latente Probabiliste** (*Probabilistic Latent Semantic Analysis*, pLSA). Contrairement aux approches purement algébriques, la pLSA repose sur une formulation probabiliste explicite fondée sur un **modèle génératif latent** (*aspect model*).

Le principe central du modèle est que chaque mot w observé dans un document d est généré à partir d'une variable latente $z \in \{1, \ldots, K\}$ représentant un thème. Dans ce cadre :

- le choix du thème dépend du document ;
- le choix du mot dépend uniquement du thème sélectionné.

Cette hypothèse conduit à la factorisation probabiliste suivante :

$$P(d, w) = P(d)\, P(w \mid d) = P(d) \sum_{z=1}^{K} P(z \mid d)\, P(w \mid z) \tag{2.4}$$

où :

- $P(z \mid d)$ représente la distribution des thèmes dans le document d ;
- $P(w \mid z)$ représente la distribution des mots dans le thème z.

La pLSA introduit ainsi une interprétation probabiliste directe des thèmes latents. Chaque document est modélisé comme un mélange de thèmes, tandis que chaque thème correspond à une distribution probabiliste sur les mots du vocabulaire.

2.2.4.1 Estimation par l'algorithme EM et limites du modèle

Les paramètres du modèle, à savoir $P(z \mid d)$ et $P(w \mid z)$, sont estimés par maximisation de la log-vraisemblance du corpus à l'aide de l'algorithme **Expectation-Maximization (EM)**. La fonction objectif s'écrit :

$$\mathcal{L}_{\text{pLSA}} = \sum_{d=1}^{M} \sum_{w=1}^{V} n(d, w) \log \left[\sum_{z=1}^{K} P(z \mid d)\, P(w \mid z) \right]$$

où $n(d, w)$ désigne le nombre d'occurrences du mot w dans le document d.

La pLSA constitue une avancée majeure par rapport à la LSA, notamment grâce à son cadre probabiliste explicite et à l'amélioration de l'interprétabilité des thèmes extraits. Néanmoins, le modèle présente deux limites théoriques importantes, qui motiveront directement le développement du LDA.

- **Absence de véritable modèle génératif des documents.**
  Les distributions thématiques $P(z \mid d)$ sont directement associées aux documents observés dans le corpus d'apprentissage. Elles ne sont pas elles-mêmes générées par une distribution probabiliste a priori. Par conséquent, la pLSA ne permet pas de généraliser naturellement à de nouveaux documents non observés pendant l'entraînement.
- **Risque élevé de surapprentissage (*overfitting*).**
  Le nombre de paramètres du modèle croît linéairement avec le nombre de documents du corpus, soit environ $K \times M + K \times V$ paramètres à estimer. Cette dépendance directe à la taille du corpus augmente fortement le risque de surapprentissage, en particulier sur les corpus de grande dimension.

2.2.4.2 Illustration : Analyse Sémantique Latente Probabiliste (pLSA)

Intuition

La pLSA peut être interprétée comme un mécanisme probabiliste d'attribution thématique. Chaque mot observé dans un document est supposé avoir été généré par un thème latent, sélectionné selon une distribution de probabilités propre au document.

Analogie. Pour chaque mot d'un article, le modèle cherche à déterminer s'il provient principalement du vocabulaire associé aux crues hydrologiques, à la gestion institutionnelle ou à la couverture médiatique. Cette attribution probabiliste est réalisée simultanément pour l'ensemble des mots du corpus à l'aide de l'algorithme EM.

Correspondance mathématique. La formule de marginalisation sur les thèmes latents z,

$$P(d, w) = P(d) \sum_{z=1}^{K} P(z \mid d)\, P(w \mid z),$$

décompose la probabilité d'observer un mot w dans un document d comme une combinaison pondérée des différents thèmes possibles. L'algorithme EM alterne alors entre :

- une étape E (*Expectation*), qui estime la probabilité qu'un mot appartienne à chaque thème ;
- une étape M (*Maximization*), qui met à jour les distributions $P(z \mid d)$ et $P(w \mid z)$.

Résultats : estimation des distributions thématiques par l'algorithme EM

$P(z \mid d)$ — **Distribution des thèmes dans chaque document.**

Doc	$P(z_1 \mid d)$ Hydro.	$P(z_2 \mid d)$ Gestion	$P(z_3 \mid d)$ Médias	Interprétation
A1	**70 %**	20 %	10 %	*Dominance des crues physiques*
A2	10 %	**75 %**	15 %	*Gestion institutionnelle*
A3	15 %	8 %	**77 %**	*Couverture médiatique*
A4	**60 %**	28 %	12 %	*Dégâts et infrastructures*
A5	20 %	40 %	40 %	*Profil mixte gestion/médias*
A6	12 %	**70 %**	18 %	*Prévention et systèmes d'alerte*

Lecture du tableau

Chaque ligne correspond à un document. Les trois colonnes indiquent la probabilité d'appartenance de l'article à chacun des thèmes latents. La somme des probabilités sur une ligne est toujours égale à 100 %.

Par exemple, A1 présente une probabilité de 70 % pour le thème Hydrologie. Cela signifie que la majorité des mots observés dans ce document est expliquée par le vocabulaire associé aux crues et aux phénomènes hydrologiques.

Différence importante avec la NMF : les valeurs présentées ici sont de véritables probabilités issues d'un modèle génératif probabiliste, et non de simples poids géométriques.

Calcul de $P(d, w)$: décomposer la vraisemblance pas à pas

En appliquant la formule de marginalisation sur $P(A_1,$ "inondation"$)$:

$$P(A_1,\ \text{inond.}) = P(A_1)\Big[\underbrace{P(z_1|A_1)}_{0.70} \cdot \underbrace{P(\text{inond.}|z_1)}_{0.28} + \underbrace{P(z_2|A_1)}_{0.20} \cdot \underbrace{P(\text{inond.}|z_2)}_{0.04} + \underbrace{P(z_3|A_1)}_{0.10} \cdot \underbrace{P(\text{inond.}|z_3)}_{0.07}\Big]$$

$$= P(A_1)\,[0.196 + 0.008 + 0.007] = P(A_1) \times 0.211$$

Le mot « inondation » dans A1 s'explique à $\frac{0.196}{0.211} \approx \mathbf{93}\,\%$ par le thème Hydrologie z_1, ce qui est cohérent avec le contenu fortement centré sur les crues physiques de ce document.

▶ **Limite principale de la pLSA**
Dans la pLSA, les distributions $P(z \mid d)$ sont des paramètres directement associés aux documents observés dans le corpus d'apprentissage. Elles ne sont pas générées par une distribution probabiliste a priori.

Par conséquent, si un nouveau document A_7 est ajouté au corpus, il devient impossible d'estimer directement ses proportions thématiques sans relancer l'algorithme EM sur l'ensemble des données.

De plus, le nombre de paramètres du modèle augmente linéairement avec le nombre de documents :

$$6 \times 3 = 18$$

paramètres pour les distributions $P(z \mid d)$, auxquels s'ajoutent :

$$3 \times 10 = 30$$

paramètres pour les distributions $P(w \mid z)$.

Sur des corpus réels contenant plusieurs dizaines ou centaines de milliers de documents, cette croissance du nombre de paramètres augmente fortement le risque de surapprentissage (*overfitting*). Cette double limite constitue l'une des principales motivations ayant conduit au développement du modèle LDA.

2.2.5 Description : Allocation de Dirichlet Latente (LDA)

L'Allocation de Dirichlet Latente (*Latent Dirichlet Allocation*, LDA), introduite par Blei, Ng, et Jordan (2003), est un modèle probabiliste génératif qui formalise la création d'un document comme un processus de tirage aléatoire. L'idée centrale du LDA est qu'un auteur rédige un texte en deux étapes : il choisit d'abord un mélange de thèmes pour son document (ex. : 70 % d'Hydrologie et 30 % de Médias), puis il pioche des mots spécifiques associés à ces thèmes. Pour modéliser cette idée, le LDA attribue une structure probabiliste rigoureuse à chaque niveau du corpus. Alors que la pLSA se contentait d'estimer des proportions fixes pour chaque document existant, le LDA introduit des lois de probabilité a priori (des distributions de Dirichlet) qui encadrent la création des thèmes et des mots.

Cette approche permet de dépasser les limites structurelles de la pLSA de trois manières :

- Limite le risque de surapprentissage : le nombre de paramètres à estimer ne dépend plus du nombre de documents du corpus, ce qui supprime le risque d'ajustement excessif (*overfitting*) propre à la pLSA ;
- Mieux généraliser à de nouveaux documents : en définissant une loi mathématique globale pour le corpus, le LDA est capable de calculer la probabilité de documents totalement nouveaux qui n'existaient pas lors de la phase d'entraînement ;
- Régularisation statistique naturelle : La loi de Dirichlet agit comme un stabilisateur. Elle force le modèle à retenir les grandes tendances globales plutôt que de se perdre dans les détails ou les cas particuliers de chaque document.

2.2.5.1 Processus génératif du modèle

Le LDA suppose que chaque document du corpus est généré selon le processus probabiliste suivant :

- Pour chaque thème $k \in \{1, \dots, K\}$: on tire une distribution lexicale $\beta_k \sim \text{Dirichlet}(\eta)$
- Pour chaque document $d \in \{1, \dots, M\}$:

a. on tire une distribution thématique :

$$\theta_d \sim \text{Dirichlet}(\alpha)$$

b. pour chaque mot $n \in \{1, \dots, N_d\}$ du document :

- on sélectionne un thème latent : $z_{d,n} \sim \text{Multinomial}(\theta_d)$
- on génère ensuite le mot observé : $w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}})$

Le document est ainsi interprété comme un mélange probabiliste de thèmes latents, chaque thème étant lui-même défini par une distribution probabiliste sur les mots du vocabulaire.

Signification des variables du modèle

Afin de clarifier la structure probabiliste du LDA, les variables peuvent être regroupées en trois catégories.

1. Variables observées

- $w_{d,n}$: mot observé à la position n dans le document d.
  Il s'agit de la seule information textuelle effectivement observée par le modèle.

2. Variables latentes

- θ_d : vecteur de dimension K représentant la distribution des thèmes dans le document d ;
- $z_{d,n}$: thème latent associé au mot situé à la position n du document d ;
- β_k : distribution des mots associée au thème k.

3. Hyperparamètres

- α : paramètre de Dirichlet contrôlant la concentration des thèmes dans les documents. Un faible α favorise des documents dominés par un petit nombre de thèmes ;
- η (parfois noté β dans la littérature) : paramètre de Dirichlet contrôlant la concentration des mots dans les thèmes. Une faible valeur de η favorise des thèmes composés d'un nombre réduit de mots fortement discriminants.

2.2.5.2 Distribution conjointe complète

La distribution conjointe du modèle pour l'ensemble du corpus s'écrit :

$$p(\mathbf{W}, \mathbf{Z}, \Theta, \mathbf{B} \mid \alpha, \eta) = \prod_{k=1}^{K} p(\beta_k \mid \eta) \prod_{d=1}^{M} p(\theta_d \mid \alpha) \prod_{n=1}^{N_d} p(z_{d,n} \mid \theta_d)\, p(w_{d,n} \mid z_{d,n}, \beta_k) \tag{2.5}$$

où :

- $\mathbf{W} = \{w_{d,n}\}$ représente l'ensemble des mots observés dans le corpus ;
- $\mathbf{Z} = \{z_{d,n}\}$ représente l'ensemble des affectations thématiques latentes associées à chaque mot ;
- $\Theta = \{\theta_d\}_{d=1}^{M}$ représente l'ensemble des distributions thématiques des documents ;

- $\mathbf{B} = \{\beta_k\}_{k=1}^{K}$ représente l'ensemble des distributions lexicales des thèmes.

Les hyperparamètres α et η gouvernent respectivement la concentration des distributions thématiques Θ et des distributions lexicales $\mathbf{B}$ via les lois de Dirichlet associées.

2.2.5.3 Illustration : Allocation de Dirichlet Latente (LDA)

Intuition

Le LDA peut être interprété comme un modèle génératif hiérarchique décrivant la manière dont un document pourrait être produit à partir de thèmes latents. Chaque article est supposé être généré en tirant d'abord une distribution thématique, puis en sélectionnant les mots un à un selon les thèmes activés.

Analogie. Pour générer le document A1, le modèle commence par tirer une composition thématique de type : 68 % Hydrologie, 22 % Gestion et 10 % Médias. Ensuite, pour chaque position du document, il sélectionne un thème latent selon cette distribution, puis choisit un mot selon la distribution lexicale β_k associée au thème actif.

Correspondance mathématique. Le processus génératif hiérarchique $\beta_k \sim \text{Dirichlet}(\eta)$, $\theta_d \sim \text{Dirichlet}(\alpha)$, $z_{d,n} \sim \text{Multinomial}(\theta_d)$, $w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}})$ constitue le cœur du modèle. En traitant θ_d comme une variable aléatoire issue d'une distribution de Dirichlet plutôt que comme un paramètre fixe, le LDA peut généraliser son apprentissage à de nouveaux documents sans réestimer entièrement le modèle.

Résultats de l'inférence ($K = 3$, $\alpha = 0.5$, $\eta = 0.1$)

- Distributions β_k — Profils lexicaux des thèmes.

Thème β_k	inond.	digue	crue	médias	journ.	gouv.	alerte	infra.	prév.	évac.
T1 Hydro.	**.26**	**.17**	.05	.02	.02	.02	.04	**.20**	.06	**.16**
T2 Gestion	.04	.02	**.18**	.03	.02	**.20**	**.22**	.07	**.21**	.01
T3 Médias	.10	.02	.06	**.32**	**.30**	.06	.03	.03	.05	.03

- Distributions θ_d — Proportions thématiques des documents.

Doc	$\theta_{d,1}$ Hydro.	$\theta_{d,2}$ Gestion	$\theta_{d,3}$ Médias	Interprétation
A1	**68 %**	22 %	10 %	$\theta_d \sim \text{Dir}(\alpha)$: les proportions thématiques sont désormais modélisées comme des variables aléatoires issues d'une distribution a priori, ce qui permet au modèle de généraliser à de nouveaux documents.
A2	9 %	**74 %**	17 %	
A3	13 %	7 %	**80 %**	
A4	**57 %**	31 %	12 %	
A5	19 %	38 %	43 %	
A6	11 %	**71 %**	18 %	

Affectation mot-par-mot $z_{d,n}$ pour A1 (échantillonnage de Gibbs)

Rôle du processus d'échantillonnage de Gibbs

L'échantillonnage de Gibbs constitue l'un des principaux algorithmes d'inférence utilisés pour estimer les variables latentes du LDA. À partir des mots observés dans les documents, il attribue progressivement à chaque mot le thème latent le plus probable compte tenu des affectations thématiques des autres mots du corpus.

Analogie. On peut imaginer un système de tri dans lequel chaque mot est déplacé d'un thème à un autre jusqu'à ce qu'une organisation thématique stable émerge progressivement du corpus.

La colonne **Confiance** indique le degré de certitude associé à l'affectation thématique proposée. Par exemple, une confiance de 92 % pour le mot « inondation » dans le thème Hydrologie signifie que ce mot est fortement caractéristique de ce thème.

L'échantillonnage de Gibbs attribue ainsi, pour chaque mot, une variable latente $z_{d,n}$ selon une distribution conditionnelle dépendant des thèmes attribués à tous les autres mots du corpus.

Mot $w_{d,n}$	Thème $z_{d,n}$	Justification	Confiance
inondation	**T1 Hydrologie**	Fort signal lexical associé aux crues	**92 %**
Meuse	**T1 Hydrologie**	Forte proximité avec le vocabulaire hydrologique	**87 %**
villages	**T1 Hydrologie**	Association modérée au thème Hydrologie	71 %
digues	**T1 Hydrologie**	Mot fortement discriminant du thème T1	**84 %**
évacuation	**T1 Hydrologie**	Vocabulaire lié à la gestion de crise hydrologique	78 %
secours	**T1 Hydrologie**	Association probable au contexte de catastrophe	65 %
pomper	**T1 Hydrologie**	Terme cohérent avec les opérations post-crue	70 %
réparer	**T2 Gestion**	Terme davantage associé aux infrastructures et à la gestion publique	58 %

2.2.6 Synthèse comparative

2.2.7 Extensions du LDA

2.2.7.1 Modèle de Topics Corrélés (CTM)

Le Modèle de Topics Corrélés (*Correlated Topic Model*, CTM), proposé par Blei et Lafferty (2007), constitue une extension du LDA destinée à modéliser explicitement les dépendances entre thèmes.

Dans le LDA classique, la distribution de Dirichlet impose implicitement une corrélation négative entre les topics : lorsqu'un document présente une forte proportion d'un thème donné, les proportions des autres thèmes diminuent mécaniquement. Cette hypothèse peut toutefois apparaître restrictive dans de nombreux corpus réels, où certains thèmes tendent naturellement à coexister.

Table 2.2: Comparaison des principaux modèles appliqués au mini-corpus

[H]

Modèle	Valeurs négatives	Modèle génératif	Nouveaux documents	Interprétation sur le mini-corpus
LSA	Oui ×	Non ×	Non ×	Dimensions algébriques contenant des coefficients négatifs, difficiles à interpréter directement comme des thèmes.
NMF	Non ✓	Non ×	Non ×	Trois thèmes clairement séparés (Hydrologie, Gestion, Médias) avec des coefficients exclusivement positifs.
pLSA	Non ✓	Partiel ∼	Non ×	Distributions probabilistes $P(z \mid d)$ interprétables, mais directement dépendantes du corpus d'apprentissage.
LDA	Non ✓	Oui ✓	Oui ✓	Distributions thématiques généralisables via $\theta_d \sim \mathrm{Dir}(\alpha)$ et affectation probabiliste mot par mot.

Afin de lever cette contrainte, le CTM remplace la distribution de Dirichlet par une **distribution logistique normale** :

$$\mathbf{u}_d \sim \mathcal{N}(\mu, \mathbf{C}), \qquad \theta_d = \frac{\exp(\mathbf{u}_d)}{\sum_{k=1}^{K} \exp(u_{dk})} \tag{2.6}$$

où :

- $\mathbf{u}_d \in \mathbb{R}^K$ désigne un vecteur latent associé au document d ;
- μ représente le vecteur des moyennes ;
- $\mathbf{C} \in \mathbb{R}^{K \times K}$ correspond à la matrice de covariance entre les thèmes.

La matrice $\mathbf{C}$ permet de capturer les corrélations thématiques observées dans le corpus. Une covariance positive entre deux topics indique qu'ils ont tendance à apparaître conjointement dans les mêmes documents.

Cette propriété est particulièrement pertinente dans l'analyse de corpus journalistiques consacrés aux catastrophes naturelles, où certains thèmes — comme les inondations, les tempêtes, les évacuations ou la gestion de crise — apparaissent fréquemment de manière conjointe.

2.2.7.2 Processus de Dirichlet Hiérarchique (HDP)

Le Processus de Dirichlet Hiérarchique (*Hierarchical Dirichlet Process*, HDP), introduit par Teh et al. (2006), constitue une extension non paramétrique du LDA.

Sa principale originalité réside dans le fait qu'il n'est plus nécessaire de fixer a priori le nombre de thèmes K. Le modèle est capable d'inférer automatiquement le nombre pertinent de topics à partir des données observées.

L'HDP repose sur une hiérarchie de processus de Dirichlet :

- un processus global partagé par l'ensemble du corpus ;
- des processus locaux spécifiques à chaque document.

Cette structure permet au modèle de considérer théoriquement un nombre infini de thèmes potentiels, même si seul un sous-ensemble limité est effectivement activé dans les données.

L'approche présente un intérêt théorique important, notamment dans les contextes exploratoires où le nombre de thèmes sous-jacents est inconnu. Toutefois, cette flexibilité s'accompagne d'un coût computationnel plus élevé ainsi que d'une interprétation parfois plus complexe des thèmes extraits.

2.2.7.3 Modèle Auteur-Topic (ATM)

Le Modèle Auteur-Topic (*Author Topic Model*, ATM), proposé par Rosen-Zvi et al. (2004), étend le LDA en intégrant explicitement l'identité des auteurs dans le processus génératif.

Dans ce modèle, chaque auteur possède sa propre distribution thématique. Lors de la génération d'un mot, le modèle sélectionne d'abord un auteur parmi ceux associés au document, puis choisit un thème à partir de la distribution thématique propre à cet auteur avant de générer le mot correspondant.

Le modèle permet ainsi d'étudier simultanément :

- les thèmes présents dans le corpus ;
- les préférences thématiques des différents auteurs.

Cette approche est particulièrement adaptée à l'analyse de corpus académiques, journalistiques ou politiques, dans lesquels l'identité des auteurs constitue une dimension informative importante pour comprendre les variations discursives et les orientations thématiques.

2.2.8 Extensions temporelles et modèles structurels

Afin d'intégrer explicitement la dimension temporelle des discours, plusieurs extensions du LDA ont été développées pour modéliser l'évolution des thèmes dans le temps. Parmi les plus influentes figurent le *Dynamic Topic Model* (DTM) proposé par Blei et Lafferty (2006), ainsi que le *Structural Topic Model* (STM), qui introduit des variables explicatives externes dans la modélisation thématique.

Le DTM permet d'étudier l'évolution progressive des thèmes au sein de corpus chronologiques en modélisant leurs transformations lexicales à travers le temps. Cette approche est particulièrement adaptée à l'analyse de phénomènes médiatiques ou sociaux évolutifs, tels que les catastrophes naturelles, les crises sanitaires ou le changement climatique.

Dans le domaine de la communication environnementale, Meier et Eskjaer (2024) mobilisent le *Structural Topic Model* (STM) afin d'analyser plus de trente années de couverture médiatique danoise consacrée au changement climatique. En intégrant des covariables telles que l'année de publication ou le type de média, les auteurs mettent en évidence une transformation progressive des priorités discursives, passant des controverses scientifiques vers les politiques climatiques, les solutions techniques et les stratégies d'adaptation.

Dans la suite de ce mémoire, une attention particulière est portée au *Dynamic Topic Model* (DTM), en raison de sa capacité à représenter explicitement l'évolution temporelle des thématiques médiatiques.

2.2.8.1 Description : Le Dynamic Topic Model (DTM)

Le *Dynamic Topic Model* (DTM) constitue une extension temporelle du LDA introduisant une dépendance explicite entre les thèmes observés à différentes périodes.

Le corpus est d'abord découpé en T tranches chronologiques (années, semestres, trimestres, etc.). Contrairement au LDA classique, les distributions lexicales des thèmes ne sont plus supposées fixes : elles évoluent progressivement au cours du temps selon un processus gaussien.

2.2.8.2 Processus génératif

Pour chaque thème k et chaque période temporelle t, le modèle suppose :

$$\beta_k^{(t)} \mid \beta_k^{(t-1)} \sim \mathcal{N}\left(\beta_k^{(t-1)}, \sigma^2 \mathbf{I}\right) \tag{2.7}$$

où :

- $\beta_k^{(t)}$ représente la distribution lexicale du thème k à la période t ;
- σ^2 contrôle la vitesse de dérive thématique ;
- $\mathbf{I}$ désigne la matrice identité.

Une faible valeur de σ^2 impose une évolution lente et progressive des thèmes, tandis qu'une valeur plus élevée autorise des changements lexicaux plus abrupts entre deux périodes successives.

Les distributions lexicales sont ensuite normalisées à l'aide d'une transformation logistique (*softmax*) afin de produire des distributions de probabilités valides sur le vocabulaire :

$$\tilde{\beta}_{k,w}^{(t)} = \frac{\exp\left(\beta_{k,w}^{(t)}\right)}{\sum_{w'=1}^{V} \exp\left(\beta_{k,w'}^{(t)}\right)} \tag{2.8}$$

Pour les documents appartenant à la période t, les proportions thématiques sont ensuite générées selon :

$$\theta_d^{(t)} \sim \mathrm{Dir}\left(\alpha^{(t)}\right), \qquad z_{d,n} \sim \mathrm{Mult}\left(\theta_d^{(t)}\right), \qquad w_{d,n} \sim \mathrm{Mult}\left(\tilde{\beta}_{z_{d,n}}^{(t)}\right) \tag{2.9}$$

La différence fondamentale avec le LDA réside dans le fait que les distributions lexicales $\beta_k^{(t)}$ dépendent explicitement des distributions de la période précédente $\beta_k^{(t-1)}$. Chaque thème évolue donc progressivement au fil du temps tout en conservant une certaine continuité historique.

où :

- $\beta_k^{(t)} \in \mathbb{R}^V$ représente la distribution lexicale non normalisée du thème k à la période t ;
- $\tilde{\beta}_{k,w}^{(t)}$ correspond au poids normalisé du mot w dans le thème k à la période t ;
- $\theta_d^{(t)}$ désigne la distribution thématique du document d appartenant à la période t ;
- σ^2 contrôle l'intensité de la dérive thématique ;
- T représente le nombre total de périodes temporelles.

Table 2.3: Répartition des six articles sur trois périodes temporelles

[H]

Période	Articles	Dominante discursive
$t = 1$ **(2010–2013)**	A1, A2	Réaction immédiate aux crues et gestion de l'urgence
$t = 2$ **(2014–2017)**	A3, A4	Couverture médiatique et bilan des dégâts
$t = 3$ **(2018–2021)**	A5, A6	Prévention, politiques publiques et dispositifs d'alerte

i Illustration sur mini-corpus — DTM

Les résultats présentés ci-dessous sont simulés à des fins pédagogiques. Les six articles fictifs introduits dans la section Section 2.2.1 sont répartis sur trois périodes temporelles distinctes afin d'illustrer l'évolution progressive des distributions lexicales $\beta_k^{(t)}$ d'une période à l'autre.

2.2.8.3 Illustration : Dynamic Topic Model sur le mini-corpus

Intuition

Le DTM peut être vu comme une version dynamique du LDA. Là où le LDA fournit une représentation statique des thèmes sur l'ensemble du corpus, le DTM permet de suivre leur évolution au fil du temps.

Analogie. Imaginons le thème « Hydrologie » comme un nuage de mots évolutif. Au début de la période étudiée, ce thème est principalement associé aux termes « inondation », « digue » et « évacuation ». Progressivement, des termes comme « alerte » et « prévention » prennent davantage d'importance, traduisant un déplacement du discours médiatique vers l'anticipation et la gestion préventive des risques.

Correspondance mathématique. Cette évolution est modélisée par le processus gaussien :

$$\beta_k^{(t)} \mid \beta_k^{(t-1)} \sim \mathcal{N}\left(\beta_k^{(t-1)}, \sigma^2 \mathbf{I}\right),$$

ce qui signifie que chaque thème hérite de la période précédente tout en pouvant évoluer progressivement sous l'effet d'une perturbation aléatoire contrôlée par σ^2.

1. Répartition temporelle du mini-corpus

2. Évolution de $\beta_k^{(t)}$: dérive lexicale des thèmes

Thème T1 — Hydrologie & Urgence (évolution sur trois périodes).

Période	inond.	digue	crue	alerte	infra.	évac.	prév.	autres
$\beta_{T1}^{(1)}$	**.28**	**.20**	.08	.04	.18	.16	.04	.02
$\beta_{T1}^{(2)}$	**.26**	**.17**	.10	.08	.20	.10	.07	.02
$\beta_{T1}^{(3)}$	.22	.12	.12	**.16**	.18	.06	**.12**	.02
Tendance	↘	↘	↗	↗↗	→	↘↘	↗↗	→

Thème T2 — Gestion & Prévention (évolution sur trois périodes).

Période	crue	gouv.	alerte	prév.	infra.	budget	autres
$\beta_{T2}^{(1)}$	.18	**.22**	.20	.18	.10	.08	.04
$\beta_{T2}^{(2)}$	.16	.20	.22	.20	.10	.08	.04
$\beta_{T2}^{(3)}$	.14	.18	**.24**	**.22**	.10	.08	.04
Tendance	↘	↘	↗	↗	→	→	→

Lecture des tableaux

Chaque ligne $\beta_{T_k}^{(t)}$ représente une distribution de probabilité sur le vocabulaire à une période donnée t ; la somme des coefficients d'une ligne est donc égale à 1.

Pour interpréter la dérive thématique, comparez les colonnes verticalement d'une période à l'autre :

- ↗ indique une augmentation progressive ;
- ↗↗ une augmentation marquée ;
- ↘ une diminution ;
- → une relative stabilité.

Par exemple, dans le thème T1, le poids du mot « alerte » passe de .04 à .08, puis à .16 : son importance relative est multipliée par quatre entre la première et la troisième période. Cette évolution traduit un déplacement du discours médiatique : l'attention se déplace progressivement des conséquences immédiates des crues (digues, évacuations, dégâts) vers les dispositifs d'anticipation et de prévention.

Ce type de transformation constitue précisément l'apport majeur du DTM : contrairement au LDA classique, le modèle permet de représenter explicitement l'évolution temporelle des thèmes.

Table 2.4: Comparaison entre le LDA et le DTM

Critère	LDA	DTM
Représentation de β_k	Distribution unique sur l'ensemble du corpus	Distribution spécifique à chaque période temporelle
Dérive lexicale	Invisible : les variations temporelles sont moyennées	Visible : les évolutions lexicales sont suivies explicitement
Hypothèse temporelle	Le temps est ignoré	Dépendance temporelle explicite entre $\beta_k^{(t-1)}$ et $\beta_k^{(t)}$
Paramètre supplémentaire	Aucun	σ^2 contrôle la vitesse de dérive thématique
Question principale	Quels thèmes structurent le corpus ?	Comment les thèmes évoluent-ils dans le temps ?

2.2.9 Choix et évaluation du nombre de topics pertinents

L'un des principaux enjeux de la modélisation thématique réside dans la détermination du nombre optimal de thèmes latents K. Dans les modèles probabilistes classiques, ce paramètre doit généralement être fixé avant l'apprentissage et influence directement la qualité, la stabilité et l'interprétabilité des résultats obtenus.

Afin de sélectionner une valeur pertinente de K, nous mobilisons la mesure de cohérence thématique C_v proposée par Roder, Both, et Hinneburg (2015). Cette métrique vise à évaluer la cohérence sémantique des thèmes en mesurant le degré de similarité entre les mots les plus représentatifs d'un même topic.

Le score C_v repose sur : - une segmentation des mots à l'aide d'une fenêtre glissante ; - la construction de vecteurs de cooccurrence ; - puis le calcul d'une similarité cosinus entre ces vecteurs contextuels.

Formellement, la mesure de confirmation indirecte utilisée par C_v s'écrit :

$$m_{ic}(\mathbf{u}, \mathbf{w}) = \frac{\sum_i u_i w_i}{\|\mathbf{u}\|_2 \cdot \|\mathbf{w}\|_2}$$

où :

- $\mathbf{u}$ et $\mathbf{w}$ désignent des vecteurs de cooccurrence ;
- les probabilités de cooccurrence sont estimées à partir d'une fenêtre glissante, généralement fixée à 11 mots ;
- $\|\cdot\|_2$ représente la norme euclidienne.

Un thème est considéré comme cohérent lorsque les mots qui le composent présentent une forte similarité sémantique, ce qui améliore leur interprétabilité humaine.

2.2.9.1 Comparaison avec la cohérence $C_{U_{\text{mass}}}$

La métrique C_v n'est toutefois pas la seule approche disponible pour évaluer la qualité des topics. Une autre mesure fréquemment utilisée est la cohérence $C_{U_{\text{mass}}}$ proposée par Mimno et al. (2011).

Contrairement à C_v, qui repose sur une mesure contextuelle de similarité sémantique, $C_{U_{\text{mass}}}$ évalue la cohérence intrinsèque des thèmes à partir des probabilités de cooccurrence observées directement dans les documents du corpus.

Elle est définie par :

$$C_{U_{\text{mass}}}(w_1, \ldots, w_N) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \frac{P(w_i, w_j) + \epsilon}{P(w_j)}$$

où :

- $P(w_i, w_j)$ désigne la probabilité de cooccurrence des mots w_i et w_j ;
- $P(w_j)$ correspond à la probabilité marginale du mot w_j ;
- ϵ représente un terme de lissage évitant les indéterminations lorsque certaines cooccurrences sont absentes.

2.2.9.2 Choix de C_v

Dans le cadre de ce mémoire, nous privilégions la cohérence C_v pour sélectionner le nombre optimal de thèmes.

Ce choix s'explique principalement par la nature discursive et journalistique du corpus étudié. Les articles de presse mobilisent en effet :

- un vocabulaire riche ;
- des formulations variées ;
- des synonymes ;
- ainsi que des expressions métaphoriques ou contextuelles.

Dans ce contexte, une approche fondée uniquement sur les cooccurrences directes au sein des documents, comme $C_{U_{\text{mass}}}$, peut sous-estimer certaines proximités sémantiques importantes. À l'inverse, la fenêtre glissante utilisée par C_v permet de capturer des relations lexicales plus fines et plus adaptées à l'interprétation humaine des thèmes.

Cette approche est largement utilisée dans les travaux récents d'analyse textuelle à grande échelle, notamment dans les recherches consacrées aux discours médiatiques et environnementaux (Meier et Eskjaer 2024).

2.2.9.3 Comparaison avec la perplexité

Le choix du nombre de thèmes peut également être évalué à l'aide de la **perplexité**, mesure issue de la théorie de l'information qui évalue la capacité prédictive d'un modèle sur des documents non observés.

Formellement, la perplexité est définie par :

$$\text{Perplexity}(D) = \exp\left(-\frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right)$$

où :

- $p(\mathbf{w}_d)$ représente la probabilité du document d sous le modèle et $\mathbf{w}$ le vecteur de mots du document d ;
- N_d désigne le nombre total de mots du document.

Une faible perplexité indique généralement une meilleure capacité de généralisation statistique. Toutefois, plusieurs travaux montrent que cette mesure tend à favoriser des modèles comportant un grand nombre de topics, sans garantir pour autant leur interprétabilité sémantique.

Pour cette raison, la perplexité est souvent considérée comme insuffisante lorsqu'il s'agit d'évaluer la qualité interprétative des thèmes extraits.

2.2.9.4 Évaluation de la stabilité thématique

Au-delà de la cohérence sémantique, nous évaluons également la stabilité des topics à l'aide de l'indice de Jaccard moyen $\bar{J}$ proposé par Jaccard (1901) et mobilisé dans les travaux de Greene, O'Callaghan, et Cunningham (2014).

Cette approche consiste à :

- entraîner plusieurs modèles avec différentes graines aléatoires ;
- comparer les ensembles de mots associés aux thèmes obtenus ;
- puis mesurer leur degré de recouvrement lexical.

Un score de Jaccard élevé indique que les thèmes extraits restent relativement stables malgré les variations liées à l'initialisation aléatoire du modèle.

2.2.10 Interprétation et visualisation des modèles thématiques

L'interprétation des modèles thématiques constitue une étape essentielle du *topic modeling*. Au-delà de la performance statistique, la qualité d'un modèle dépend également de la capacité des chercheurs à comprendre, nommer et analyser les thèmes extraits.

Dans cette perspective, Sievert et Shirley (2014) proposent l'outil interactif **LDAvis**, conçu pour faciliter l'exploration visuelle des modèles LDA. L'approche repose sur une représentation bidimensionnelle des thèmes ainsi que sur une mesure de pertinence lexicale permettant d'identifier les mots les plus représentatifs de chaque topic.

La pertinence d'un terme combine deux dimensions complémentaires : - sa fréquence globale dans le thème ; - son caractère distinctif par rapport aux autres thèmes du corpus.

Cette stratégie permet de limiter la domination des mots très fréquents mais peu informatifs, tout en mettant davantage en évidence les termes réellement discriminants pour l'interprétation des topics.

Les outils de visualisation jouent ainsi un rôle central dans la validation exploratoire des modèles thématiques. Ils facilitent :

- l'identification des thèmes dominants ;
- l'analyse des proximités entre topics ;
- la détection de recouvrements sémantiques ;
- ainsi que l'interprétation qualitative des distributions lexicales.

Ces approches sont particulièrement importantes dans les recherches interdisciplinaires, où des spécialistes issus de domaines différents (statistique, traitement automatique du langage, sciences sociales ou communication) doivent collaborer à l'analyse et à l'interprétation des résultats produits par les modèles thématiques.

2.3 La révolution des Transformers (2017 à aujourd'hui)

2.3.1 Des prolongements lexicaux aux Transformers

Les méthodes classiques de traitement du langage ont longtemps buté sur la représentation en sac de mots, qui ignore l'ordre et le contexte. Une première avancée décisive est venue des plongements lexicaux[3] (*word embeddings*) : Word2Vec (Mikolov et al. (2013)) apprend, via un réseau de neurones peu profond, à prédire un mot à partir de son contexte (CBOW) ou l'inverse (Skip-gram), capturant ainsi la proximité sémantique dans un espace géométrique ; GloVe (Pennington, Socher, et Manning (2014)) pousse cette logique plus loin en factorisant la matrice de cooccurrences globale du corpus, combinant la richesse statistique de la LSA avec la flexibilité des plongements. Ces représentations restaient néanmoins statiques et incapables de s'adapter au sens d'un mot selon son contexte phrastique.

La véritable révolution est venue de Vaswani et al. (2017) avec l'article «**Attention Is All You Need**». En introduisant le mécanisme d'auto-attention, le Transformer permet de saisir les relations lointaines entre les mots sans la lourdeur des réseaux récurrents. Cette architecture, fondée sur l'attention multi-têtes et le codage positionnel, a ouvert la voie à des modèles pré-entraînés majeurs comme BERT (Devlin et al. 2019), supplantant les RNN et les LSTM dans la quasi-totalité des tâches du TAL.

2.3.2 BERT et les modèles contextuels profonds

BERT (*Bidirectional Encoder Representations from Transformers*), proposé par Devlin et al. (2019), constitue l'une des avancées majeures issues de l'architecture Transformer. Le modèle est pré-entraîné sur de très larges corpus textuels, notamment Wikipedia et des collections de livres, à partir de deux objectifs d'apprentissage complémentaires : le *Masked Language Model* (MLM) et la prédiction de phrase suivante (*Next Sentence Prediction*, NSP).

Le principe du *Masked Language Model* consiste à masquer aléatoirement certains mots d'une phrase afin de demander au modèle de les prédire à partir du contexte environnant. La tâche de *Next Sentence Prediction* vise, quant à elle, à déterminer si deux phrases se suivent réellement dans le corpus d'apprentissage. Contrairement aux modèles séquentiels précédents, BERT repose sur une représentation bidirectionnelle du langage : chaque mot est encodé en tenant compte simultanément du contexte situé à gauche et à droite de la phrase. Cette propriété permet de produire des représentations contextuelles profondes, dans lesquelles un même mot peut recevoir des vecteurs différents selon son usage et son environnement linguistique.

Cette capacité à modéliser finement le contexte améliore considérablement les performances sur de nombreuses tâches de traitement automatique du langage, notamment la classification de texte, les systèmes de question-réponse, la reconnaissance d'entités nommées et l'inférence en langage naturel. Les modèles dérivés de BERT ont rapidement établi de nouveaux résultats de référence (*state of the art*) dans la majorité des tâches de TAL, contribuant ainsi à transformer durablement le domaine du traitement automatique du langage naturel.

[3] Un plongement lexical (*word embedding*) est une représentation d'un mot sous forme de vecteur de nombres réels dans un espace multidimensionnel.

2.4 Modèles neuronaux de topic modelling

2.4.1 BERTopic

BERTopic, proposé par Grootendorst (2022), constitue aujourd'hui l'une des approches neuronales les plus avancées en modélisation thématique. Le modèle combine les représentations contextuelles profondes issues des Transformers avec des techniques modernes de réduction de dimensionnalité et de regroupement non supervisé.

L'approche se déroule en quatre grandes étapes successives.

1. La génération d'embeddings.

Dans un premier temps, chaque document d est transformé en un vecteur dense à l'aide d'un modèle Transformer pré-entraîné, généralement Sentence-BERT. Pour un document d, la représentation vectorielle s'écrit :

$$\mathbf{e}_d = f_\theta(d) \in \mathbb{R}^p$$

où:

- f_θ désigne le modèle Transformer pré-entraîné;
- $\mathbf{e}_d$ représente l'embedding du document d ;
- p correspond à la dimension de l'espace vectoriel latent.

Contrairement aux approches classiques fondées sur les fréquences de mots, ces représentations capturent directement le contexte sémantique global des phrases et des documents.

2. La réduction de dimensionnalité par UMAP.

Les embeddings produits par le Transformer sont ensuite projetés dans un espace de dimension plus faible à l'aide de l'algorithme UMAP. L'objectif d'UMAP est de préserver simultanément :

- la structure locale des voisinages ;
- ainsi que l'organisation globale de l'espace sémantique.

L'algorithme cherche à minimiser une divergence entre les probabilités de voisinage observées dans l'espace original de haute dimension ; et celles reconstruites dans l'espace réduit.

La fonction objectif peut s'écrire :

$$\mathcal{L} = \sum_{(i,j)} \left[w_{ij} \log \frac{w_{ij}}{\hat{w}_{ij}} + (1 - w_{ij}) \log \frac{1 - w_{ij}}{1 - \hat{w}_{ij}} \right]$$

où :

- $\mathcal{L}$ représente la fonction de perte (*loss function*) optimisée par UMAP ;
- (i, j) désigne une paire de documents dans l'espace des embeddings ;
- w_{ij} correspond à la probabilité de voisinage entre les points i et j dans l'espace original de haute dimension ;
- $\hat{w}_{ij}$ représente la probabilité de voisinage reconstruite dans l'espace réduit produit par UMAP ;

- log désigne le logarithme naturel ;
- le premier terme : $w_{ij} \log \frac{w_{ij}}{\hat{w}_{ij}}$ pénalise la perte de proximité entre deux points initialement proches ;
- le second terme : $(1-w_{ij}) \log \frac{1-w_{ij}}{1-\hat{w}_{ij}}$ pénalise les faux rapprochements entre points initialement éloignés.

Cette optimisation permet de conserver les principales relations géométriques de l'espace sémantique avant l'étape de clustering.

3. Le regroupement par HDBSCAN

Les documents projetés sont ensuite regroupés à l'aide de HDBSCAN proposé par Campello, Moulavi, et Sander (2013). Il constitue une extension hiérarchique de DBSCAN capable d'identifier automatiquement des groupes de densité variable dans l'espace des embeddings. Cette approche présente plusieurs avantages importants dans le cadre du *topic modeling*.

Premièrement, le nombre de clusters n'est pas fixé à l'avance : il émerge directement des données. Deuxièmement, les documents atypiques ou ambigus peuvent être explicitement identifiés comme du bruit et reçoivent alors le label -1, plutôt que d'être artificiellement forcés dans un groupe inadapté. L'algorithme repose sur la notion de distance mutuellement accessible (*mutual reachability distance*) définie par :

$$d_{\text{mreach}}(a,b) = \max\left(\text{core}_k(a),\ \text{core}_k(b),\ d(a,b)\right)$$

où :

- a et b désignent deux documents (ou deux points) dans l'espace des embeddings ;
- $d(a,b)$ correspond à la distance classique entre les deux documents, généralement calculée à l'aide de la distance euclidienne ou cosinus ;
- $\text{core}_k(a)$ représente la *core distance* du document a, c'est-à-dire la distance entre a et son k-ième plus proche voisin ;
- $\text{core}_k(b)$ représente la même quantité que $\text{core}_k(a)$ pour le document b ;
- k correspond au nombre de voisins utilisé pour estimer la densité locale autour d'un point ;
- l'opérateur $\max(\cdot)$ signifie que la distance mutuellement accessible entre deux points est définie comme la plus grande de ces trois valeurs.

Cette formulation permet d'éviter que deux points situés dans des régions peu denses soient considérés artificiellement proches. HDBSCAN favorise ainsi les regroupements correspondant à des zones sémantiques denses et homogènes.

4). La représentation des thèmes

Une fois les clusters constitués, BERTopic représente chaque thème à l'aide du *class-based* TF-IDF (c-TF-IDF) (Grootendorst 2022). L'approche consiste à concaténer l'ensemble des documents d'un même cluster pour former un unique « méga-document », puis à appliquer une variante adaptée du calcul TF-IDF. Pour un mot x au sein d'un cluster c, le score est défini par :

$$\text{c-TF-IDF}(w,c) = f_{w,c} \times \log\left(1 + \frac{A}{\text{DF}_x}\right)$$

où :

- $f_{w,c}$ représente la fréquence brute du terme x dans le cluster c ;
- DF_w désigne la fréquence document globale (*Document Frequency*), soit le nombre total de documents de l'ensemble du corpus contenant le mot w ;

- A correspond au nombre moyen de mots par cluster à l'échelle du corpus, agissant comme un facteur de normalisation pour lisser l'impact de la taille des clusters.

Dans cette formulation, le terme logarithmique fonctionne comme un puissant facteur de pénalisation globale. Plus un mot est omniprésent à travers les différents documents du corpus, plus son poids discriminant diminue. Les termes obtenant le score c-TF-IDF le plus élevé sont donc à la fois fréquents au sein du cluster et sélectifs par rapport au reste du corpus, constituant ainsi les meilleurs descripteurs sémantiques du thème.

2.4.2 Traitement des outliers

Un atout majeur de HDBSCAN est d'identifier les documents atypiques plutôt que de les assigner de force. Ces documents labellisés -1 représentent toutefois une part non négligeable du corpus, et les exclure de l'analyse serait préjudiciable. BERTopic propose trois stratégies de réassignation fondées sur des principes distincts.

Stratégie par embeddings. Chaque document considéré comme outlier est réassigné au topic dont le centroïde d'embedding est le plus proche, selon la similarité cosinus :

$$\hat{k}_i = \arg\max_{k \neq -1} \cos(\mathbf{e}_d,\ \bar{\mathbf{e}}_k)$$

où : $\bar{\mathbf{e}}_k = \frac{1}{|C_k|} \sum_{d \in C_k} \mathbf{e}_d$ est le centroïde du cluster (topic) k. Cette stratégie repose sur la propriété établie par Reimers et Gurevych (2019) selon laquelle la similarité cosinus entre vecteurs SBERT constitue une mesure fiable de proximité sémantique. éloignés des thèmes existants.

Cette stratégie est particulièrement efficace lorsque le contenu sémantique du document reste clair, mais que celui-ci se situe dans une région localement peu dense de l'espace vectoriel.

Stratégie par c-TF-IDF. La seconde stratégie repose non plus sur les embeddings, mais directement sur les représentations lexicales construites par le c-TF-IDF. Le document atypique est réassigné au topic maximisant la similarité cosinus entre :

- son vecteur c-TF-IDF ;
- et le vecteur représentatif de chaque thème.

Formellement :

$$\hat{k}_i = \arg\max_{k \neq -1} \cos\big(\mathbf{v}_d^{\text{c-tfidf}},\ \mathbf{r}_k^{\text{c-tfidf}}\big)$$

où :

- $\mathbf{v}_d^{\text{c-tfidf}} \in \mathbb{R}^V$ est le vecteur c-TF-IDF du document outlier d : chaque composante correspond au score c-TF-IDF d'un mot du vocabulaire V dans ce document, calculé comme s'il constituait à lui seul un cluster temporaire ;
- $\mathbf{r}_k^{\text{c-tfidf}} \in \mathbb{R}^V$ est le vecteur c-TF-IDF du topic t : il est obtenu en concaténant l'ensemble des documents assignés au cluster t en un unique méga-document, puis en appliquant le calcul c-TF-IDF défini par Grootendorst (2022) :

$$\text{c-TF-IDF}(w, k) = f_{w,k} \times \log\left(1 + \frac{A}{\text{DF}_w}\right)$$

où $f_{w,k}$ est la fréquence du mot w dans le cluster (topic) k, DF_w sa fréquence globale dans le corpus, et A le nombre moyen de mots par cluster.

L'avantage de cette méthode est qu'elle est entièrement indépendante de l'espace UMAP : elle raisonne directement sur les distributions lexicales, ce qui la rend robuste aux distorsions géométriques introduites par la réduction de dimensionnalité.

Stratégie par distributions de probabilités. Lorsque le modèle est entraîné avec `calculate_probabilities=True` HDBSCAN produit pour chaque document une distribution de probabilité d'appartenance à chaque cluster. Pour les outliers, on retient simplement le topic de probabilité non nulle la plus élevée~:

$$\hat{k}_i = \arg\max_{k \neq -1} \; P(k \mid d)$$

Cette méthode est la plus fine conceptuellement : elle exploite l'incertitude native du modèle plutôt que de construire une heuristique de distance a posteriori. En contrepartie, elle est plus coûteuse en mémoire et ne fonctionne que si les probabilités ont été calculées à l'entraînement.

Au-delà de ces trois stratégies, Groot, Aliannejadi, et Haas (2022) ont proposé de remplacer l'algorithme natif du HDBSCAN par celui du K-means, ce qui obligerait tous les articles à être assignés à des clusters. Cette approche permettrait ainsi de résoudre le problème des outliers. Mais pose la difficulté de choisir le nombre de de topics (clusters) à l'avance. Cependant, cette approche modifie profondément la logique du modèle. Contrairement à HDBSCAN, KMeans suppose des clusters sphériques et homogènes, ce qui peut être moins adapté à la structure complexe des espaces d'embeddings sémantiques. C'est pourquoi cette option ne sera pas envisager dans ce travail.

2.4.2.1 Limites du BERTopic

Malgré ses performances, BERTopic présente plusieurs limites méthodologiques. Le modèle est fortement dépendant de la qualité des embeddings produits par le Transformer utilisé~: des représentations mal adaptées au domaine étudié peuvent conduire à des regroupements sémantiques peu cohérents. Il est également sensible aux hyperparamètres d'UMAP et de HDBSCAN~: de faibles variations de `n_neighbors`, `min_dist` ou `min_cluster_size` peuvent modifier substantiellement le nombre et la structure des topics obtenus. L'étape de réduction de dimensionnalité introduit par ailleurs une part de stochasticité, ce qui peut nuire à la reproductibilité exacte des résultats. Enfin, BERTopic traite chaque document comme mono-thème (Grootendorst (2022)), ce qui ne correspond pas toujours à la réalité des articles de presse, souvent plurithématiques.

2.4.2.2 Variantes et améliorations du modèle

Des travaux consacrés à l'optimisation des hyperparamètres de BERTopic dans des corpus scientifiques, notamment ceux de Egger et Yu (2022), montrent que le choix des paramètres associés à UMAP et HDBSCAN influence fortement la qualité des thèmes extraits. Les auteurs soulignent en particulier que les paramètres contrôlant la réduction de dimensionnalité, la densité des clusters ou encore la taille minimale des groupes peuvent modifier de manière substantielle la cohérence et la granularité des topics identifiés. Ces résultats mettent ainsi en évidence l'importance d'une démarche rigoureuse d'optimisation des hyperparamètres dans l'utilisation de BERTopic.

Par ailleurs, plusieurs travaux récents ont étudié l'adaptation de BERTopic à des contextes plus complexes, notamment les corpus de textes courts Islam et al. (2023) ou les corpus fortement hétérogènes couvrant

plusieurs domaines thématiques Bianchi, Terragni, et Hovy (2021). Ces recherches montrent que les performances du modèle diminuent lorsque les documents contiennent peu d'information textuelle ou lorsque les espaces sémantiques deviennent trop dispersés.

Afin de pallier ces limites, différentes stratégies de réorganisation thématique ont été proposées, parmi lesquelles des procédures de *re-clustering* visant à regrouper ou subdiviser certains topics afin d'améliorer leur cohérence interne et leur granularité sémantique.

2.4.3 Top2vec

Angelov (2020) introduit Top2Vec, une méthode de modélisation thématique qui apprend conjointement les plongements de documents et de mots dans un espace sémantique continu (via Doc2Vec). Contrairement au modèle LDA, Top2Vec s'affranchit de la nécessité de spécifier le nombre de sujets à l'avance : il utilise des algorithmes de réduction dimensionnelle et de clustering (typiquement HDBSCAN) pour identifier automatiquement des groupes de documents sémantiquement proches.

Mathématiquement, Top2Vec et BERTopic diffèrent dans leur définition même d'un topic. Dans Top2Vec, un topic correspond à une région dense de l'espace vectoriel des embeddings. Chaque topic est représenté par un vecteur centroid :

$$\mathbf{z}_k = \frac{1}{|C_k|} \sum_{d \in C_k} \mathbf{e}_d$$

où $\mathbf{e}_d$ représente l'embedding du document d et C_k l'ensemble des documents appartenant au topic k. Les mots caractéristiques du topic sont ensuite sélectionnés selon leur proximité vectorielle avec ce centroïde.

À l'inverse, BERTopic construit explicitement une représentation lexicale des clusters à l'aide du c-TF-IDF. Les thèmes extraits ne correspondent donc pas uniquement à des régions géométriques dans l'espace des embeddings, mais également à des distributions lexicales discriminantes facilitant leur interprétation.

Top2Vec présente néanmoins plusieurs limites. D'une part, les topics y sont principalement définis à partir de proximités géométriques dans l'espace vectoriel des embeddings. Cette approche peut réduire l'interprétabilité lexicale des thèmes : des mots proches sémantiquement dans l'espace latent ne forment pas toujours une thématique clairement identifiable pour un analyste humain.

D'autre part, Top2Vec offre une modularité méthodologique plus limitée que BERTopic. Les différentes étapes du pipeline y sont fortement intégrées, ce qui réduit les possibilités d'ajustement indépendant des embeddings, du clustering ou encore des mécanismes de représentation thématique.

À l'inverse, BERTopic permet de modifier séparément :

- les modèles d'embeddings ;
- les techniques de réduction de dimensionnalité ;
- les algorithmes de clustering ;
- ainsi que les méthodes de représentation des topics.

Cette modularité constitue un avantage important dans les analyses comparatives et expérimentales, où l'optimisation indépendante de chaque composante peut améliorer significativement la qualité des résultats.

Enfin, Top2Vec demeure généralement moins adapté aux analyses temporelles complexes ainsi qu'aux contextes nécessitant une forte explicabilité des résultats, notamment dans les travaux interdisciplinaires mobilisant des experts non spécialistes du traitement automatique du langage.

2.4.4 BERTopic dynamique

Le BERTopic dynamique (*Dynamic Topic Modeling*) Grootendorst (2022) étend le modèle BERTopic classique en intégrant explicitement la dimension temporelle afin d'analyser l'évolution des thématiques dans le temps. Après la génération des thèmes globaux, les documents sont partitionnés par périodes temporelles. Les représentations textuelles de chaque thème sont ensuite recalculées localement pour chaque période à l'aide d'une variante dynamique du *class-based* TF-IDF (c-TF-IDF).

Pour une période temporelle t, le poids d'un mot w dans un thème c est estimé par :

$$\mathrm{c-TF-IDF}_{w,c,t} = f_{w,c,t} \times \log\left(1 + \frac{A}{\mathrm{DF}_w}\right)$$

où :

- $f_{w,c,t}$ représente la fréquence brute du mot w au sein du thème (cluster) c spécifiquement durant la période t ;
- DF_w désigne la fréquence document globale (*Document Frequency*) du mot w calculée sur l'ensemble du corpus, toutes périodes et thèmes confondus ;
- A correspond au nombre moyen de mots par classe à l'échelle globale, agissant comme un facteur de régularisation pour lisser l'impact de la taille des clusters.

Cette approche permet de suivre précisément l'évolution lexicale des thèmes au fil du temps tout en conservant une structure macro-thématique stable (les documents ne changent pas de cluster). Les variations de distributions de mots d'une période à l'autre peuvent ensuite être quantifiées à l'aide de mesures de distance, telles que la divergence de Jensen-Shannon, afin d'identifier les phases de rupture ou de transition sémantique.

2.5 Comparaison des approches probabilistes et neuronales

La confrontation entre les approches probabilistes, comme le LDA ou le DTM, et les approches neuronales récentes, telles que BERTopic ou Top2Vec, occupe aujourd'hui une place centrale dans les débats méthodologiques du traitement automatique du langage.

Les modèles probabilistes présentent plusieurs avantages importants. Ils reposent sur un cadre mathématique explicite et sur des hypothèses clairement définies concernant le processus génératif des textes. Cette structure théorique favorise une interprétation relativement transparente des résultats, ce qui explique leur succès durable dans les travaux de sciences sociales, d'analyse médiatique ou de recherche académique. Les distributions thématiques produites par ces modèles restent généralement faciles à expliquer, à discuter et à relier aux phénomènes étudiés.

Ces approches présentent toutefois certaines limites. Elles demeurent sensibles au choix des hyperparamètres, en particulier au nombre de topics K, dont la détermination reste souvent délicate. De plus, les modèles probabilistes classiques peinent à capturer des relations sémantiques complexes lorsque celles-ci dépassent les simples cooccurrences lexicales observées dans les documents.

Les approches neuronales cherchent précisément à dépasser ces limitations. En s'appuyant sur des représentations vectorielles denses pré-entraînées sur de très grands corpus, elles parviennent généralement à produire des thèmes plus cohérents sur le plan sémantique. Les embeddings contextuels permettent notamment de mieux prendre en compte les synonymes, les variations lexicales et les proximités de sens qui échappent souvent aux modèles probabilistes traditionnels.

Cette amélioration de la richesse sémantique s'accompagne néanmoins d'une perte partielle d'interprétabilité. Les modèles neuronaux reposent sur des architectures complexes, parfois difficiles à expliquer de manière intuitive, ce qui peut rendre l'analyse des résultats moins transparente pour des chercheurs non spécialistes du TAL ou de l'apprentissage profond.

Dans le cadre de l'analyse de grands corpus journalistiques consacrés aux questions environnementales et aux catastrophes naturelles, les modèles probabilistes conservent ainsi un avantage important lorsque l'objectif principal est l'interprétation et la discussion des résultats avec des experts issus d'autres disciplines. Comme le soulignent Tanguy et al. (2025), la transparence méthodologique et la capacité à produire des thèmes facilement interprétables demeurent des critères essentiels dans les recherches interdisciplinaires mobilisant des spécialistes des médias, des sciences sociales ou de l'environnement.

2.6 Applications à la communication sur les risques naturels et climatiques

2.6.1 Invisibilité médiatique de certains risques

Les travaux de Rouquette et Bihay (2022) s'intéressent à la question des « risques naturels médiatiquement invisibles ». Les auteurs montrent que certains phénomènes pourtant susceptibles d'avoir des conséquences importantes restent peu présents dans l'espace médiatique. C'est notamment le cas des phénomènes lents ou difficilement perceptibles, comme le retrait-gonflement des argiles ou certaines formes de sécheresse progressive.

Selon eux, cette faible visibilité s'explique en partie par la temporalité particulière de ces risques. Contrairement aux catastrophes soudaines, telles que les inondations ou les tempêtes, ces phénomènes évoluent lentement et produisent moins d'images spectaculaires susceptibles d'attirer l'attention médiatique. Leur forte complexité scientifique contribue également à rendre leur traitement journalistique plus difficile.

Ces constats rejoignent les travaux fondateurs de Vasterman (2005) sur la notion de **media hype**, qui montre que la couverture médiatique des catastrophes suit des dynamiques d'amplification auto-entretenues : un événement déclencheur spectaculaire génère une attention médiatique disproportionnée, au détriment des risques chroniques ou diffus. Dans ce cadre, les aléas à manifestation lente — sécheresses, érosion côtière, tassements de terrain — peinent structurellement à atteindre le seuil de visibilité nécessaire pour intégrer l'agenda médiatique.

Cette sélectivité médiatique est par ailleurs amplifiée par des logiques éditoriales bien documentées. Boykoff et Boykoff (2007) montrent que les médias de masse tendent à privilégier la nouveauté, le conflit et la proximité géographique dans le traitement des risques environnementaux, ce qui favorise mécaniquement les catastrophes soudaines et internationalement visibles au détriment des processus lents et locaux. Joye (2010) prolonge cette analyse en introduisant la notion de **distant suffering** : la souffrance lointaine n'est médiatisée que lorsqu'elle atteint un seuil de dramatisation suffisant, ce qui explique en partie la surreprésentation des catastrophes tropicales majeures dans les corpus journalistiques francophones, au détriment des risques propres aux latitudes tempérées.

Les auteurs soulignent ainsi l'existence de déséquilibres importants dans la couverture médiatique des risques naturels. Leur étude, principalement qualitative, met en évidence l'intérêt d'approches quantitatives et computationnelles, notamment le topic modeling, pour analyser de manière systématique les thèmes dominants et les évolutions discursives associées aux catastrophes naturelles.

2.6.2 Indicateurs médiatiques de sensibilisation au changement climatique

Au-delà de la simple visibilité des risques, plusieurs travaux ont cherché à mesurer dans quelle mesure la couverture médiatique contribue effectivement à la sensibilisation du public au changement climatique. Boykoff (2010) proposent à cet égard un cadre d'analyse fondé sur le suivi longitudinal du volume et du cadrage des articles consacrés au changement climatique dans la presse internationale. Leurs résultats montrent que cette couverture médiatique a connu une croissance significative entre 2004 et 2008, avant de se stabiliser, voire de diminuer dans certains pays sous l'effet de crises économiques et politiques concurrentes.

Dans le prolongement de ces travaux, Schmidt, Ivanova, et Schäfer (2013) analyse les facteurs expliquant les variations nationales de l'intensité de la couverture médiatique climatique. L'auteur met en évidence plusieurs déterminants majeurs, parmi lesquels la structure du système médiatique, le positionnement politique des rédactions ainsi que la survenue d'événements climatiques extrêmes locaux. Ces résultats suggèrent que la sensibilisation médiatique aux risques naturels ne relève pas d'un processus linéaire, mais résulte plutôt d'interactions complexes entre événements réels, dynamiques politiques et logiques propres aux industries médiatiques.

2.6.3 Évolution diachronique de la couverture climatique

La synthèse de littérature proposée par Lalande et Yates (2025) met en évidence plusieurs grands axes de recherche en communication climatique, parmi lesquels la perception publique du changement climatique, les stratégies discursives utilisées par les médias et les institutions, ainsi que les obstacles cognitifs à la réception des messages scientifiques. Les autrices insistent également sur l'apport croissant des méthodes de traitement automatique du langage dans l'étude des grands corpus médiatiques et institutionnels.

Dans cette perspective, Meier et Eskjaer (2024) mobilisent le *Structural Topic Model* afin d'étudier plus de trente années de couverture médiatique danoise consacrée au changement climatique. Leur analyse met en évidence une transformation progressive des discours médiatiques : les controverses scientifiques dominantes dans les premières périodes cèdent progressivement la place aux questions liées aux politiques publiques, aux solutions technologiques et aux stratégies d'adaptation climatique. Les auteurs observent également une montée progressive de la polarisation discursive à partir des années 2010, reflétant les tensions politiques et sociales croissantes autour des enjeux climatiques.

De leur côté, Huang et Xiao (2015) appliquent le *topic modeling* à des données issues de Twitter dans le cadre de la gestion des catastrophes naturelles. Leur approche combine le LDA avec des informations géospatiales afin de détecter les zones les plus touchées et de suivre l'évolution des besoins exprimés par les populations, notamment en matière d'abris, de nourriture ou de secours d'urgence. Cette étude met en évidence l'intérêt de combiner médias traditionnels et médias sociaux dans l'analyse des catastrophes naturelles : les médias institutionnels permettent généralement une couverture plus structurée et contextualisée, tandis que les réseaux sociaux offrent un accès plus immédiat aux réactions, aux besoins et aux informations circulant en temps réel sur le terrain.

2.7 Problématique et positionnement de notre recherche

La revue de littérature met en évidence plusieurs limites méthodologiques et applicatives que ce mémoire cherche à explorer et à dépasser.

Sur le plan méthodologique, peu de travaux proposent une comparaison approfondie entre les modèles probabilistes classiques, tels que le LDA ou le DTM, et les approches neuronales récentes comme BERTopic ou Top2Vec, appliquées à un même corpus médiatique réel. Les études existantes évaluent généralement

ces modèles de manière séparée ou sur des jeux de données très différents, ce qui rend les comparaisons difficiles.

Par ailleurs, les évaluations mobilisées restent souvent partielles. Les travaux se concentrent soit sur la cohérence sémantique des topics, soit sur leurs performances computationnelles, sans intégrer simultanément plusieurs dimensions essentielles comme :

- la cohérence thématique ;
- la stabilité des résultats entre différentes exécutions ;
- la capacité à représenter des évolutions temporelles fines ;
- ou encore l'interprétabilité des thèmes pour des chercheurs non spécialistes du traitement automatique du langage.

Ces questions apparaissent pourtant centrales dans les recherches interdisciplinaires consacrées aux médias, à l'environnement ou aux risques naturels.

Sur le plan applicatif, les usages de BERTopic dans l'étude des risques naturels et climatiques demeurent encore relativement limités. La majorité des recherches mobilisant ce modèle portent principalement sur des corpus académiques ou sur des données issues des réseaux sociaux, dont la structure linguistique diffère fortement de celle des articles de presse journalistiques.

De plus, dans de nombreuses études, la dimension temporelle est souvent reconstruite a posteriori à partir des résultats de BERTopic, alors que des modèles comme le DTM proposent directement une modélisation dynamique intégrée des évolutions thématiques.

Ces constats orientent directement les choix méthodologiques retenus dans ce mémoire.

Le LDA est utilisé comme point de départ en raison de la solidité de son cadre probabiliste et de son haut niveau d'interprétabilité. Son fonctionnement repose sur des distributions de probabilité relativement faciles à interpréter (Blei, Ng, et Jordan (2003) ; Roder, Both, et Hinneburg (2015) ; Sievert et Shirley (2014)) pour des chercheurs issus des sciences sociales ou de la communication, ce qui constitue un avantage important dans un travail interdisciplinaire.

Le DTM est retenu pour sa capacité à modéliser explicitement l'évolution temporelle des thèmes. Contrairement à d'autres extensions du LDA, il permet de suivre directement les transformations lexicales des topics au fil du temps sans nécessiter l'ajout de covariables externes. Son implémentation dans **tomotopy** offre par ailleurs une accessibilité pratique supérieure aux versions originales en C++.

Du côté des approches neuronales, BERTopic est privilégié pour sa modularité et sa flexibilité méthodologique. Le modèle permet notamment de modifier indépendamment :

- les embeddings ;
- les techniques de réduction de dimensionnalité ;
- les méthodes de clustering ;
- ainsi que les mécanismes de représentation des topics.

Cette modularité facilite les comparaisons expérimentales et permet d'adapter plus finement le modèle aux spécificités du corpus étudié. BERTopic présente également des performances sémantiques généralement supérieures aux approches probabilistes classiques, notamment sur les corpus hétérogènes et riches lexicalement.

Top2Vec, bien qu'intéressant sur le plan conceptuel, n'a finalement pas été retenu comme modèle principal en raison de sa moindre modularité et de difficultés plus importantes d'interprétation lexicale.

Enfin, des modèles comme le CTM ou le HDP, malgré leur intérêt théorique, ont été écartés en raison de leur complexité computationnelle élevée et de leur interprétation parfois plus délicate sur des corpus médiatiques volumineux.

2.7.1 Contribution de ce mémoire

Afin de répondre à ces différents enjeux, ce mémoire adopte une démarche progressive, comparative et interdisciplinaire.

L'analyse débute par une phase de préparation et de tri lexical du corpus (Section 3.3), avant la mise en œuvre de plusieurs familles de modèles thématiques. Les approches probabilistes sont représentées par le LDA statique et le Dynamic Topic Model (DTM), tandis que les approches neuronales sont étudiées à travers BERTopic et son extension dynamique.

L'objectif n'est pas uniquement de comparer les performances techniques des modèles, mais également d'évaluer leur capacité à produire des résultats interprétables et utiles dans l'analyse des discours médiatiques liés aux catastrophes naturelles et aux risques environnementaux.

En confrontant ces différentes approches sur un même corpus journalistique francophone, ce travail cherche ainsi à mieux comprendre les forces, les limites et les complémentarités des modèles thématiques contemporains dans l'étude de la sensibilisation médiatique aux risques naturels et climatiques.

3 Méthodologie

3.1 Présentation du corpus

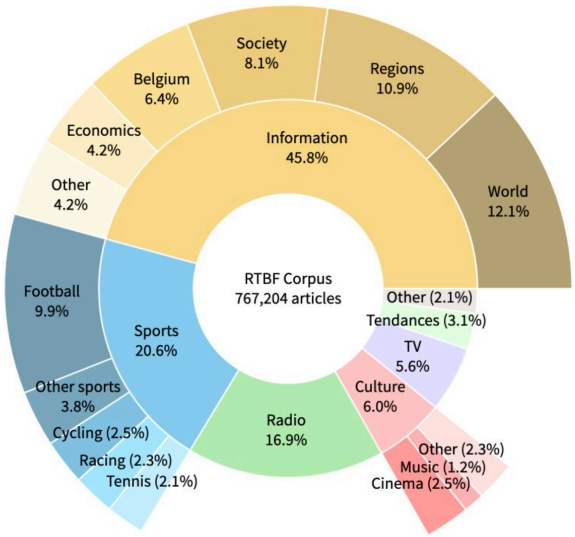


Figure 3.1: Répartition temporelle des articles du corpus RTBF (Escouflaire et al., 2023).

Ce mémoire s'appuie sur le corpus RTBF décrit par Escouflaire et al. (2023), qui regroupe 767,204 articles journalistiques publiés entre 2008 et 2021[1] par le service d'information de la Radio-Télévision Belge Francophone (RTBF), pour un total de plus de 214 millions de mots. Ce corpus couvre un spectre éditorial très large — politique, économie, culture, sport, société et constitue de ce fait une base particulièrement riche pour l'analyse thématique à long terme. Le fichier source, disponible sur la plateforme Dataverse de l'Union européenne au format XLSX, est réparti en cinq tranches annuelles dont la distribution temporelle est présentée à la Figure 3.1. Chaque article est décrit par les colonnes suivantes :

- **ID**: id interne de l'article qui sont des nombres entre 3 et 8 généres par le media source ;
- **title** : le titre de l'article ;
- **pub_date** : la date de publication ;
- **signature** : nom du journaliste qui avait écrit l'article, la radio ou le programme TV dont il est dérivé ;
- **feed** : section du site de la RTBF où l'article a été publié ;
- **category** : classification relative au topic traité par l'article ;
- **keyword** : mot ou phrase attribué à un article par un journaliste ou la source qui l'avait écrite ;

[1] convient de noter que les données de l'année 2021 sont partielles ; les articles concernant les inondations de juillet 2021 en Belgique ne sont pas inclus dans ce corpus.

- **text_HTLM** : texte directement reçu du site de la RTBF, qui contient encore les balises HTML ;
- **text_cleaned** : texte après suppression des balises HTML et correction des anomalies d'encodage ; la signature de l'article et les liens vers d'autres contenus du site ont été ajoutés en fin de texte ;
- **text_preprocessed** : version du texte obtenue après un prétraitement approfondi — suppression des signatures, des liens, remplacement des nombres par le token `NUM` et des URL par le token `URL`. Les auteurs conseillent d'utiliser cette version avec prudence, car certaines de ces transformations peuvent nuire à des analyses spécifiques.

Pour notre analyse, nous allons utiliser le texte contenu dans **text_preprocessed** celui-ci étant la plus adapté dans notre cas.

3.2 Chargement des données

La version du corpus utilisée est celle disponible sur le portail Dataverse de l'Union européenne, répartie en cinq fichiers Excel distincts. Chaque tranche est chargée séparément, puis les cinqfichiers sont concaténés pour former un corpus unique de 767,204 articles.

3.3 Tri du corpus

La méthode de tri présentée ci-dessous résulte d'un processus itératif progressivement affiné sur l'ensemble du corpus. L'objectif était de constituer une base documentaire cohérente, centrée sur les risques naturels et climatiques, tout en limitant la présence d'articles hors sujet ou trop éloignés de la problématique étudiée.

Dans un premier temps, un dictionnaire thématique a été élaboré à partir de listes de mots-clés associés aux différentes catégories d'aléas étudiées. Ces listes comprenaient notamment des termes liés aux inondations, aux tempêtes, aux sécheresses, aux incendies ou encore aux phénomènes climatiques extrêmes. L'application de ce dictionnaire au corpus complet a permis de réaliser une première sélection automatique des articles potentiellement pertinents.

Dans un second temps, plusieurs phases d'évaluation qualitative par échantillonnage ont été réalisées. Des sous-ensembles d'articles ont été examinés manuellement afin de vérifier la pertinence réelle des textes sélectionnés. Cette étape a permis d'identifier certains termes trop génériques produisant du bruit, mais également d'ajouter des expressions plus spécifiques qui n'étaient pas présentes dans les premières versions du dictionnaire.

Le processus de tri a ensuite été complété par une phase de modélisation thématique exploratoire à l'aide du modèle LDA. L'analyse des termes les plus représentatifs de chaque topic a permis d'identifier plus précisément les catégories discursives réellement captées par le corpus filtré. Cette étape a joué un rôle important dans l'ajustement final des paramètres de sélection et dans la validation de la cohérence thématique globale du corpus.

Cette démarche progressive de filtrage, de validation manuelle et d'exploration thématique a permis de construire un corpus plus homogène et mieux adapté à l'analyse des discours médiatiques relatifs aux risques naturels et environnementaux.

3.3.1 Tri de niveau 1

Le tri de niveau 1 consiste à filtrer le corpus sur la variable *feed*, qui comprend sept catégories~: *RTBF-SPORT*, *RTBFINFO*, *RTBFCULTURE*, *RADIO*, *TV*, *OTHER* et *RTBFTENDANCE*. Seule la catégorie **RTBFINFO** est conservée ; les six autres sont écartées en raison de leur contenu hors périmètre thématique (sport, culture, divertissement et tendances).

Si certains articles issus de ces catégories peuvent contenir une mention incidente d'un phénomène naturel — une rencontre sportive perturbée par des intempéries, par exemple — ces occurrences restent contextuelles et ne constituent pas un traitement éditorial de la catastrophe. Dans le cadre d'une modélisation de thématiques consacrée aux risques naturels, leur inclusion introduirait du bruit thématique sans valeur analytique. Seule la catégorie **RTBFINFO** garantit un traitement journalistique direct et substantiel des événements, comme l'illustrent les exemples présentés dans le tableau Table 6.3 en annexe.

Cette première sélection porte le corpus à **351,029 articles**, soit **45,8 %** des 767,204 articles initiaux.

3.3.2 Tri de niveau 2

Pour cette étape, nous nous appuyons sur les classifications internationales des aléas naturels proposées par les organismes de référence (Nations Unies (2016) ; United Nations Office for Disaster Risk Reduction (s.d.) ; Groupe d'experts intergouvernemental sur l'évolution du climat (2014) ; Office québécois de la langue française (s.d.)).

Les événements climatiques extrêmes sont organisés selon une structure hiérarchique à deux niveaux : (1) cinq sous-catégories thématiques (inondation, tempête, intempéries, sécheresse, vague de chaleur) et (2) deux grandes familles conceptuelles — **chaleur**, regroupant sécheresse et vagues de chaleur, et **intempéries**, regroupant inondations, tempêtes et intempéries diverses.

Cette organisation répond à un double objectif. D'une part, elle suit la recommandation en sciences de l'environnement de regrouper les aléas selon leurs mécanismes physiques dominants (IPCC 2021)~: les phénomènes liés à la chaleur découlent d'anomalies thermiques prolongées, tandis que les phénomènes d'intempéries relèvent de dynamiques atmosphériques rapides ou hydrométéorologiques. D'autre part, elle optimise l'analyse textuelle en combinant une catégorisation fine — utile pour la détection spécifique des aléas — et une simplification analytique — utile pour les comparaisons globales et la modélisation. Ce type d'approche hiérarchique améliore la robustesse de l'assignation et réduit le risque de dilution sémantique entre catégories proches (Sebastiani (2002)).

Afin de limiter les faux positifs, seuls les termes qui ne peuvent pas être utilisés de manière métaphorique ou dans un autre sens sont conservés dans le dictionnaire. Des termes comme *chaleur*, *sécheresse*, *ouragan*, *tempête* ou *crues* sont volontairement exclus en raison de leur polysémie potentielle. Le tableau Table 3.1 présente la liste définitive des termes retenus par catégorie d'aléa.

Le processus de classification repose sur une approche lexicale combinée à un algorithme de recherche multi-motifs. Le texte de chaque article est d'abord construit par concaténation du titre et du contenu prétraité, puis soumis à une normalisation linguistique stricte : décomposition Unicode de type NFKD pour éliminer les diacritiques, passage en minuscules, suppression des caractères non alphanumériques et réduction des espaces multiples. Les termes du dictionnaire sont normalisés selon le même pipeline. Des variantes morphologiques simples — pluriels, formes concaténées — sont générées pour chaque terme afin de couvrir les variations lexicales sans recourir à une lemmatisation complète.

Table 3.1: Termes clés retenus par catégorie d'aléa

Catégorie	Termes retenus
Inondation	inondation(s), crue subite, crues subites, inondations éclair, inondations côtières, pluies torrentielles, dégâts des eaux, montée des eaux, débordement d'un cours d'eau, rivière(s) en crue, submersion marine, pluies diluviennes, zones inondables, élévation du niveau de la mer
Vague de chaleur	vague(s) de chaleur, canicule(s), chaleur extrême, température(s) extrême(s), record(s) de chaleur, dôme(s) de chaleur, stress thermique, anomalie thermique, épisode(s) caniculaire(s), îlot de chaleur urbain
Tempête	orage, cyclone(s), typhon(s), rafales de vents, vents violents, tempête(s) tropicale(s), bourrasques violentes, perturbation atmosphérique, ouragan de catégorie, onde(s) de submersion, onde(s) de tempête
Intempéries	intempéries, conditions météorologiques extrêmes, précipitations intenses/extrêmes, fortes pluies, neige abondante, orages violents, grêle (violente, intense, destructrice), grêlons, chutes de grêle, supercellule orageuse, pluie verglaçante
Sécheresse	sécheresse(s), déficit(s) hydrique(s), stress hydrique, manque d'eau, pénurie d'eau, canicule prolongée, sécheresse des sols / météorologique / hydrologique / agricole, feux de forêt, mégasécheresse, désertification, aridification

L'ensemble des motifs est ensuite intégré dans un automate d'Aho et Corasick (1975), permettant la détection simultanée de multiples chaînes en un seul parcours linéaire du texte. Cette structure garantit une complexité temporelle en $O(\ell)$, indépendamment du nombre de motifs, où ℓ désigne la longueur du document. La stratégie retenue implémente une **classification multi-étiquettes** (multi-label) : un même article peut être associé à plusieurs catégories d'aléas lorsqu'il en mentionne plusieurs simultanément. Les documents ne contenant aucune occurrence pertinente sont exclus des analyses ultérieures.

3.4 Préparation du corpus après tri pour le Topic Modeling (Modèles Probabilistes).

3.4.1 Nettoyage et normalisation du texte

Pour garantir la qualité des thématiques extraites, nous appliquons un prétraitement rigoureux via la bibliothèque **clean-text**. Cette étape permet :

- de corriger les anomalies d'encodage (par exemple, la transformation des entités HTML ou des erreurs UTF-8 de type *Ã©* en *é*) ;
- de normaliser la casse en passant l'ensemble du corpus en minuscules ;
- de préserver les accents et la ponctuation essentielle tout en supprimant les bruits de structure tels que les sauts de ligne.

À la suite de ce nettoyage, nous utilisons des expressions régulières pour remplacer tous les jours de la semaine par le jeton générique **week**. Cette généralisation s'impose car la fréquence élevée des noms de jours, observée lors de nos tests préliminaires, parasitait la formation des clusters thématiques. En les

regroupant sous un concept unique, nous permettons au modèle de se concentrer sur des termes à plus forte valeur sémantique.

3.4.2 Tokenisation, lemmatisation et filtrage

Le texte est ensuite segmenté en tokens, débarrassé des mots vides (*stopwords*) et soumis à une lemmatisation. Cette étape réduit les mots à leur forme canonique, ce qui facilite le regroupement des termes appartenant à une même famille sémantique.

3.4.2.1 Extraction des n-grammes (bi- et trigrammes)

Afin de capturer des concepts complexes qui perdraient leur sens s'ils étaient décomposés — tels que « réseaux sociaux » ou « changement climatique » —, nous procédons à la formation de n-grammes en deux étapes.

Identification par le score PMI. Nous utilisons le score de *Pointwise Mutual Information* (PMI) pour détecter les associations de mots statistiquement significatives. Ce score mesure dans quelle mesure deux mots w_i et w_j cooccurrent plus souvent qu'attendu sous l'hypothèse d'indépendance~:

$$\mathrm{PMI}(w_i, w_j) = \log \frac{P(w_i, w_j)}{P(w_i)\,P(w_j)} \tag{3.1}$$

où $P(w_i, w_j)$ est la probabilité de cooccurrence dans une fenêtre contextuelle, et $P(w_i)$, $P(w_j)$ leurs probabilités marginales respectives. Un PMI élevé signale une association forte et non aléatoire~: le bigramme « changement climatique » obtient un score élevé car ces deux mots apparaissent ensemble bien plus souvent que leur fréquence individuelle ne le laisserait prévoir. En pratique, un seuil minimal $\mathrm{PMI}_{\min}$ est fixé au-dessus duquel une paire est retenue comme n-gramme candidat. Les trigrammes sont construits par application itérative du même score sur les bigrammes déjà identifiés.

Filtrage syntaxique par annotation morpho-syntaxique (*POS tagging*). L'annotation morpho-syntaxique attribue à chaque token son rôle grammatical~: nom (`NOUN`), verbe (`VERB`), adjectif (`ADJ`), déterminant (`DET`), etc. Nous utilisons le modèle `fr_core_news_md` de la bibliothèque spaCy, entraîné sur des corpus journalistiques français. Pour chaque n-gramme candidat identifié par le PMI, nous vérifions que sa structure grammaticale correspond à l'un des schémas nominaux valides suivants~:

$$\mathrm{NOUN}, \quad \mathrm{ADJ} + \mathrm{NOUN}, \quad \mathrm{NOUN} + \mathrm{NOUN}, \quad \mathrm{NOUN} + \mathrm{PREP} + \mathrm{NOUN} \tag{3.2}$$

Les structures verbales, adverbiales ou mixtes sont exclues. Ce filtre garantit que les n-grammes conservés constituent des descripteurs lexicaux stables et directement exploitables dans la matrice document-terme X.

3.4.3 Vectorisation

Cette étape transforme le texte filtré en données mathématiques exploitables par les algorithmes de modélisation. Elle se déroule en deux temps :

- **Construction du dictionnaire** : chaque mot unique du corpus filtré reçoit un identifiant numérique ;

- **Génération du sac de mots** (*Bag of Words*) : le texte est converti en un vecteur de tuples (identifiant du mot, fréquence), où la fréquence indique le nombre d'occurrences du mot dans le document.

Cette représentation permet de passer d'un format textuel à une matrice de fréquences, en faisant abstraction de l'ordre des mots pour se concentrer sur leur importance statistique.

3.5 Optimisation du modèle et Analyse de sensibilité

3.5.1 Model LDA

3.5.1.1 Optimisation des Hyperparamètres et Validation Croisée

Le choix du nombre optimal de topics K ne peut être dissocié des autres paramètres structurels du modèle LDA. Afin de garantir la stabilité et la convergence de l'algorithme, nous mettons en œuvre une recherche par grille (*grid search*) sur les hyperparamètres suivants.

Le paramètre de Dirichlet α contrôle la distribution des thèmes par document, c'est-à-dire la concentration de $\theta_d \sim \text{Dir}(\alpha)$. Deux modes de spécification sont possibles.

Dans le mode **symétrique**, le même hyperparamètre est appliqué à tous les thèmes sans distinction :

$$\alpha = (\alpha, \alpha, \dots, \alpha) \in \mathbb{R}^K \tag{3.3}$$

Le modèle ne fait a priori aucune hypothèse sur l'importance relative des thèmes. C'est la configuration par défaut de la plupart des implémentations (Gensim, Mallet).

Dans le mode **asymétrique**, chaque thème k possède son propre hyperparamètre α_k, appris à partir des données :

$$\alpha = (\alpha_1, \alpha_2, \dots, \alpha_K) \in \mathbb{R}^K, \quad \alpha_k \neq \alpha_{k'} \text{ en général} \tag{3.4}$$

Certains thèmes reçoivent alors a priori une probabilité d'apparition plus forte que d'autres, ce qui reflète une réalité structurelle des corpus journalistiques : des sujets comme la météo ou la politique y sont chroniquement surreprésentés par rapport à des événements rares tels que les tsunamis.

Nous testons les trois configurations présentées dans le tableau ci-dessous :

Table 3.2: Configurations du paramètre α testées dans l'analyse de sensibilité

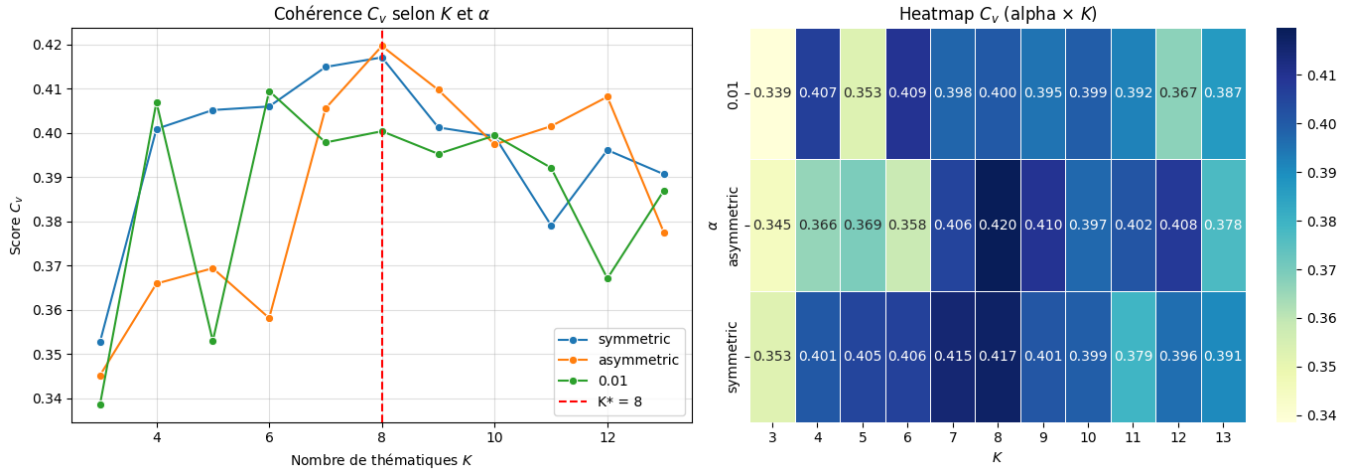
Mode	Vecteur α	Effet attendu sur notre corpus
`symmetric`	$(\alpha, \dots, \alpha)$	Tous les thèmes traités équitablement
`asymmetric`	$(\alpha_1, \dots, \alpha_K)$ appris	S'adapte aux thèmes dominants du corpus
`0.01 (fixe)`	$(0{,}01, \dots, 0{,}01)$	Force des documents très spécialisés (1–2 thèmes max)

Le nombre de passes (*passes*) définit le nombre de fois que l'algorithme parcourt l'intégralité du corpus. Une valeur plus élevée stabilise les probabilités $P(z \mid d)$ dans les corpus médiatiques denses. **Le**

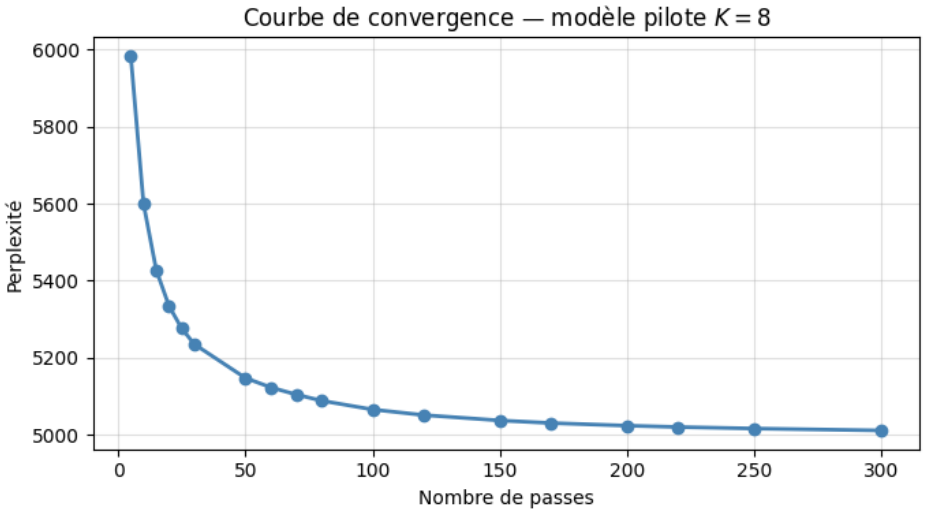
nombre d'itérations (*iterations*) contrôle la fréquence maximale des boucles de Gibbs ou de l'inférence variationnelle pour chaque document. La grille de recherche complète est présentée dans le Table 3.3.

Le nombre de passes (200) a été déterminé empiriquement à partir d'un modèle pilote ($K = 8$, α symétrique par défaut de Gensim) : la Figure 3.2b montre que la perplexité se stabilise aux alentours de 200 passes, seuil retenu comme critère d'arrêt. Dans le cadre de l'entraînement LDA, la perplexité est calculée à chaque passe sur un ensemble de documents : tant qu'elle décroît significativement, le modèle continue d'affiner ses distributions θ_d et β_k ; lorsqu'elle se stabilise, les passes supplémentaires n'apportent plus de gain informatif. Toutefois, la perplexité tend à favoriser des modèles avec un grand nombre de topics sans garantir leur interprétabilité ; elle est donc utilisée ici exclusivement comme critère d'arrêt, et non comme critère de sélection de K, rôle dévolu à la cohérence C_v (Roder, Both, et Hinneburg (2015)).

Le nombre d'itérations est fixé à 1,000, valeur délibérément conservative (50 par défaut dans Gensim). En pratique, l'inférence variationnelle s'interrompt dès que le critère de convergence interne ($\gamma_{\text{threshold}} = 0{,}001$) est atteint, ce qui survient généralement bien avant ce plafond sur des corpus journalistiques (Blei, Ng, et Jordan (2003)).



(a) Analyse alpha



(b) Analyse du nombre de passes

Figure 3.2: Optimisation des Hyperparamètres (LDA)

Table 3.3: Grille de recherche des hyperparamètres du modèle LDA

Paramètre	Grille de recherche	Rôle mathématique
K **(Topics)**	3 à 14	Dimensionnalité de l'espace latent
α	`symmetric`, `asymmetric`, 0,01	Prior de la distribution $\theta_d \sim \text{Dir}(\alpha)$
Passes	200	Époques d'apprentissage global sur le corpus
Itérations	1 000 (fixes)	Convergence locale par inférence variationnelle

L'analyse de la grille d'hyperparamètres aboutit à la conclusion suivante : le score C_v le plus faible est systématiquement obtenu avec $\alpha = 0{,}01$ (Figure 3.2a), tandis que la configuration asymétrique donne les meilleurs résultats. Ce constat suggère que les articles du corpus ne traitent pas les thématiques de manière uniforme, ce qui rend l'α asymétrique particulièrement adapté à notre corpus journalistique.

Le modèle final retenu présente les caractéristiques consignées dans le tableau suivant :

Paramètre	Valeur optimale
Prior Alpha (α)	Asymétrique
Nombre de passes	200
Nombre d'itérations	1 000
Nombre de topics (K)	8
Score de cohérence (C_v)	**0,4196**

Table 3.4: Configuration finale du modèle LDA

3.5.1.2 Analyse de sensibilité et robustesse du modèle

3.5.1.2.1 Stabilité des topics

Afin d'éprouver la robustesse de notre modélisation, nous mettons en œuvre une procédure de sous-échantillonnage aléatoire proche de la logique du bootstrap (Efron (1979)). Nous générons 100 sous-échantillons aléatoires par tirages sans remise. Pour chaque échantillon, la graine aléatoire est fixée à 123 et les hyperparamètres optimaux précédemment identifiés sont conservés, tandis que le nombre de topics K varie de 5 à 12.

À des fins d'efficience computationnelle, le nombre de passes est réduit à 10 et le nombre d'itérations à 50. Ce calibrage allégé est délibéré : conformément aux recommandations de Greene, O'Callaghan, et Cunningham (2014), l'objectif n'est pas d'obtenir des modèles pleinement convergés, mais de vérifier si une organisation thématique cohérente émerge de manière récurrente, indépendamment du sous-ensemble de documents considéré (Jacobi, Atteveldt, et Welbers (2016)).

Cette procédure poursuit deux objectifs : évaluer la stabilité du nombre optimal de topics à travers les réplicats ; et analyser l'évolution des scores de cohérence via le calcul d'intervalles de confiance, afin de confirmer si la structure à $K = 8$ topics est représentative de l'ensemble du corpus.

3.5.1.2.2 Stabilité sémantique

Le processus d'inférence du LDA est intrinsèquement stochastique : lors de l'initialisation, chaque mot est assigné aléatoirement à un thème ; à chaque itération, l'algorithme tire au sort de nouvelles affectations selon les distributions de probabilité estimées. Il convient donc de s'assurer que les thèmes extraits ne sont pas un artefact de l'initialisation, mais reflètent une structure latente stable du corpus. Pour ce faire, nous appliquons le modèle optimal (alpha asymétrique, 200 passes, 1,000 itérations, $K = 8$) avec cinq graines aléatoires distinctes : $S \in \{0, 21, 42, 84, 123\}$.

Choix de la mesure de stabilité. Deux familles de métriques sont classiquement mobilisées (Greene, O'Callaghan, et Cunningham (2014)) :

- Le **Rand Index** (Rand (1971)) et sa version ajustée (ARI, Hubert et Arabie (1985)) mesurent l'accord entre deux partitions de documents : ils répondent à la question « les mêmes documents sont-ils regroupés ensemble dans les deux modèles ? » ;
- L'**indice de Jaccard** mesure le recouvrement lexical entre deux ensembles de mots : il répond à la question « les mêmes mots définissent-ils les mêmes topics ? ».

Dans notre cas, l'objet de la comparaison est la **composition lexicale** des topics, et non la répartition des documents entre clusters. Le Rand Index est donc écarté au profit de l'indice de Jaccard, conformément aux recommandations de Greene, O'Callaghan, et Cunningham (2014).

Nous mesurons ainsi la similarité entre modèles via l'**indice de Jaccard** (Jaccard (1901)), qui quantifie le recouvrement lexical entre deux thématiques A et B représentées par leurs N mots les plus probables ($N = 50$) :

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

où $|A \cap B|$ représente l'intersection (mots communs) et $|A \cup B|$ l'union (total des mots uniques) des deux ensembles. Un indice proche de 1 indique une stabilité parfaite des thématiques malgré la modification du point de départ aléatoire.

Afin de quantifier la robustesse globale du modèle, nous calculons l'indice de Jaccard moyen ($\bar{J}$) sur l'ensemble des paires de graines aléatoires P. Pour chaque paire de modèles (s_i, s_j), nous identifions la meilleure correspondance sémantique en maximisant la similarité entre chaque thème k du premier modèle et l'ensemble des thèmes k' du second :

$$\bar{J} = \frac{1}{|P|} \sum_{(s_i,\, s_j) \in P} \frac{1}{K} \sum_{k=1}^{K} J\Big(T_k^{(s_i)},\, T_k^{(s_j)}\Big)$$

où :

- $|P|$ est le nombre total de paires de *seeds* testées ;
- K est le nombre de thèmes (ici $K = 8$) ;
- $J(T_k)$ est l'indice de Jaccard calculé sur les $N = 50$ mots les plus probables d'un topic sur deux graines aléatoires différentes (s_i, s_j).

Pour l'interprétation des résultats, nous adoptons les seuils empiriques de Greene, O'Callaghan, et Cunningham (2014) présentés dans le tableau ci-dessous :

Table 3.5: Seuils d'interprétation de l'indice de Jaccard moyen

Score $\bar{J}$	Interprétation
$\bar{J} > 0{,}6$	**Forte stabilité** : les thèmes sont structurels et indépendants de la graine aléatoire.
$0{,}4 < \bar{J} \leq 0{,}6$	**Stabilité modérée** : la structure globale est présente mais sensible aux conditions d'initialisation.
$\bar{J} \leq 0{,}4$	**Instabilité** : les thèmes sont considérés comme des artefacts de l'initialisation aléatoire.

3.5.2 Model DTM(Dynamic topic Model)

3.5.2.1 Optimisation des Hyperparamètres et Validation Croisée

Dans le cadre du Dynamic Topic Model (DTM), nous avons choisi de conserver le même nombre de thématiques que celui retenu pour le modèle LDA statique, soit $K = 8$.

Ce choix repose d'abord sur l'hypothèse structurelle propre au DTM, selon laquelle les thèmes persistent au cours du temps tout en évoluant progressivement sur le plan lexical. Le modèle cherche ainsi à suivre la trajectoire des thématiques à travers les différentes périodes étudiées, sans supposer leur apparition ou leur disparition brutale. Le maintien d'un nombre fixe de topics favorise donc une lecture continue des dynamiques discursives.

Cette décision répond également à un objectif de cohérence méthodologique entre les analyses statiques et temporelles. En conservant la même structure thématique dans le LDA et dans le DTM, il devient possible de comparer directement les résultats obtenus et d'interpréter les évolutions observées à partir d'une base commune. Une modification du nombre de topics aurait introduit une rupture de segmentation rendant les comparaisons beaucoup plus difficiles.

Par ailleurs, l'objectif principal de cette analyse n'est pas de redéfinir la structure globale du corpus à chaque période, mais plutôt d'étudier la manière dont les thèmes évoluent dans le temps. En maintenant une structure fixe à huit thématiques, nous isolons davantage la dimension temporelle et pouvons observer plus clairement les phénomènes de dérive sémantique, de recomposition lexicale ou de variation de prépondérance des thèmes au fil des années.

Afin d'assurer une continuité méthodologique avec le modèle LDA, le prétraitement textuel a été appliqué de manière identique dans le cadre du DTM. Les termes transversaux précédemment identifiés comme peu informatifs (« nbr », « autre », « week », « nouveau ») ont ainsi été supprimés, de même que les mots apparaissant dans moins de cinq documents. Cette étape vise à concentrer l'analyse sur les éléments lexicaux réellement structurants pour les thématiques étudiées.

Une adaptation technique a toutefois été nécessaire concernant le filtrage des termes très fréquents. Dans Gensim, cette opération repose sur la fonction `filter_extremes` et sur le paramètre proportionnel

`no_above`, fixé ici à 50 %. La bibliothèque **tomotopy**, utilisée pour le DTM, ne propose pas d'équivalent direct à ce mécanisme. Nous avons donc utilisé le paramètre `rm_top`, qui supprime un nombre fixe de termes parmi les plus fréquents du corpus.

Afin d'identifier la valeur la plus adaptée, une recherche par grille (*grid search*) a été réalisée sur les valeurs suivantes :

$$\{10,\ 20,\ 30,\ 40,\ 50,\ 60,\ 80,\ 100\}$$

Chaque configuration a ensuite été évaluée à partir d'un score de cohérence thématique.

Dans cette phase d'optimisation, nous avons privilégié la métrique de cohérence $C_{U_{\text{mass}}}$ plutôt que le score C_v.

Ce choix s'explique d'abord par la nature temporelle du DTM. Contrairement au C_v, qui repose sur des fenêtres glissantes et des similarités cosinus particulièrement adaptées aux modèles statiques, le $C_{U_{\text{mass}}}$ s'appuie directement sur les cooccurrences internes observées dans le corpus d'apprentissage. Cette propriété le rend plus cohérent avec la logique probabiliste et séquentielle du DTM.

Le $C_{U_{\text{mass}}}$ présente également un avantage computationnel important. Son calcul exploite directement les probabilités estimées par le modèle, ce qui réduit fortement le coût algorithmique de l'évaluation par rapport au C_v.

Enfin, plusieurs travaux, notamment ceux de Stevens, Anderson, et Liberman (2012) et Mimno et al. (2011), montrent que cette métrique est particulièrement robuste pour identifier des thèmes statistiquement cohérents au sein d'un corpus donné. À l'inverse, le score C_v, davantage influencé par des proximités sémantiques perçues, peut parfois produire des évaluations moins stables dans des contextes fortement évolutifs.

L'optimisation du paramètre `rm_top` a finalement été réalisée en fixant arbitrairement le nombre de périodes temporelles à $T = 8$. Cette stratégie permettait d'isoler l'effet du filtrage lexical sans introduire de variations supplémentaires liées au découpage chronologique du corpus. Les différences observées dans les scores de cohérence pouvaient ainsi être attribuées principalement à la qualité du nettoyage lexical et non à des modifications de la structure temporelle du modèle.

L'optimisation du paramètre `rm_top` est réalisée en fixant arbitrairement le découpage temporel à $T = 8$, ce qui permet d'isoler l'effet du filtrage lexical sans le confondre avec celui de la segmentation chronologique. Comme l'illustre la Figure 3.3, le score de cohérence est maximisé pour un retrait des **dix termes les plus fréquents** ($n = 10$).

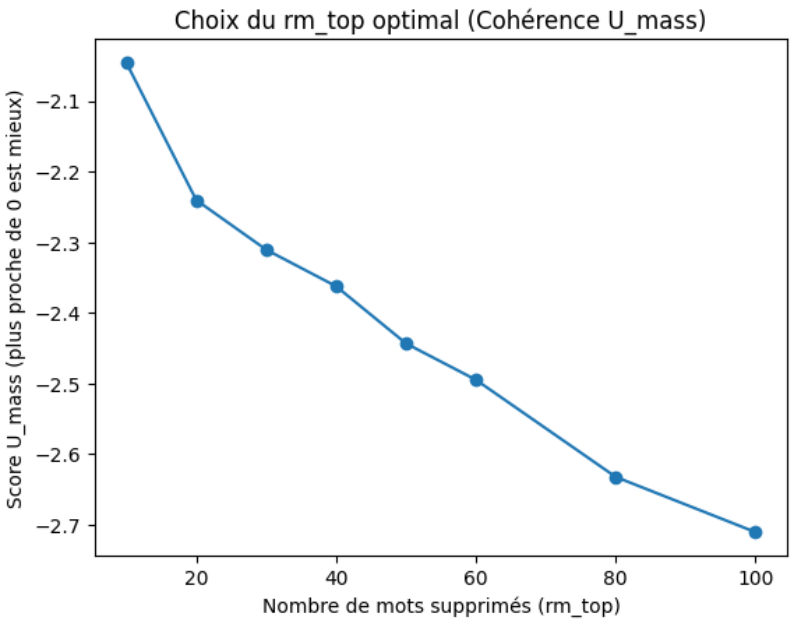


Figure 3.3: Évolution du score $C_{U_{\text{mass}}}$ selon le nombre de mots supprimés (rm_top)

3.6 Préparation du corpus pour le Topic Modeling neuronal

Contrairement aux approches probabilistes, les modèles neuronaux de *topic modelling* ne se fondent pas sur les cooccurrences de mots : ils s'appuient sur des représentations vectorielles denses — les *embeddings* — capables de capturer des relations sémantiques inaccessibles aux modèles classiques. Deux documents peuvent ainsi traiter du même sujet sans partager un seul mot : un modèle neuronal les rapprochera là où le LDA les ignorerait.

Les analyses menées avec le LDA et le DTM ont par ailleurs permis d'identifier des termes transversaux peu discriminants (*nbr*, *week*, *autre*, *nouveau*) qui brouillent la séparation sémantique des clusters ; ces termes sont retirés dès la phase de prétraitement. Ce qui change fondamentalement par rapport aux modèles précédents, c'est la philosophie du nettoyage : plutôt que de réduire agressivement les mots par lemmatisation, nous cherchons à préserver leurs formes naturelles. Les Transformers exploitent le contexte lexical complet, et une lemmatisation trop agressive appauvrit précisément l'information que le modèle est conçu à exploiter.

3.6.1 Nettoyage et normalisation du corpus

Le prétraitement textuel appliqué aux modèles neuronaux reprend les principales étapes décrites pour les modèles probabilistes, notamment la correction des anomalies d'encodage, la normalisation des caractères, la conversion en minuscules, la suppression des retours à la ligne ainsi que l'élimination des éléments non textuels, des mots vides génériques et des termes identifiés comme bruit lexical. Aucune lemmatisation agressive n'est toutefois appliquée dans cette phase afin de préserver les formes naturelles des mots et la richesse contextuelle exploitée par les architectures Transformer.

3.6.2 Génération des représentations vectorielles

Chaque document est transformé en un vecteur dense à l'aide du modèle **Sentence-BERT (SBERT)** dans sa version multilingue adaptée au français. Pour un document d, la représentation vectorielle s'écrit :

$$\mathbf{e}_d = f_\theta(d) \in \mathbb{R}^{384}$$

où f_θ désigne le modèle Transformer pré-entraîné utilisé pour générer les embeddings, et où la dimension 384 correspond à la taille de l'espace vectoriel latent produit par Sentence-BERT.

Appliqué à notre corpus de 8,958 articles, ce processus produit une matrice d'embeddings :

$$E \in \mathbb{R}^{8958 \times 384}$$

dans laquelle chaque ligne représente la projection vectorielle d'un article dans l'espace sémantique appris par le modèle.

3.6.3 Optimisation des hyperparamètres par Optuna

Le pipeline BERTopic implique une dizaine d'hyperparamètres répartis entre UMAP, HDBSCAN et le vectoriseur. Leur interaction rend une recherche exhaustive impossible : les choix pour UMAP contraignent directement les distances dans lesquelles HDBSCAN opère. Nous recourons à **Optuna** (Akiba et al. (2019)) avec le sampler TPE (*Tree-structured Parzen Estimator*) en mode multivarié, ce qui permet de capturer les dépendances entre paramètres plutôt que de les traiter indépendamment.

3.6.3.1 Espace de recherche

Table 3.6: Espace de paramètres testés

Composant	Paramètre	Plage
UMAP	`n_neighbors`	$[5, 50]$
UMAP	`n_components`	$[5, 20]$
UMAP	`min_dist`	$[0.0, 0.3]$
UMAP	`metric`	cosine, euclidean
HDBSCAN	`min_cluster_size`	$[20, 80]$
HDBSCAN	`min_samples`	$\lfloor \text{min\_cluster\_size}/2 \rfloor$
HDBSCAN	`cluster_selection_epsilon`	$[0.0, 0.15]$
HDBSCAN	`cluster_selection_method`	eom, leaf
Vectoriseur	`ngram_max`	$\{1, 2, 3\}$
Vectoriseur	`min_df`	$[2, 8]$
Vectoriseur	`max_df`	$[0.6, 0.95]$

Le paramètre `min_samples` est contraint dynamiquement à ne pas dépasser $\lfloor \text{min\_cluster\_size}/2 \rfloor$, car une valeur supérieure produirait systématiquement trop d'outliers.

3.6.3.2 Fonction objectif

La qualité d'un modèle est évaluée par un score composite~:

$$\mathcal{S} = C_v \cdot (1 - r_{\text{out}}) \cdot D(K) \cdot P(K) \cdot \Delta(K) \tag{3.5}$$

où~:

- $C_v \in [0,1]$ est la cohérence sémantique des topics, mesurée sur les documents non-outliers ;
- $r_{\text{out}} \in [0,1]$ est le taux d'outliers ;
- $D(K) = \frac{\log(1+\min(K,40))}{\log(41)}$ est un bonus de diversité qui récompense un nombre raisonnable de topics~;
- $P(K) = \begin{cases} 1{,}0 & \text{si } 10 \leq K \leq 30 \\ 0{,}6 & \text{sinon} \end{cases}$ pénalise les solutions avec trop peu ou trop de topics ;
- $\Delta(K) = \begin{cases} 1{,}0 & \text{si } p_{\max} < 0{,}30 \\ 0{,}5 & \text{si } p_{\max} < 0{,}50 \\ 0{,}1 & \text{sinon} \end{cases}$ est la pénalité de dominance, où $p_{\max}$ est la part du corpus assignée au topic le plus volumineux.

Cette dernière pénalité répond à un problème observé : sans elle, HDBSCAN peut créer un cluster dominant absorbant la majorité des documents, signe que les paramètres permettent une fusion excessive des clusters.

3.6.4 Sélection de la stratégie pour les outliers.

Trois stratégies sont évaluées en parallèle sur le modèle final~: par embeddings, par c-TF-IDF, et par distributions de probabilités (voir tableau ci-dessous). La stratégie retenue est celle qui minimise le nombre de documents restant labellisés -1 après réassignation, sous réserve que la cohérence C_v ne se dégrade pas de manière significative. Le modèle mis à jour via `update_topics()` constitue la version finale utilisée pour l'analyse thématique.

Stratégie	Principe	Avantage	Limite
Embeddings	Proximité cosinus au centroïde du cluster dans l'espace vectoriel	Sémantiquement fidèle aux représentations SBERT	Sensible aux distorsions géométriques introduites par UMAP
c-TF-IDF	Similarité cosinus sur la représentation lexicale des topics	Indépendante de la géométrie UMAP	Ignore le contexte sémantique profond des documents
Distributions	Probabilité d'appartenance native produite par HDBSCAN	Exploite l'incertitude native du modèle	Nécessite `calculate_probabilities=True` à l'entraînement

Table 3.7: Comparaison des trois stratégies de réassignation des outliers

Le modèle mis à jour via `update_topics()` constitue la version finale utilisée pour l'analyse thématique présentée dans la section suivante.

3.7 Analyse de sensibilité: granularité temporelle.

3.7.1 Modèle DTM

L'objectif de cette section est de déterminer le découpage temporel le plus pertinent pour analyser l'évolution des thématiques du corpus, tout en évitant l'introduction de bruit statistique. Comme le soulignent Sjöbergh et Nanas (2012), un découpage inadéquat peut altérer significativement la qualité des topics estimés et leur interprétabilité. Le choix de la granularité temporelle constitue un paramètre critique : un découpage trop grossier (par exemple $T = 2$) risque de masquer des dynamiques importantes, tandis qu'un découpage trop fin (par exemple $T = 13$, correspondant au nombre total d'années du corpus) peut conduire à des estimations instables faute d'un nombre suffisant de documents par période.

Nous adoptons une stratégie d'analyse de sensibilité fondée sur la variation du nombre de périodes temporelles, en utilisant la bibliothèque **tomotopy** (Blei et Lafferty (2006)). Ce choix est motivé par son implémentation optimisée en Python, sa relative simplicité d'utilisation et l'absence de nécessité de recourir à une implémentation externe en C++. La stratégie repose sur quatre étapes.

1. Configuration du modèle et critères de convergence. La configuration du modèle DTM repose sur les paramètres présentés dans le tableau ci-dessous. Les seuils `min_cf` et `min_df` éliminent les mots trop rares ; le paramètre `rm_top` supprime les termes très fréquents mais peu informatifs. Le processus d'apprentissage repose sur un échantillonnage de Gibbs : la phase de *burn-in* correspond aux premières itérations durant lesquelles le modèle converge vers une distribution stationnaire et dont les résultats ne sont pas conservés ; les itérations suivantes (*train*) approximent les distributions a posteriori des variables latentes.

Table 3.8: Configuration du modèle DTM pour l'analyse de sensibilité temporelle

Paramètre	Valeur	Interprétation
K	8	Nombre de topics
T	3 à 13	Nombre de périodes temporelles testées
`min_cf`	5	Fréquence minimale dans le corpus
`min_df`	5	Fréquence minimale en nombre de documents
`rm_top`	10	Suppression des mots les plus fréquents
`seed`	123	Graine aléatoire (reproductibilité)
`burn_in`	200	Itérations de stabilisation (non conservées)
`train`	1 500	Itérations d'échantillonnage retenues

Afin d'évaluer la qualité d'ajustement, nous utilisons la log-vraisemblance par mot (`ll_per_word`), formellement définie par :

$$\ell = \frac{1}{N} \sum_{d=1}^{M} \log p(\mathbf{w}_d \mid \alpha, \eta)$$

où $N = \sum_{d=1}^{M} N_d$ est le nombre total de tokens dans le corpus. Dans tomotopy, cette quantité est estimée à chaque itération de l'échantillonnage de Gibbs (Griffiths et Steyvers (2004)) par la méthode de **Chib-style estimation** (Chib (1995)), en intégrant sur les variables latentes $\mathbf{Z}$ et Θ :

$$\log p(\mathbf{W}) = \log \int p(\mathbf{W} \mid \mathbf{Z}, \mathbf{B})\, p(\mathbf{Z} \mid \Theta)\, p(\Theta \mid \alpha)\, p(\mathbf{B} \mid \eta)\, \mathrm{d}\Theta\, \mathrm{d}\mathbf{B}$$

Une valeur moins négative de `ll_per_word` indique un meilleur ajustement. Cet indicateur est utilisé à deux niveaux~: vérifier la convergence pour chaque configuration temporelle, et ajuster le nombre d'itérations d'entraînement si nécessaire.

2. Extraction des distributions de mots. Pour chaque topic k et chaque période t, nous extrayons les cinquante mots les plus probables et normalisons la distribution pour garantir $\sum_w P(w \mid z = k, t) = 1$.

3. Mesure de la variation thématique. Pour quantifier l'évolution des topics entre périodes successives, nous utilisons la divergence de Jensen-Shannon (Lin (1991) ; Shannon (1948)) :

$$JS(P \parallel Q) = \frac{1}{2} D_{\mathrm{KL}}(P \parallel \bar{M}) + \frac{1}{2} D_{\mathrm{KL}}(Q \parallel \bar{M}), \quad \bar{M} = \tfrac{1}{2}(P + Q)$$

où D_{KL} désigne la divergence de Kullback-Leibler (Kullback et Leibler (1951)). Cette mesure présente trois propriétés particulièrement utiles : la symétrie ($JS(P \parallel Q) = JS(Q \parallel P)$), le bornage ($0 \leq JS \leq 1$) et la robustesse aux distributions nulles. Pour chaque topic k, nous calculons :

$$JS_{k,t} = JS\big(P(w \mid z = k, t),\ P(w \mid z = k, t+1)\big)$$

puis une mesure globale :

$$\overline{JS} = \frac{1}{K(T-1)} \sum_{k=1}^{K} \sum_{t=1}^{T-1} JS_{k,t}$$

4. Analyse de l'importance des topics. En parallèle, nous mesurons le poids de chaque topic dans chaque période :

$$\bar{P}(z = k \mid t) = \frac{1}{N_t} \sum_{d \in t} P(z = k \mid d)$$

où N_t est le nombre de documents dans la période t. Cette mesure permet d'identifier les thèmes dominants, les émergences et les déclins.

3.7.2 Modèle BERTopic dynamique

Le même découpage temporel T que celui établi pour le DTM est retenu pour le BERTopic dynamique. Ce choix repose sur trois arguments complémentaires. D'abord, il garantit une comparabilité directe entre les deux approches dynamiques : en analysant les mêmes périodes, les évolutions observées peuvent être confrontées sans biais lié au découpage. Ensuite, il évite de répéter une procédure déjà éprouvée. Enfin, il assure une cohérence narrative dans la présentation des résultats : un lecteur qui suit la progression des modèles retrouve les mêmes repères temporels tout au long de l'analyse.

3.8 Procédure de labellisation des thématiques.

La première approche envisagée reposait sur une annotation humaine réalisée par trois experts issus de domaines complémentaires :

- **Le Professeur Eugen Picalabelu**, promoteur de ce mémoire et professeur de statistique à la LSBA ;
- **Le Docteur Louis Escouflaire**, Postdoctoral Researcher at MIT, spécialiste en traitement automatique du langage naturel et sciences sociales computationnelles à l'UNamur ;
- **Le Docteur Damien Delforge**, Research Associate at UCLouvain, co-promoteur de ce mémoire.

Le protocole prévoyait de soumettre à ces experts les mots les plus probables de chaque thématique, puis de comparer leurs labels à ceux produits par des modèles de langage (ChatGPT et Claude AI), afin d'évaluer la fiabilité de l'annotation automatique via les coefficients de Cohen κ (Cohen (1960)) et de Fleiss (Fleiss (1971)) κ_{Fleiss}.

Cette approche a été testée et finalement **rejetée** pour deux raisons structurelles.

- Lorsque le nombre de mots fournis aux annotateurs était faible, ceux-ci ne parvenaient pas à différencier les thématiques entre elles : le signal lexical était insuffisant pour discriminer des topics proches (par exemple, « inondations » et « crues »).
- Lorsqu'on portait le volume de mots à 25 —seuil retenu dans notre protocole de mesure de stabilité—, les experts parvenaient à proposer des labels pour chaque thématique prise isolément, mais rencontraient des difficultés significatives pour distinguer l'évolution d'une même thématique **entre différentes périodes temporelles** t dans les modèles dynamiques (DTM est le cas testé). La granularité temporelle rendait la tâche d'annotation trop complexe et peu reproductible.

3.8.1 Approche retenue : labellisation automatique uniforme via l'API OpenAI

Face à ces limites, nous avons opté pour une approche **uniforme et reproductible**, applicable à l'ensemble des modèles (LDA, DTM, BERTopic, Dynamic BERTopic) sans adaptation spécifique à chacun. La labellisation est réalisée via l'API OpenAI avec le modèle **GPT-4o-mini**, en utilisant le package Python `openai`.

Pour chaque thématique, la requête (*prompt*) fournit au modèle :

- les **25 premiers mots** les plus caractéristiques de la thématique ;
- **3 articles de référence** représentatifs de cette thématique.

La structure du prompt est identique pour tous les modèles. Seul le nom du modèle change d'une requête à l'autre :

Tu es un expert en analyse thématique académique. Voici les mots-clés d'un topic [nom du modèle] : [les 25 premiers mots] Voici quelques documents représentatifs : [les 3 articles de référence]

Donne :

- **un label thématique court**
- **en français**
- **académique**
- **maximum 6 mots**
- **sans guillemets**

Pour les modèles dynamiques (DTM et Dynamic BERTopic), la même requête est exécutée **pour chaque période temporelle** t, afin de capturer l'évolution du label au fil du temps.

Ce choix présente trois avantages par rapport à l'annotation humaine. D'abord, il est **scalable** : le nombre de thématiques et de périodes n'impose aucune contrainte opérationnelle. Ensuite, il est **reproductible** car la même requête produit des résultats stables pour un même ensemble de mots. Enfin, il est **cohérent** : en utilisant un prompt commun à tous les modèles, les labels sont comparables entre méthodes, ce qui facilite l'analyse transversale des résultats.

3.9 Analyse de la dynamique lexicale : le taux de renouvellement

Une fois le découpage temporel optimal établi, une question centrale émerge : l'évolution temporelle d'une thématique traduite-elle une véritable mutation du vocabulaire, ou simplement une redistribution de l'importance des termes au sein d'une thématique figé? En d'autres termes, nous cherchons à distinguer une simple **oscillation statistique** d'une dérive sémantique réelle.

Pour répondre à cette interrogation, nous complétons l'analyse de stabilité de la divergence de Jensen-Shannon par une mesure du **taux de renouvellement (T_r)**. Cette approche nous permet de quantifier précisément les mutations sémantiques en isolant la proportion de nouveaux termes qui intègrent le « cœur » d'une thématique (les mots les plus probables) entre deux périodes consécutives.

L'objectif est d'identifier l'apparition de nouveaux concepts ou l'émergence de nouveaux champs lexicaux qui n'existaient pas, lors de la période précédente. Cette mesure nous permet de confirmer si une thématique s'enrichit de nouvelles problématiques ou si elle se contente de réorganiser ses acquis.

Formellement, soit $W_{k,t}$ l'ensemble de mots n les plus probables ('top_n) pour une thématique k à l'instant t. le taux de renouvellement entre t et $t+1$ est défini par:

$$T_r(k, t \to t+1) = \frac{|W_{k,t+1} \setminus W_{k,t}|}{n}$$

où:

- $W_{k,t+1} \setminus W_{k,t}$ représente l'ensemble des n mots présents à l'instant $t+1$ mais abscents à l'instant t (nouveaux entrants)
- n est la taille du voisinage sémantique observée (fixé ici à $n = 25$)
- $|.|$ Désigne le cardinal(nombre d'éléments) de l'ensemble. Le nombre de mots conservés (stabilité lexicale) est alors donné par l'intersection:

$$\text{Mots restants} = |W_{k,t+1} \cap W_{k,t}|$$

Nous l'interprétons comme suit: Un taux ($T_r = 0$) indique une stabilité parfaite du vocabulaire dominant, tandis qu'un taux approchant $T_r = 1$, signale une mutation profonde, où le sujet change de nature sémantique. Dans notre étude, nous fixons un seuil d'alerte à 25%: tout dépassement de cette valeur est interprété comme une rupture discursive ou l'émergence d'un nouveau sous-thème au sein de la trajectoire du DTM. Cette méthode permet de repérer non seulement quand une thématique évolue, mais aussi d'identifier par quels mots (mots entrants vs mots sortants) cette transformation s'ouvre.

Justification des paramètres (n et seuil de rupture)

La taille du voisinage sémantique est fixée à $n = 25$, en cohérence avec le nombre de mots fournis au modèle de labellisation automatique (GPT-4o-mini). Aligner la mesure sur le champ de vision de l'annotateur garantit que les mutations détectées correspondent à des réalités perceptibles lors de l'étiquetage. Une fenêtre plus large introduirait des variations statistiques imperceptibles à ce niveau de granularité ; une fenêtre plus étroite masquerait des transitions sémantiques significatives (Greene, O'Callaghan, et Cunningham (2014)).
Le seuil de 25 % se justifie par le contexte propre à notre corpus. Les résultats de l'analyse de la divergence de Jensen-Shannon, très proche de zéro sur l'ensemble de la période, indiquent une forte inertie lexicale du discours médiatique. Dans un tel contexte, le remplacement d'un mot sur quatre ne constitue pas un simple bruit statistique~: il signale une rupture discursive réelle ou l'émergence d'un nouveau sous-thème (Blei et Lafferty (2006)).

4 Résultats.

4.1 Analyse du corpus (global vs selection)

Au regard de la méthode de tri sélectionnée, et malgré une croissance régulière du volume d'articles traitant des catastrophes naturelles, cette thématique demeure marginale au sein de la production éditoriale de la RTBF. En effet, avec une moyenne de seulement 1,167 % du corpus total sur la période, ces sujets apparaissent sous-représentés face à la masse globale des publications. Cette proportion reste toutefois évolutive (voir Figure 4.2) et marquée par des pics conjoncturels, notamment en lien avec des événements climatiques extrêmes comme les inondations sévères ayant touché la province de Liège en 2018.

4.2 Analyse exploratoire du corpus sélectionnés.

4.2.1 Analyse des cooccurences entre catégories d'aléas

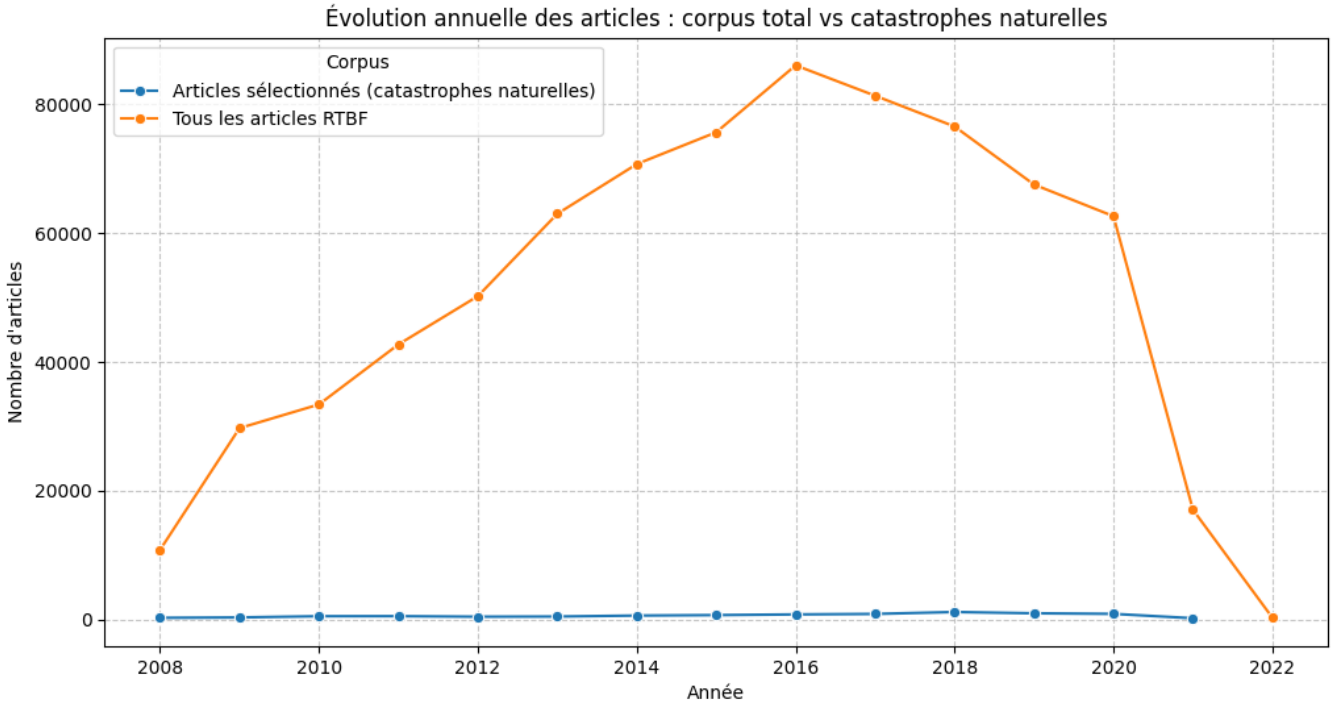
Cette étape vise à explorer les interconnexions sémantiques entre les différentes sous-catégories d'aléas. Notre méthodologie de tri ayant délibérément conservé les "doublons" c'est-à-dire les articles indexés sous plusieurs étiquettes, nous cherchons à identifier les phénomènes qui sont structurellement associés dans le discours médiatique. L'analyse de ces cooccurrences permet de mettre en lumière la manière dont la RTBF traite les événements complexes : elle révèle si certains aléas, comme les tempêtes et les inondations, sont systématiquement présentés comme des risques composés plutôt que comme des événements isolés.

4.2.1.1 Analyse par paires

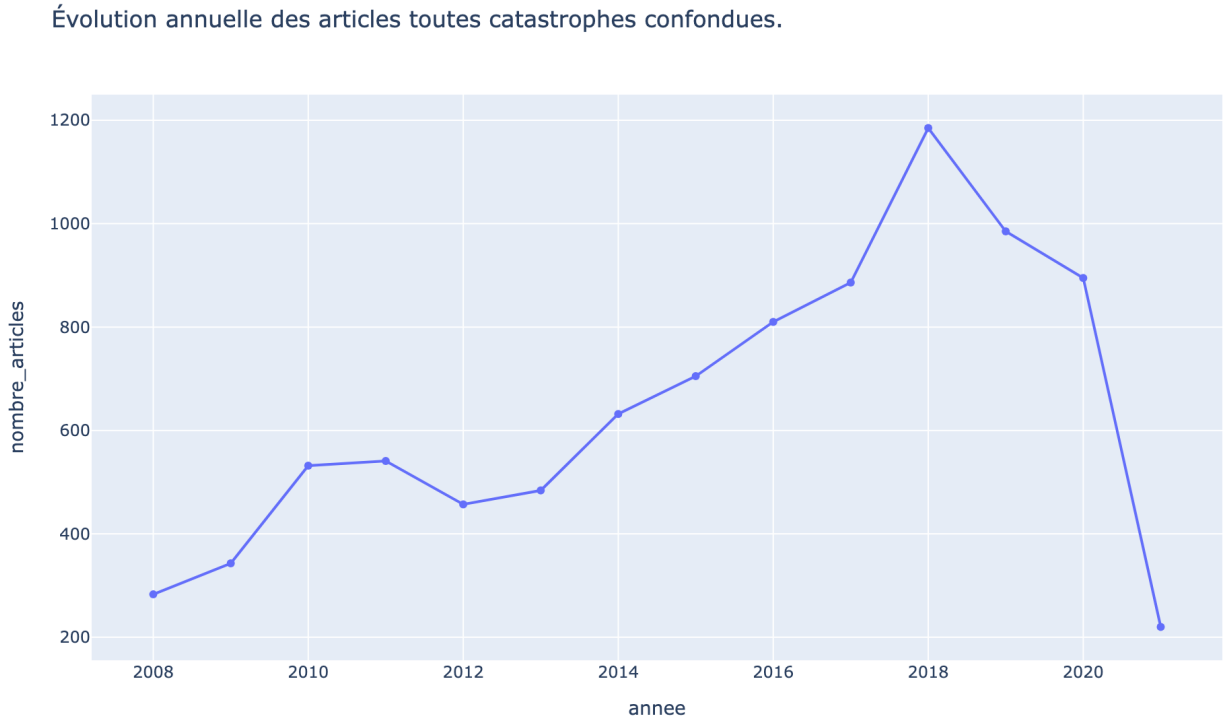
Table 4.1: Matrice de cooccurrence des termes liés aux catastrophes naturelles dans le corpus RTBF

	Inondation	Intempéries	Sécheresse	Tempête	Vague de chaleur
Inondation	**2994**	953	260	959	131
Intempéries	953	**2770**	122	976	67
Sécheresse	260	122	**2068**	189	336
Tempête	959	976	189	**3337**	121
Vague de chaleur	131	67	336	121	**1156**

La Table 4.1 met en évidence la structure sémantique du corpus. Les diagonales indiquent le volume total d'articles par mot-clé (la tempête dominant avec 3337 occurrences). Les relations les plus fortes apparaissent entre "Tempête", "Intempéries" et "Inondation", formant un noyau de reporting lié aux événements climatiques pluvieux, tandis que la "Vague de chaleur" reste un sujet plus isolé statistiquement.



(a) Corpus total vs Catastrophes



(b) Catastrophes sélectionnées uniquement

Figure 4.1: Évolution comparative des publications (2008–2021)

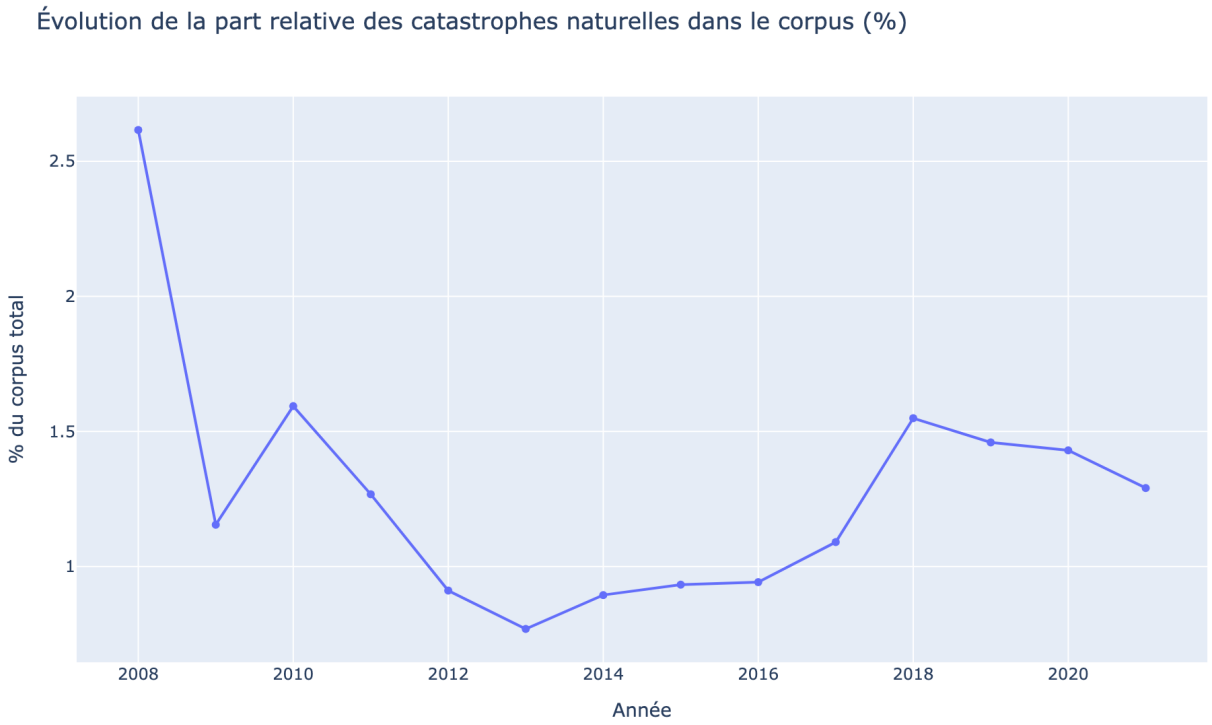


Figure 4.2: Évolution de la part relative des catastrophes naturelles dans le corpus (%)

4.2.1.2 Analyse multi-niveaux

Cette visualisation permet d'appréhender graphiquement les associations entre les sous-catégories d'aléas. L'épaisseur des liens (ou bandes) entre deux nœuds est proportionnelle à la fréquence de leur mention conjointe dans le corpus. Dans la Figure 4.3, nous confirmons l'hypothèse formulée lors de l'examen de la Table 4.1 : les relations les plus denses unissent le triptyque "Tempête", "Intempéries" et "Inondation". Un second pôle d'association significatif émerge entre la "Sécheresse", la "Vague de chaleur" et, de manière plus surprenante, l' "Inondation" (témoignant probablement d'articles traitant des extrêmes climatiques ou de l'alternance sécheresse/crues). Enfin, nous observons l'existence de réseaux interconnectés où l'ensemble des sous-catégories est mobilisé simultanément, bien que cette configuration demeure marginale.

Cette analyse est complétée par les graphiques Figure 4.4. Le graphique de gauche présente la distribution des occurrences individuelles : chaque catégorie est comptabilisée de façon autonome (via l'éclatement des listes), ce qui implique une multi-comptabilisation pour les articles traitant de plusieurs thématiques. Le graphique de droite expose, quant à lui, la hiérarchie des combinaisons de catégories, mettant en lumière les profils de cooccurrence les plus fréquents.

Bien que cette approche nous ait permis d'identifier des phénomènes structurellement associés dans le discours médiatique, la part des cooccurrences reste globalement faible. En effet, nous observons qu'en matière de catastrophes naturelles, la RTBF a tendance à traiter les thématiques de manière isolée (en volume d'articles). Ce constat est étayé par la Figure 4.4, où l'on remarque que les volumes les plus importants concernent les catégories prises individuellement (graphique de gauche), loin devant les scénarios combinés (graphique de droite).

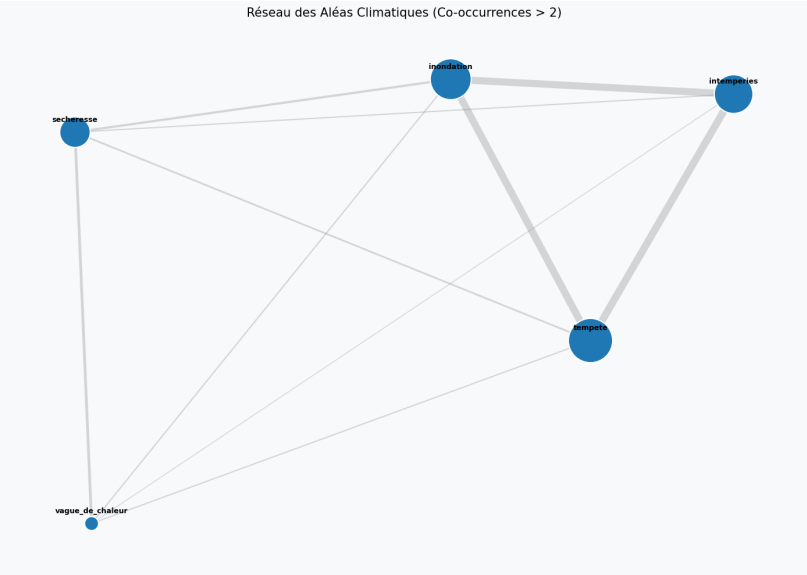


Figure 4.3: Réseau des aléas climatiques

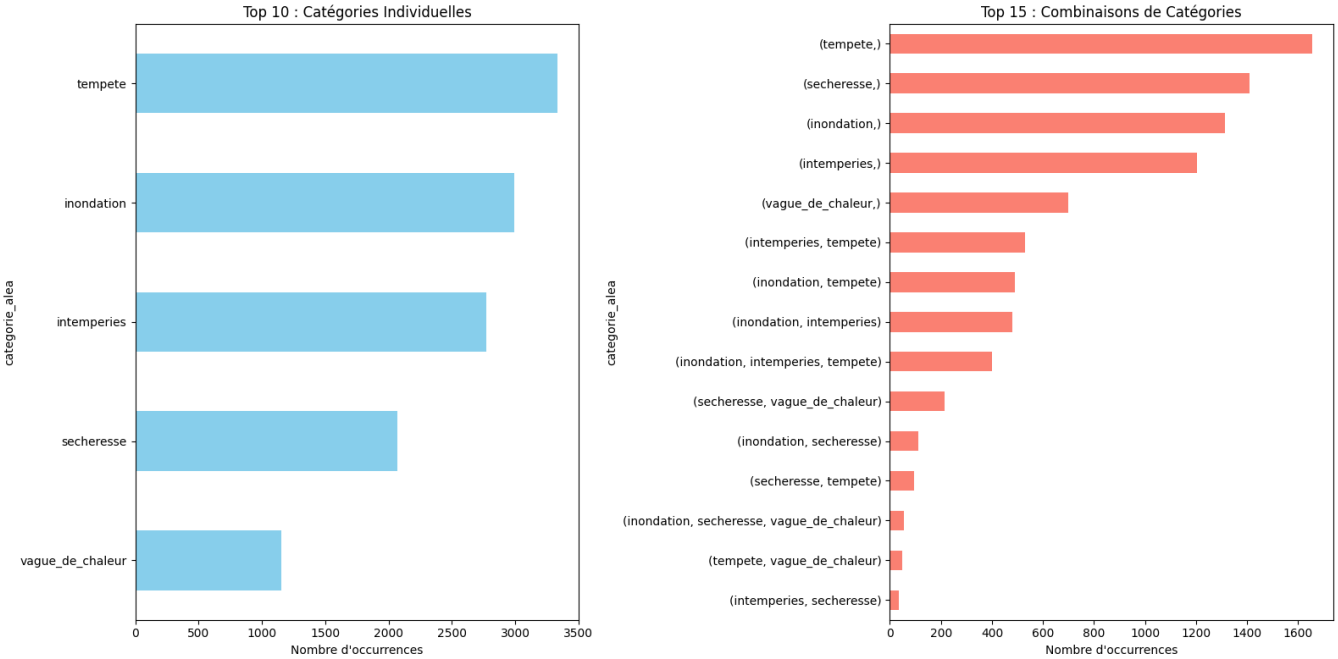


Figure 4.4: Comptage des aléas climatiques : occurrences individuelles (gauche) et combinées (droite)

4.3 Analyse de l'évolution temporelle des catégories d'aléas

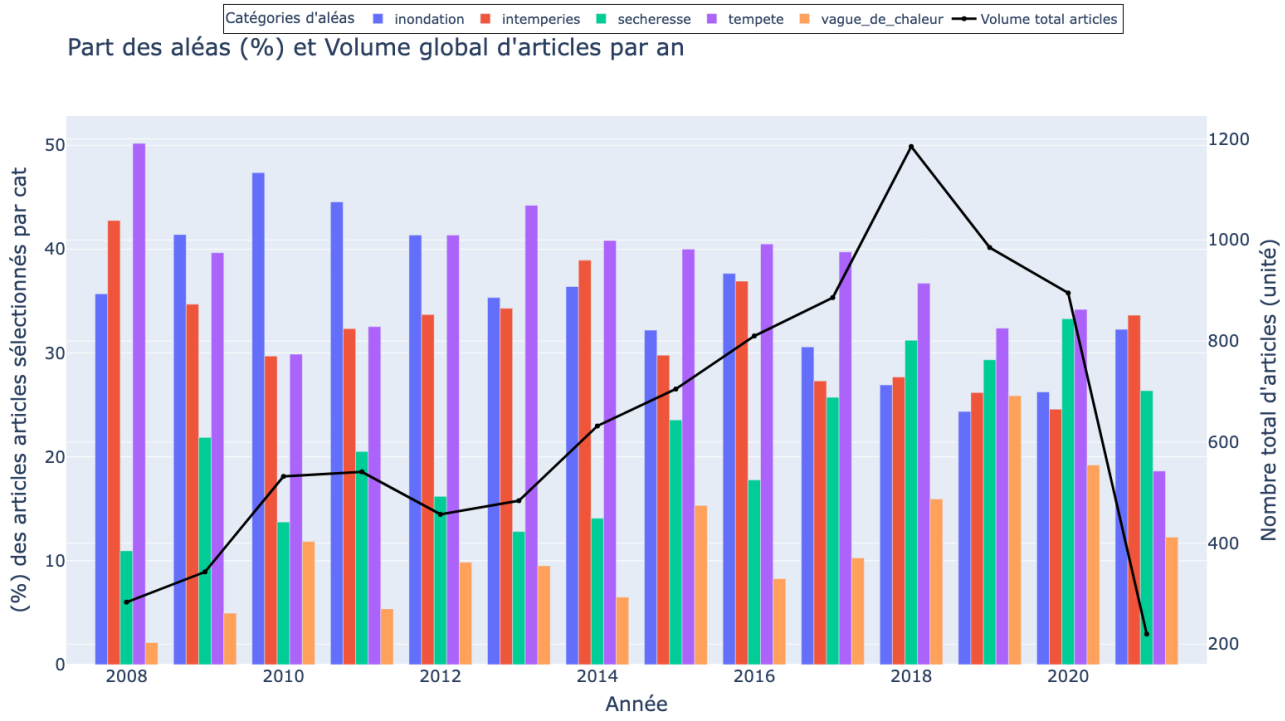
Cette section examine la dynamique du traitement médiatique par la RTBF sur la période 2008–2021. L'objectif est de mesurer la progression du corpus lié aux risques climatiques afin de déterminer si ces thématiques gagnent en visibilité au fil du temps. Nous chercherons également à identifier quels aléas spécifiques s'imposent dans l'agenda médiatique ou connaissent une accélération significative de leur couverture.

> **!** Attention : Total des pourcentages
>
> Il est possible que la somme des pourcentages par année soit supérieure à 100 %. En effet, pour affiner l'analyse, nous avons "éclaté" les catégories groupées. Un article traitant de plusieurs aléas est donc comptabilisé dans chaque catégorie concernée, ce qui peut gonfler le total relatif par année.

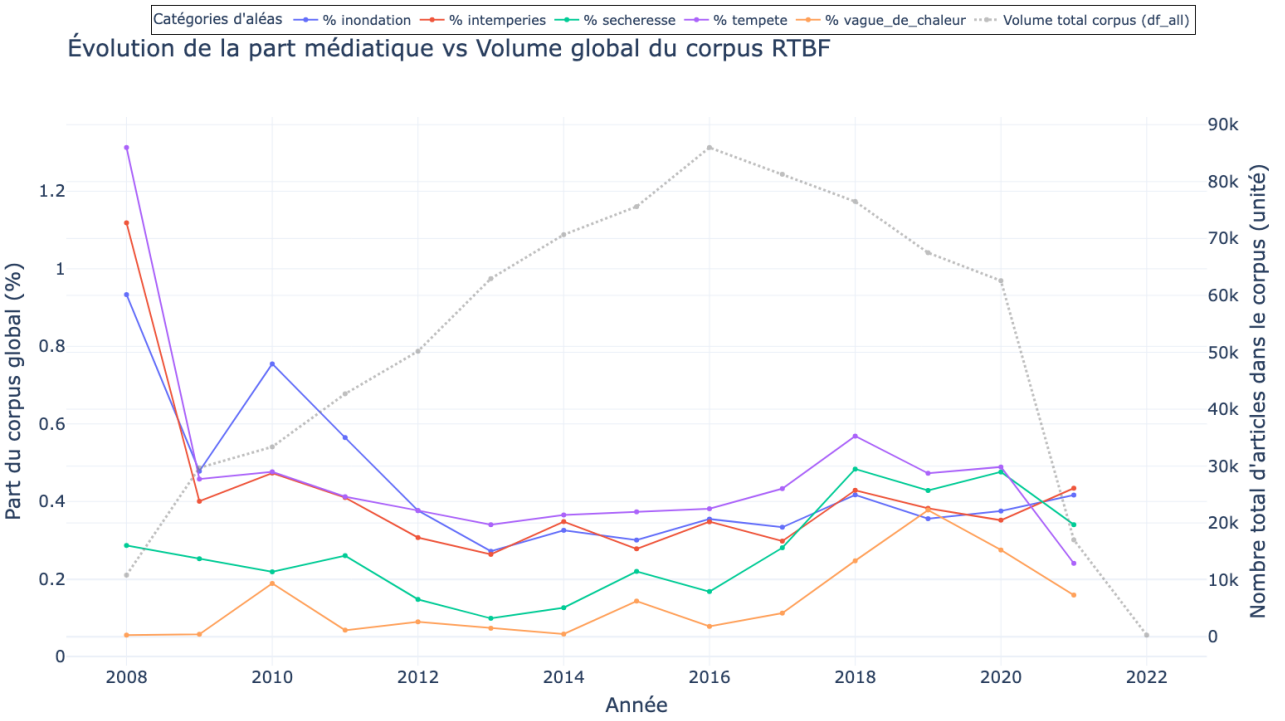
4.3.1 Sous catégories

De façon générale, le volume d'articles liés aux catastrophes naturelles évolue avec le temps. Bien que cette évolution reste faible par rapport au corpus global, comme mentionné dans la Section 4.1 et illustré par la Figure 4.1a, nous faisons les observations suivantes en nous concentrant sur les articles sélectionnés après le tri 2 (Figure 4.5a) :

- Depuis 2013, le volume d'articles liés à la sécheresse a connu une augmentation continue, signifiant peut-être un engouement nouveau de la presse pour ce sujet ou une augmentation du nombre d'occurrences des termes liés à la sécheresse.
- Le corpus se concentre globalement sur des événements en lien avec les aléas tempêtes et inondations, où le seuil d'articles pour la majorité des années se situe autour de 30 % (Figure 4.5a).
- De 2016 à 2020, la part consacrée aux vagues de chaleur augmente de façon continue.
- Nous notons la particularité de l'année 2018, où le pic d'articles toutes catégories confondues est atteint.
- À partir de 2016, les aléas prennent de plus en plus d'ampleur dans le corpus (Figure 4.5b), mouvement principalement marqué par la sécheresse et les vagues de chaleur. En 2018, la part consacrée à la sécheresse dépasse même celle des inondations pour presque égaler celle des tempêtes, qui dominaient relativement le corpus jusqu'ici (Figure 4.5b).



(a) Pourcentage du nombre d'articles après tri 2



(b) Pourcentage global du corpus

Figure 4.5: Évolution comparative des publications articles sélectionnés (2008–2021)

4.3.2 Grandes catégories

Nous utilisons les grandes catégories d'aléas créées précédemment (Section 3.3.2) pour visualiser et comprendre l'évolution de ces phénomènes de façon globale. Par « grandes catégories », nous entendons :

- Les intempéries : regroupant les inondations, les tempêtes et les intempéries diverses ;
- La chaleur : incluant la sécheresse et les vagues de chaleur.

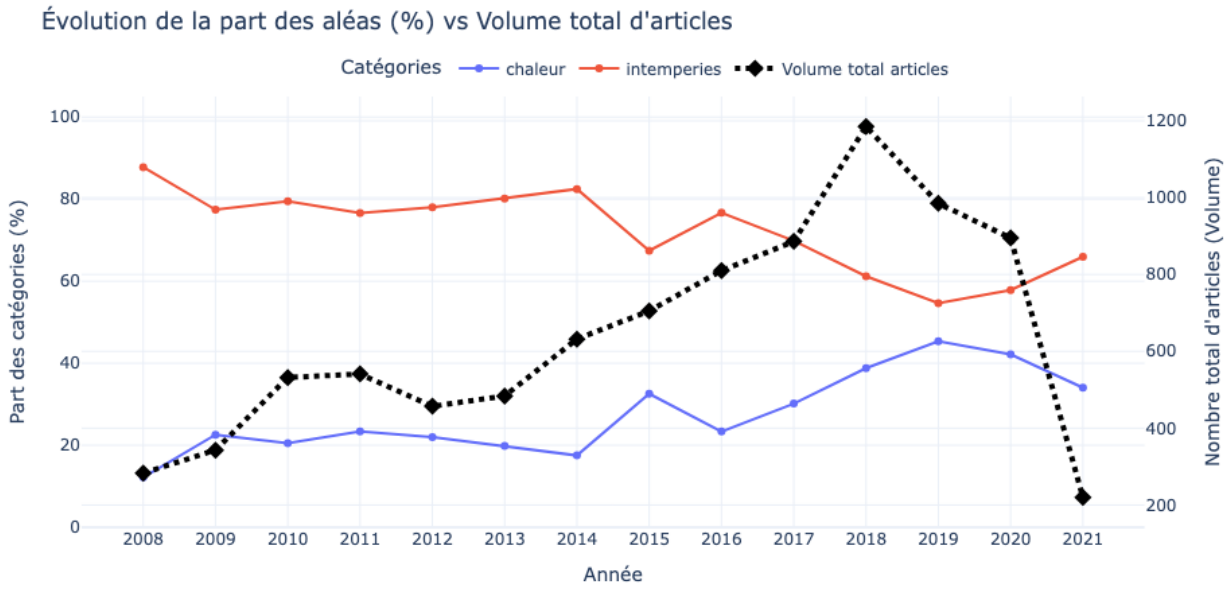


Figure 4.6: Analyse de évolution des grandes catégories

La Figure 4.6 confirme les analyses de la Section 4.3.1 et montre une nette domination des aléas liés aux intempéries par rapport à ceux liés à la chaleur. On observe toutefois un basculement à partir de 2016 : alors que la catégorie « intempéries » connaît une relative perte de vitesse, les aléas liés à la chaleur amorcent une montée en puissance significative.

4.3.3 Analyse de contexte de diffusion

Cette section vise à comprendre l'environnement de diffusion des thématiques liées aux aléas climatiques : dans quels types d'émissions ces sujets sont-ils traités et dans quelle proportion ? L'objectif est d'identifier les formats médiatiques où le débat est le plus soutenu. Pour ce faire, nous nous appuyons sur :

- La Section 3.3.1, où seule la catégorie « RTBF Info » a été conservée (dont la description détaillée est fournie en Figure 3.1) ;
- La Section 3.3.2, qui définit les grandes catégories d'aléas utilisées pour l'analyse.

De façon générale, le contexte de traitement médiatique est très varié (Figure 4.7), bien que les intempéries demeurent le sujet prédominant dans tous les formats. On observe une prédominance des faits liés à l'actualité internationale, la catégorie « MONDE » dominant nettement le classement. En revanche, si nous isolons les catégories « RÉGIONS » et « BELGIQUE » (respectivement 3^{e} et 4^{e} en volume d'articles), nous constatons une sous-représentation de ces aléas à l'échelle locale par rapport au flux global. Rouquette et Bihay (2022) soulignaient déjà ce phénomène à travers l'exemple du retrait-gonflement des argiles (RGA)

: certains aléas sont davantage traités par la presse locale que nationale en raison de leur faible ampleur initiale ou du caractère diffus des dégâts au niveau macro-économique. Cette analyse pourrait expliquer l'observation faite ici. Enfin, la deuxième place occupée par la catégorie « SOCIÉTÉ » suggère que les aléas climatiques sont de plus en plus perçus comme des faits de société globaux plutôt que de simples événements météorologiques isolés.

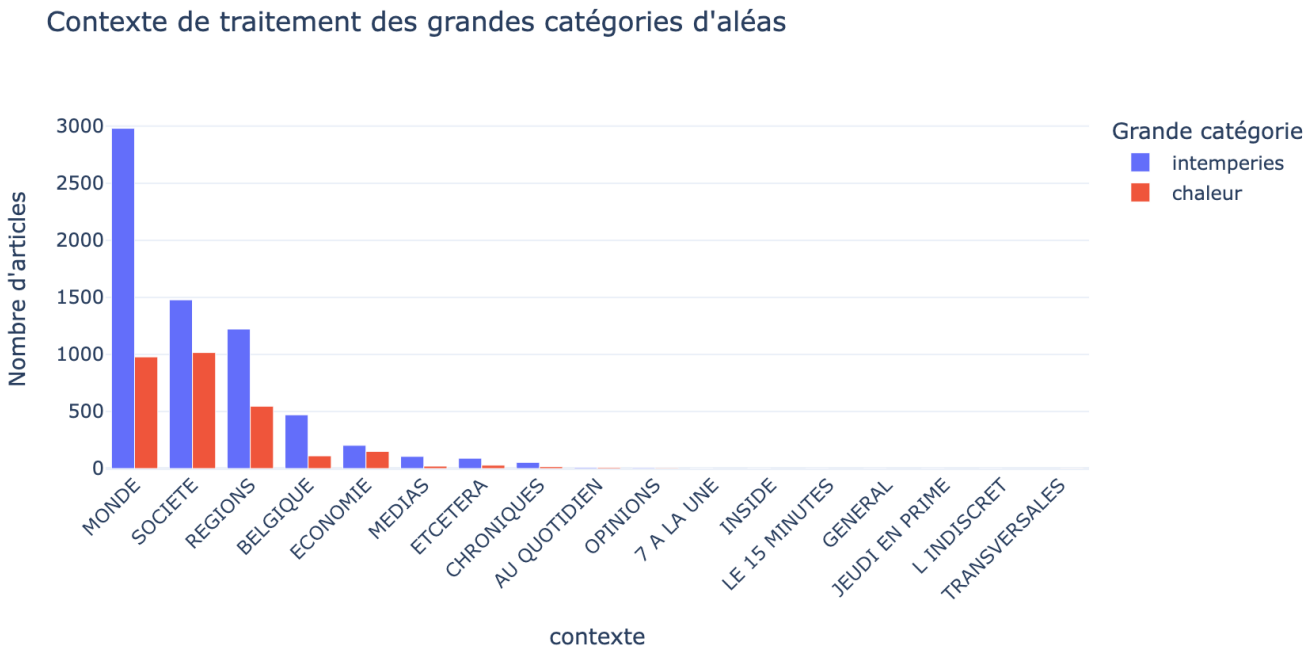


Figure 4.7: Analyse du contexte médiatique

4.4 Modélisation de thématiques

L'analyse exploratoire a mis en évidence un double mouvement : d'une part, la persistance des aléas traditionnels (tempêtes et inondations) comme piliers de la couverture médiatique ; d'autre part, l'émergence marquée des risques liés au stress thermique (sécheresse, chaleur) à partir de 2016. Cette progression quantitative pose la question de la nature du discours : le traitement médiatique se limite-t-il au récit factuel de la catastrophe, ou intègre-t-il désormais une dimension analytique et climatique plus globale? C'est ce que les modèles de *topic modelling* vont permettre d'explorer.

Objectif de la modélisation

L'objectif est de passer d'une analyse de *mots-clés* à une analyse de *concepts* (thématiques), afin de déceler les structures latentes du discours médiatique et d'en suivre l'évolution dans le temps.

4.4.1 Modèle LDA

Nous utiliserons comme point de départ le modèle Latent Dirichlet Allocation (LDA), tel que présenté dans la Section 2.2.5.

4.4.1.1 Analyse de sensibilité du modèle (stabilité des topics)

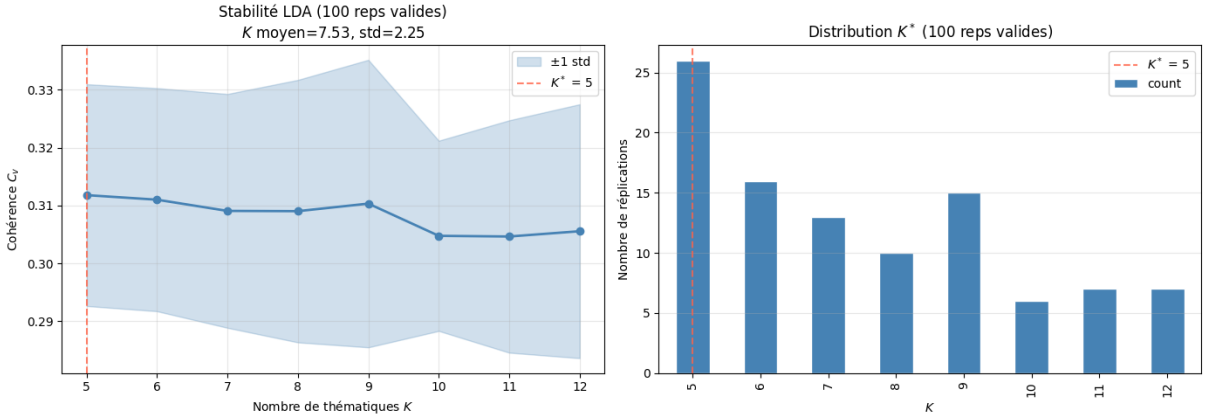


Figure 4.8: Analyse de la partition du corpus

L'analyse de sensibilité met en évidence la robustesse globale de la démarche de sélection du nombre de thématiques. Bien que la valeur $K^* = 5$ corresponde au mode de la distribution des réplications (26 réplications sur 100), les valeurs optimales observées s'étendent jusqu'à $K = 12$. La moyenne des réplications atteint ainsi $\bar{K} = 7{,}53$ avec un écart-type de 2,25, ce qui suggère une certaine hétérogénéité structurelle du corpus.

Par ailleurs, les scores de cohérence C_v demeurent relativement stables entre $K = 5$ et $K = 12$, avec des valeurs proches de 0,31. Cette stabilité indique l'absence d'un optimum unique clairement dominant. Dans ce contexte, le choix de retenir $K = 8$ apparaît comme un compromis méthodologiquement cohérent : cette valeur reste proche de la moyenne observée des réplications tout en permettant une représentation thématique plus riche du corpus.

Le modèle final, entraîné sur l'intégralité du corpus avec des hyperparamètres plus exigeants (200 passes et 1,000 itérations), atteint une cohérence de $C_v = 0{,}4196$, nettement supérieure à celle obtenue lors des phases de sous-échantillonnage. Cette différence s'explique principalement par le fait que les expérimentations de stabilité ont été réalisées avec des paramètres volontairement réduits afin de limiter le coût computationnel.

Afin d'approfondir cette analyse de robustesse, les résultats de la procédure de ré-échantillonnage sont présentés dans le Table 4.2. Les intervalles de confiance à 95,% sont construits selon le cadre classique de l'inférence sur la moyenne (Student (1908)). Pour chaque valeur de K, la cohérence moyenne $\bar{C}_v$ est estimée à partir de $n = 100$ réplications indépendantes. L'erreur type est définie par :

$$se = \frac{s}{\sqrt{n}}$$

où s désigne l'écart-type empirique des scores de cohérence observés sur les 100 réplications.

Les bornes des intervalles de confiance sont ensuite calculées selon :

$$\left[\bar{C}_v - t_{n-1,\,0{,}975} \times se \;;\; \bar{C}_v + t_{n-1,\,0{,}975} \times se\right]$$

où $t_{n-1,\,0{,}975}$ correspond au quantile à 97,5,% de la loi de Student à $n - 1 = 99$ degrés de liberté, soit :

$$t_{99;\,0{,}975} = 1{,}984$$

L'utilisation de la loi de Student est ici privilégiée à l'approximation normale classique ($z = 1{,}96$) pour des raisons suivantes : lorsque la variance σ^2 de la population est inconnue et estimée par l'écart-type empirique s, la statistique de test suit exactement une loi de Student à $n-1$ degrés de liberté, et non une loi normale (Student (1908)). La loi normale ne constituerait qu'une approximation asymptotique valide lorsque $n \to \infty$, ce qui n'est pas formellement satisfait dans notre cas ($n = 100$), même si la différence numérique entre $t_{99;\,0{,}975} = 1{,}984$ et $z_{0{,}975} = 1{,}960$ reste marginale en pratique.

Table 4.2: Intervalles de confiance à 95,% (Student t) sur 100 réplications

K	Moyenne ($\bar{C}_v$)	Écart-type (s)	Erreur type (se)	IC inf. (95,%)	IC sup. (95,%)
5	0,3118	0,0192	0,0019	0,3080	0,3156
6	0,3110	0,0193	0,0019	0,3072	0,3148
7	0,3091	0,0202	0,0020	0,3051	0,3131
8	0,3090	0,0227	0,0023	0,3045	0,3135
9	0,3103	0,0248	0,0025	0,3054	0,3153
10	0,3048	0,0164	0,0016	0,3015	0,3080
11	0,3047	0,0201	0,0020	0,3007	0,3086
12	0,3056	0,0219	0,0022	0,3012	0,3099

L'examen du Table 4.2 apporte plusieurs éléments permettant de justifier le choix méthodologique retenu (le nombre de topic K final).

Tout d'abord, les erreurs types observées restent extrêmement faibles, autour de 0,002 pour l'ensemble des configurations testées. Cette faible dispersion indique une bonne convergence des réplications et suggère que 100 sous-échantillonnages suffisent pour obtenir une estimation stable des cohérences moyennes.

Ensuite, les différences de cohérence entre les différentes valeurs de K demeurent très limitées. Entre $K = 5$ et $K = 8$, la baisse observée de la cohérence moyenne n'est que de 0,0028, soit moins de 1,%. Les intervalles de confiance se chevauchent largement, ce qui confirme l'absence de différence statistiquement significative entre ces configurations.

Enfin, la proximité des intervalles de confiance associés à $K = 5$ et $K = 8$ montre que l'augmentation du nombre de thématiques n'entraîne pas de dégradation importante de la qualité sémantique du modèle. Le choix de retenir $K = 8$ permet ainsi de conserver une cohérence satisfaisante tout en offrant une représentation plus riche et plus nuancée de la diversité structurelle du corpus mise en évidence lors de l'analyse de sensibilité.

4.4.1.2 Analyse de sensibilité du modèle (stabilité sémantique)

L'analyse de la figure (Figure 4.9) permet de valider statistiquement la robustesse de notre modèle. La similarité moyenne de Jaccard, calculée sur les 10 paires issues des 5 graines aléatoires ($C(5,2) = \frac{5\times4}{2} = 10$ paires), sur nos 8 thématiques s'établit à $\bar{J} = 0{,}462$.

Selon les seuils empiriques de Greene, O'Callaghan, et Cunningham (2014), ce score indique une **stabilité modérée** ($0{,}4 < \bar{J} \leq 0{,}6$). Ce constat s'explique par plusieurs facteurs :

- **Complexité sémantique** : les journalistes emploient un vocabulaire varié (synonymes, métaphores), ce qui induit naturellement une légère fluctuation des mots les plus probables d'un seed à l'autre.

- **Granularité des thèmes** : avec $K = 8$ thèmes, certains sujets présentent des frontières poreuses (par exemple, un thème « politique » et un thème « économie » partageant des termes relatifs au budget des catastrophes).

L'examen de la stabilité locale révèle par ailleurs une structure thématique contrastée : en fixant le seuil de robustesse à 0,5, **trois thématiques sur huit** sont statistiquement stables (Topics 0, 5 et 7), tandis que cinq présentent une instabilité notable.

L'examen du 4.3 révèle la présence quasi systématique de termes peu informatifs — « nbr », « week », « autre » — dans l'ensemble des thématiques, constituant un bruit lexical susceptible de nuire à la qualité de la modélisation. Une vérification complémentaire confirme qu'aucun mot supplémentaire n'a été écarté automatiquement par les filtres de Gensim lors de cette configuration.

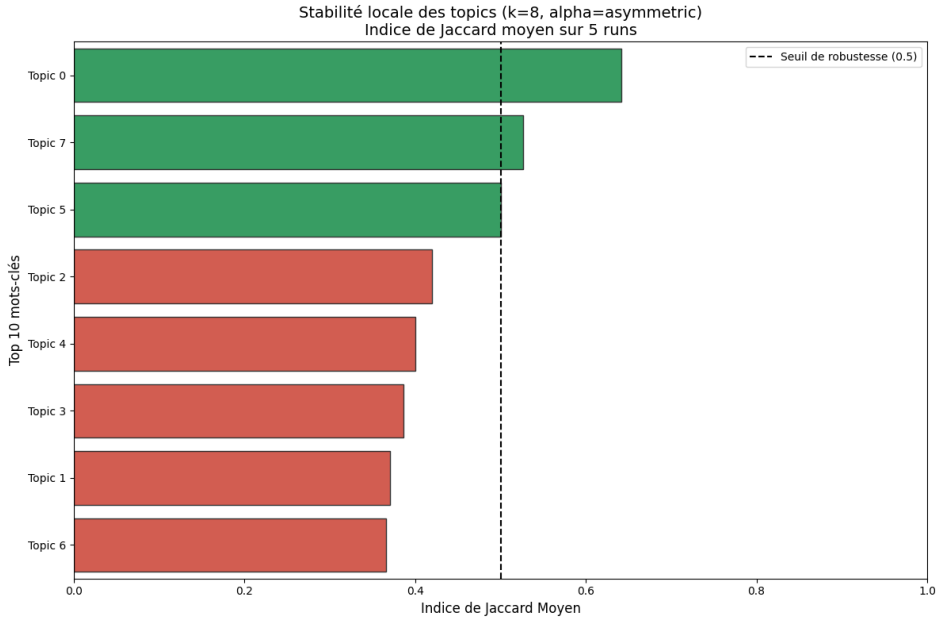


Figure 4.9: Stabilité locale des topics (seuil d'indice fixé à 0.5) avant filtrage

Topic	Mots les plus probables (top 10)
Topic 0	nbr, week, pompier, autre, heure, eau, rue, orage, nouveau, inondation
Topic 1	nbr, week, mort, région, ville, autre, personne, sud, inondation, incendie
Topic 2	nbr, année, pays, autre, week, nbr_million, nouveau, rapport, jour, premier
Topic 3	nbr, animal, autre, certain, journaliste, drone, nouveau, cas, week, région
Topic 4	nbr, nouveau, autre, projet, certain, premier, année, grand, commun, eau
Topic 5	eau, nbr, autre, année, agriculteur, certain, petit, arbre, culture, nouveau
Topic 6	nbr, autre, antarctique, zone, glacier, eau, région, scientifique, étude, année
Topic 7	eau, nbr, autre, certain, année, nouveau, grand, week, commun, maison

Table 4.3: Top 10 des mots par topic avant filtrage

Pour remédier à ce bruit lexical, les filtres suivants ont été appliqués :

- suppression manuelle des termes peu informatifs (« nbr », « autre », « week », « nouveau ») ;
- suppression automatique des mots apparaissant dans plus de 50,% des documents (trop fréquents) et dans moins de 5,% des documents (trop rares).

Une vérification manuelle du vocabulaire avant et après filtrage confirme que le filtre automatique (`filter_extremes`) n'a écarté **aucun terme supplémentaire** : les seuls mots retirés du dictionnaire sont ceux explicitement supprimés lors de l'étape manuelle. Ce résultat indique que le corpus, après le tri de niveau 2 (Section 3.3.2), présente déjà une distribution lexicale relativement équilibrée — les termes résiduels apparaissent tous dans une proportion comprise entre 5,% et 50,% des documents, sans qu'aucun ne dépasse les seuils de bruit définis.

Le nouveau graphique (Figure 4.10) ci-dessous montre les résultats obtenus après application de ces filtres. La similarité moyenne de Jaccard progresse à $\bar{J} = 0{,}487$, contre 0,462 avant filtrage.

Ce nouveau filtrage a permis plusieurs améliorations :

- **Progression de la similarité globale** : $\bar{J}$ passe de 0,462 à 0,487.
- **Stabilité locale améliorée** : trois thématiques sur huit franchissent désormais le seuil de robustesse de 0,5 (Topics 2, 4 et 7, en vert dans la figure ci-dessous). Les scores des topics stables sont par ailleurs nettement plus élevés ($\bar{J} \approx 0{,}75$ pour Topic 2).
- **Réduction du bruit lexical** : l'élimination des termes non discriminants permet aux topics d'être définis par des mots plus spécifiques (Table 4.4), ce qui améliore l'interprétabilité et la différenciation sémantique entre thématiques.

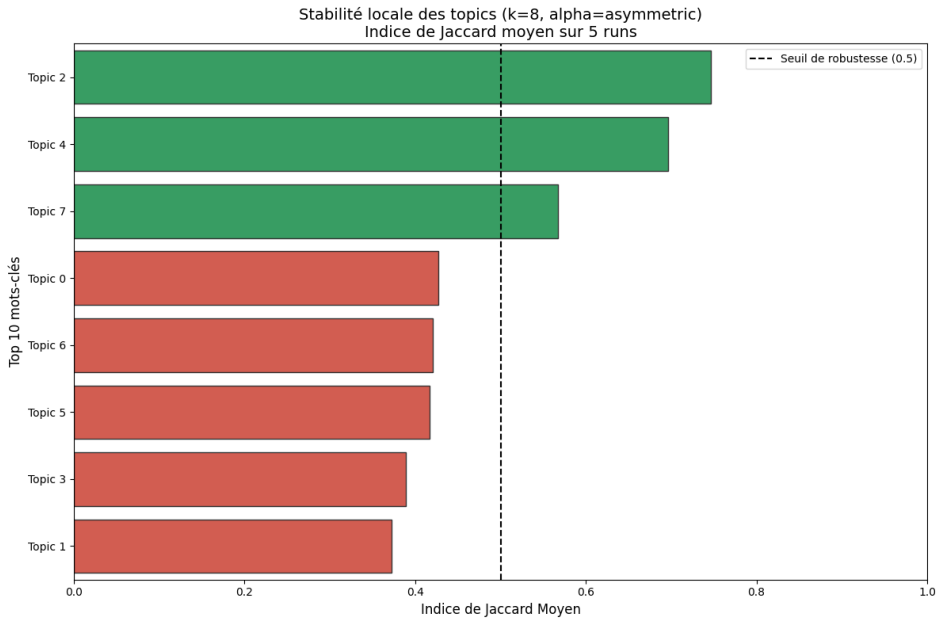


Figure 4.10: Stabilité locale des topics(seuil d'indice fixé à 0.5) après filtrage

Table 4.4: Top 10 des mots par topic après filtrage

Topic	Mots les plus probables (top 10)
Topic 0	mort, région, ville, inondation, personne, sud, nord, etat, pluie, maison
Topic 1	certain, projet, cas, commun, premier, bruxelle, train, dernier, travail, côté
Topic 2	ville, mort, personne, pays, gouvernement, grand, premier, capital, certain, président
Topic 3	personne, femme, nbr_million, pays, cas, mort, enfant, président, année, premier
Topic 4	année, température, jour, région, situation, certain, agriculteur, petit, normal, chaleur
Topic 5	région, vol, heure, accident, mort, mission, zone, appareil, avion, premier
Topic 6	année, rapport, pays, climat, scientifique, nbr_million, certain, prix, étude, grand
Topic 7	pompier, incendie, commun, rue, région, maison, heure, situation, ville, inondation

Ces résultats confirment que le filtrage appliqué au corpus constitue un prétraitement nécessaire pour l'amelioration des résultats : la structure thématique est désormais plus robuste aux conditions d'initialisation, sans qu'aucun terme pertinent n'ait été supprimé automatiquement par les seuils de fréquence.

4.4.1.3 Analyse et visualisation des résultats.

Une fois le modèle stabilisé à huit thématiques, nous utilisons l'outil pyLDAvis (Sievert et Shirley (2014)) pour projeter les topics dans un espace bidimensionnel via une mise à l'échelle multidimensionnelle (MDS). La distance entre les cercles indique leur degré de distinction sémantique ; la taille de chaque cercle reflète l'importance relative du topic au sein du corpus.

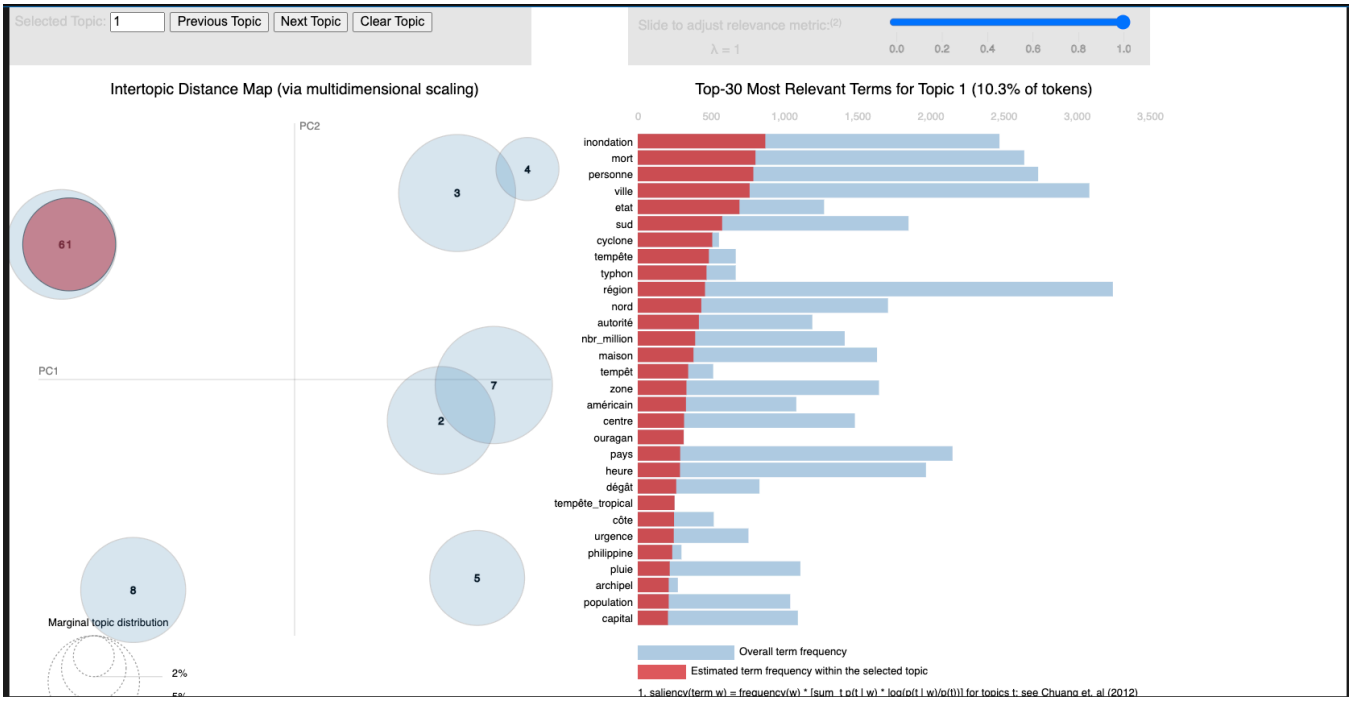


Figure 4.11: Visualisation des thématiques — modèle LDA ($K = 8$,α = asymmetric)

L'analyse de la Figure 4.11 met en évidence une structuration du corpus autour de plusieurs ensembles sémantiques relativement distincts, tout en révélant certaines zones de chevauchement entre thématiques proches.

Structuration générale de l'espace thématique.

La projection MDS met globalement en évidence une bonne séparation entre la majorité des topics. Plusieurs zones de proximité apparaissent toutefois sur la carte, suggérant l'existence de frontières thématiques plus poreuses entre certains ensembles discursifs. Trois rapprochements principaux peuvent être observés : entre les Topics 1 et 6, entre les Topics 3 et 4, ainsi qu'entre les Topics 2 et 7. Ces proximités traduisent l'existence de vocabulaires partagés ou de registres discursifs partiellement communs.

Trois configurations principales émergent de la visualisation.

Le premier ensemble, situé dans la partie gauche de la carte, regroupe les Topics 1 et 6. Le Topic 1 représente environ 10,3 % des tokens du corpus, tandis que le Topic 6 en représente environ 14,4 %. Leur forte proximité spatiale suggère une importante similarité lexicale autour des catastrophes naturelles, des situations d'urgence et de leurs conséquences humaines.

Un second ensemble apparaît dans la zone centrale de la carte. Les Topics 3 et 4 se chevauchent partiellement dans la partie supérieure, avec une domination plus marquée du Topic 3. Plus au centre, les Topics 2 et 7 présentent également un recouvrement partiel, avec des poids relativement proches dans le corpus. Cette proximité semble refléter des liens entre les dimensions institutionnelles, médiatiques et socio-économiques des risques environnementaux.

Enfin, deux topics occupent des positions plus isolées dans l'espace MDS. Le Topic 5, situé en bas à droite, et le Topic 8, situé en bas à gauche, apparaissent nettement séparés des autres groupes. Leur éloignement suggère une plus forte spécificité thématique et une cohérence lexicale interne plus marquée.

L'examen détaillé des labels et des documents représentatifs (Table 6.4 et Table 6.5 en annexe) permet ensuite de préciser le contenu sémantique associé à chacune des huit thématiques identifiées.

1. Topics directement liés aux catastrophes naturelles

Deux topics se distinguent par leur ancrage direct dans les catastrophes naturelles et les événements extrêmes.

Le **Topic 1** (*Catastrophes naturelles et impacts humains*) rassemble principalement des articles consacrés à des événements météorologiques majeurs à l'échelle internationale, comme le cyclone Pam au Vanuatu (2015) ou le typhon Mangkhut aux Philippines (2018). Le vocabulaire dominant renvoie aux destructions, aux populations touchées et aux opérations de secours.

Le **Topic 6** (*Catastrophes naturelles et pertes humaines*) se concentre davantage sur les catastrophes associées à un lourd bilan humain, notamment certains accidents ou naufrages médiatisés. La proximité observée entre les Topics 1 et 6 dans l'espace MDS s'explique ainsi par un lexique commun autour des victimes, des décès, des secours et des situations de crise.

2. Topics liés aux enjeux climatiques et environnementaux

Le **Topic 5** (*Changements climatiques et agriculture durable*) traite principalement des effets du changement climatique sur l'agriculture et les ressources naturelles. Les documents représentatifs évoquent notamment les sécheresses, les vendanges précoces, les nappes phréatiques ou encore les difficultés liées aux ressources hydriques.

Le **Topic 7** (*Changements climatiques et impacts forestiers*) adopte une perspective plus scientifique et environnementale, centrée sur les forêts, la biodiversité, la déforestation ou encore la séquestration du carbone. Plusieurs documents mobilisent un vocabulaire issu de la recherche scientifique et des publications académiques.

Le chevauchement partiel entre ces deux topics apparaît cohérent : ils appartiennent tous deux au champ discursif du changement climatique, mais selon des angles différents, l'un davantage tourné vers les impacts agricoles et socio-économiques, l'autre vers les dynamiques environnementales et écologiques.

3. Une thématique spécifiquement locale

Le **Topic 8** (*Interventions des pompiers en intempéries*) se distingue par son fort ancrage dans le contexte belge. Les documents représentatifs concernent essentiellement des épisodes locaux d'intempéries impliquant les services de secours, notamment dans le Hainaut ou le Brabant wallon. Le vocabulaire mobilisé est très opérationnel et renvoie fréquemment aux pompiers, aux caves inondées, aux plans communaux d'urgence ou aux interventions de terrain.

Cette forte spécificité lexicale explique probablement son isolement relatif dans la projection MDS.

4. Topics périphériques à la problématique principale

Enfin, les Topics 3 et 4 mettent en évidence certaines limites inhérentes au *topic modeling* appliqué à un corpus journalistique généraliste.

Le **Topic 3** (*Politique, médias et liberté d'expression*) et le **Topic 4** (*Santé publique et impacts environnementaux*) capturent des dimensions discursives présentes dans le corpus RTBFINFO mais moins directement liées à notre objet d'étude centré sur les risques naturels et climatiques.

Leur présence s'explique principalement par la nature généraliste du corpus journalistique utilisé. Les articles relatifs aux catastrophes environnementales coexistent en effet avec des contenus politiques, sanitaires ou sociétaux plus larges, ce qui favorise l'émergence de thématiques périphériques dans un cadre d'apprentissage non supervisé.

Cette forme de contamination thématique reste néanmoins relativement limitée et ne remet pas en cause la cohérence globale des autres topics, qui demeurent fortement liés aux enjeux climatiques, environnementaux et aux catastrophes naturelles étudiées dans ce mémoire.

4.4.1.4 Analyse du découpage temporel (`timepoint`) du DTM

L'objectif central de cette démarche est d'identifier le découpage temporel le plus pertinent pour suivre l'évolution des thématiques dans le temps tout en conservant une continuité sémantique suffisante entre les périodes. La Figure 4.12a présente ainsi la répartition des articles par année selon les huit périodes temporelles retenues dans cette étude.

Pour évaluer la stabilité et l'évolution du discours médiatique d'une période à l'autre, nous mobilisons la divergence de Jensen-Shannon (JS_{moyen}). Cette mesure quantifie la distance sémantique entre deux distributions lexicales successives. Une valeur proche de 1 traduirait une rupture discursive majeure entre deux périodes, tandis qu'une valeur faible indique une forte continuité lexicale et thématique.

Les résultats présentés dans la Figure 4.12b montrent que les valeurs observées demeurent globalement faibles, comprises entre 0,043 et 0,053. Cela suggère que le discours médiatique relatif aux catastrophes naturelles conserve une forte inertie lexicale sur l'ensemble de la période étudiée. Les transformations discursives apparaissent donc davantage progressives que brutales.

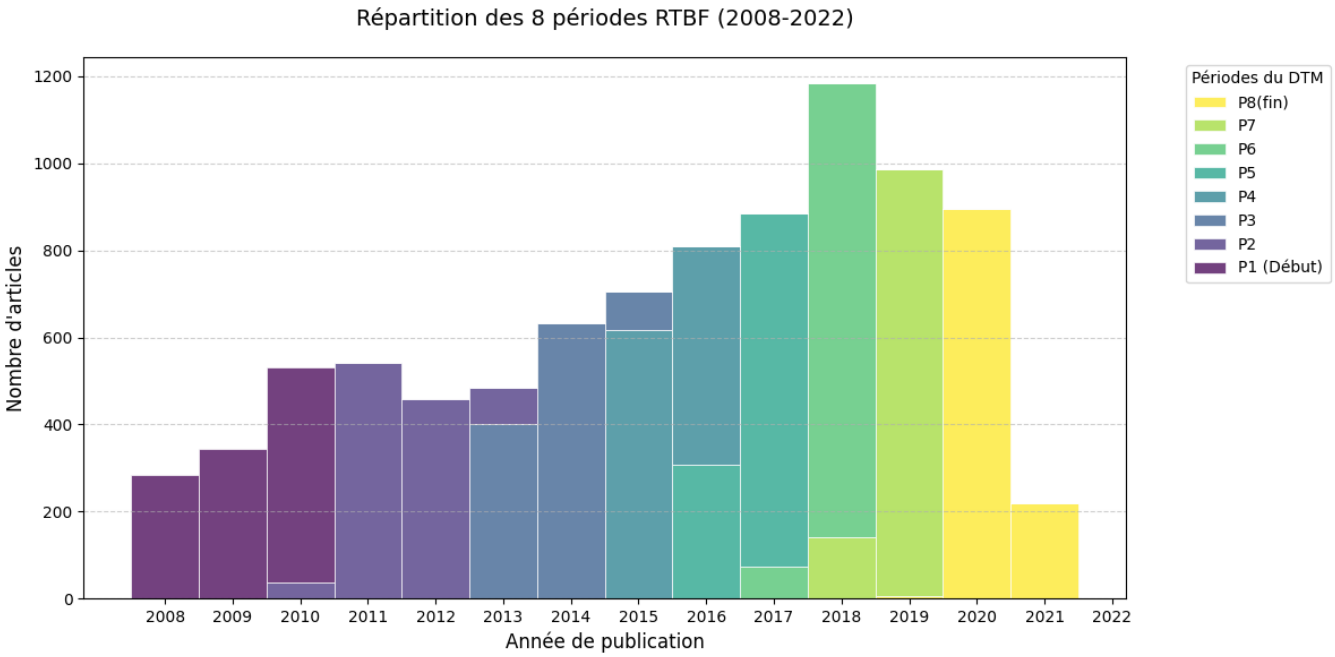
Dans cette analyse, l'objectif n'est toutefois pas de rechercher la stabilité maximale, mais plutôt d'identifier le découpage temporel permettant de mieux faire émerger les évolutions thématiques du corpus. Le Table 4.5 compare ainsi les performances du modèle pour deux configurations contrastées : $T = 8$, correspondant au niveau maximal de divergence observé, et $T = 13$, qui présente au contraire les valeurs les plus faibles.

Table 4.5: Comparaison des performances du DTM selon la granularité temporelle

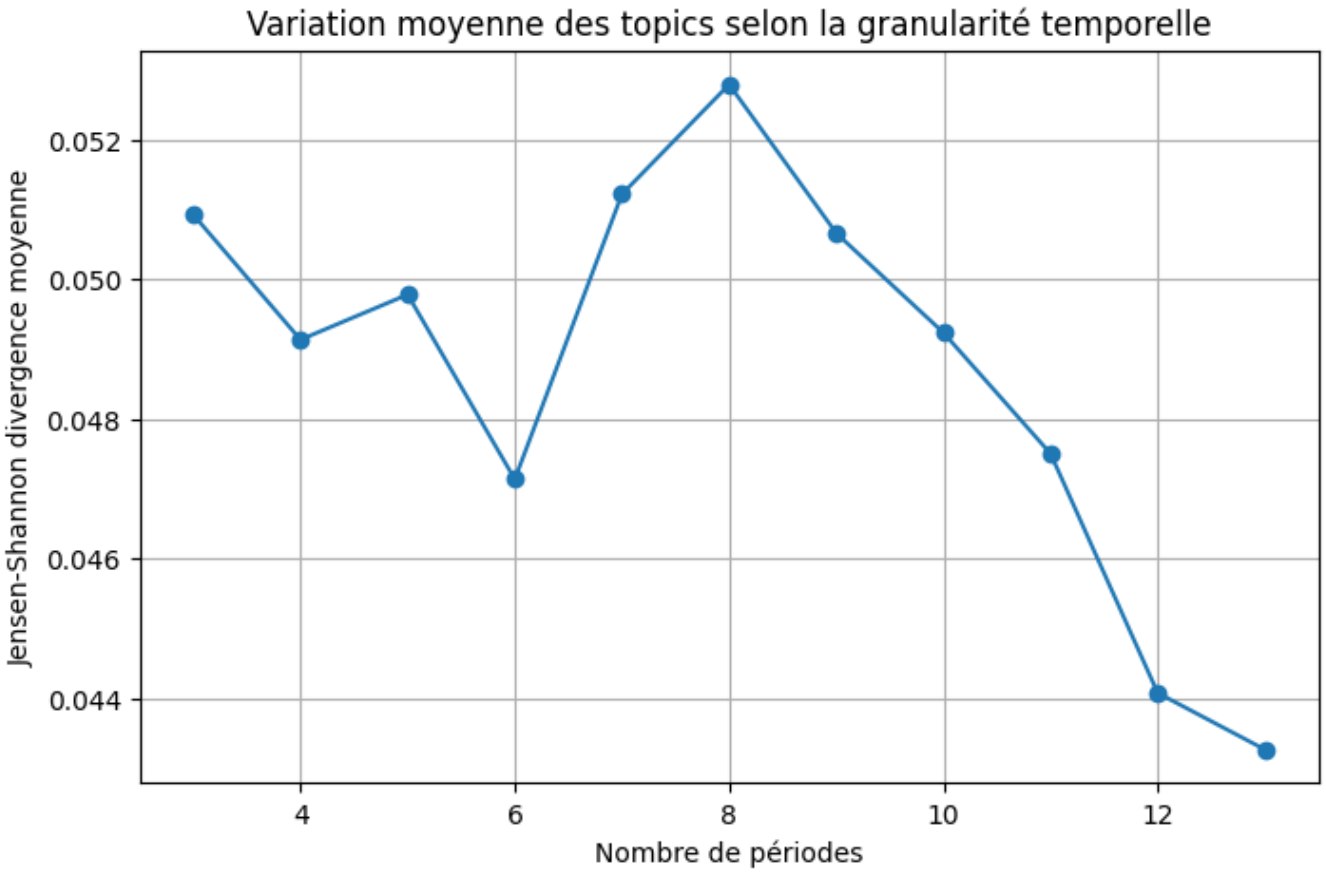
Métrique	$T = 8$	$T = 13$
Log-Likelihood	−10,469	**−10,424**
Cohérence $C_{U_{\text{mass}}}$	−2,0601	**−1,9655**
Divergence JS (JS_{moyen})	**0,0527**	0,0432

La configuration reposant sur $T = 8$ **périodes** est finalement retenue comme le meilleur compromis méthodologique pour cette étude.

Premièrement, cette configuration maximise la divergence moyenne de Jensen-Shannon ($JS_{\text{moyen}} = 0{,}0527$), ce qui indique qu'elle est la plus sensible aux variations discursives du corpus. Ce découpage permet donc de mieux mettre en évidence les transitions sémantiques, les glissements lexicaux et les recompositions thématiques au fil du temps. À l'inverse, un découpage plus fin comme $T = 13$ tend à lisser davantage les différences entre périodes successives, au risque de rendre certaines évolutions moins perceptibles.



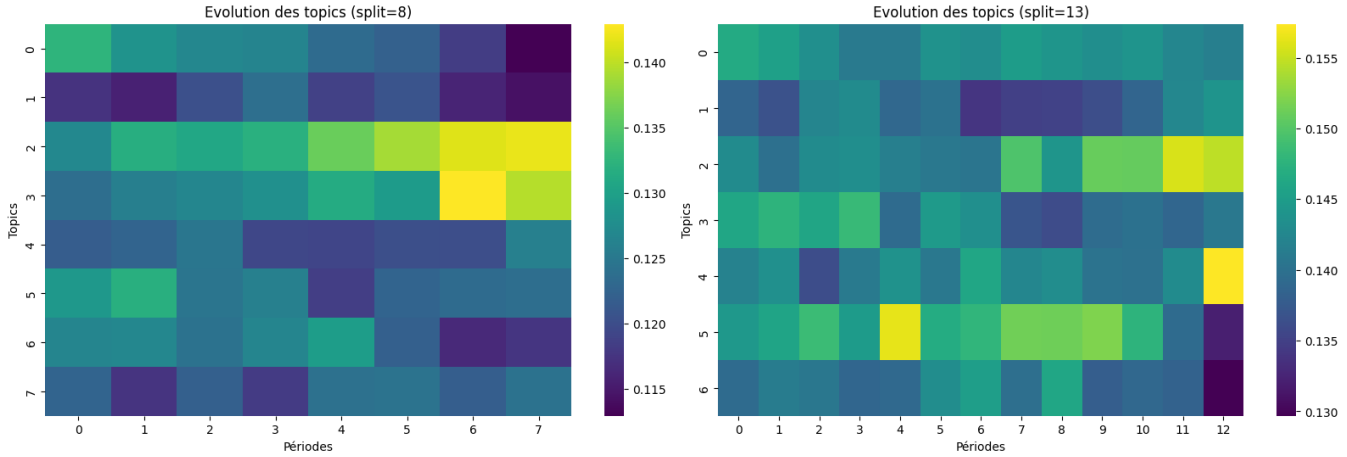
(a) Répartition des 8 périodes DTM



(b) Variation des topics selon la granularité

Figure 4.12: Analyse du découpage temporel optimal du DTM

Deuxièmement, même si la configuration à $T = 13$ présente une légère amélioration de la log-vraisemblance et du score de cohérence $C_{U_{\text{mass}}}$, ce gain reste relativement limité au regard de la perte de sensibilité temporelle observée. Or, dans une perspective diachronique, l'objectif principal du DTM est précisément de détecter les transformations discursives et non uniquement d'optimiser la cohérence statistique interne du modèle. Le choix de retenir $T = 8$ correspond ainsi à un arbitrage assumé entre qualité probabiliste du modèle et capacité à révéler les dynamiques temporelles du corpus médiatique.



(a) Évolution des topics par période ($T = 8$)

(b) Évolution des topics par période ($T = 13$)

Figure 4.13: Comparaison des heatmaps d'évolution thématique selon le découpage temporel

La comparaison des figures Figure 4.13a et Figure 4.13b met en évidence un phénomène de **lissage excessif** lorsque le nombre de périodes temporelles passe à $T = 13$.

Dans la heatmap associée à $T = 8$ (Figure 4.13a), les contrastes entre périodes apparaissent relativement marqués. Certains topics, notamment les Topics 2 et 3, connaissent une progression plus nette au cours des périodes 5 à 7, avec des poids thématiques atteignant environ 0,140. À l'inverse, les Topics 4, 6 et 7 présentent des phases de recul plus visibles, caractérisées par des valeurs plus faibles autour de 0,115. Ces variations produisent des contrastes graphiques plus prononcés et traduisent une meilleure sensibilité du modèle aux fluctuations discursives du corpus.

La situation apparaît différente dans la heatmap obtenue avec $T = 13$ périodes (Figure 4.13b). Les trajectoires thématiques y deviennent plus homogènes et les transitions plus progressives. Bien que l'amplitude globale des poids thématiques reste relativement comparable (0,130 à 0,155), les phases de montée ou de déclin des thèmes s'étalent davantage dans le temps. Les changements discursifs semblent alors plus diffus et les points de bascule deviennent moins clairement identifiables.

Ce phénomène s'explique en grande partie par la diminution du nombre moyen de documents disponibles dans chaque segment temporel. Avec un découpage en 13 périodes appliqué à un corpus de 8,958 articles, chaque tranche contient en moyenne environ 689 documents, contre environ 1,120 documents pour un découpage en 8 périodes. La réduction de la densité documentaire diminue la robustesse statistique propre à chaque intervalle temporel.

Dans ce contexte, le DTM tend naturellement à renforcer la continuité entre périodes successives afin de stabiliser les estimations. Cette régularisation provient directement du processus gaussien de dérive thématique :

$$\beta_k^{(t)} \mid \beta_k^{(t-1)} \sim \mathcal{N}\left(\beta_k^{(t-1)}, \sigma^2 \mathbf{I}\right)$$

introduit par Blei et Lafferty (2006). Lorsque les segments deviennent trop petits, le modèle s'appuie davantage sur cette dépendance temporelle pour compenser le manque d'information disponible dans chaque période. Les trajectoires thématiques produites apparaissent alors artificiellement plus lisses et plus régulières.

Le découpage en $T = 8$ **périodes** est donc retenu pour la suite de l'analyse. Il offre un compromis plus satisfaisant entre stabilité statistique et sensibilité aux évolutions discursives. Les transitions thématiques y apparaissent plus nettes, les variations plus lisibles, et les points d'inflexion médiatiques liés aux événements climatiques extrêmes plus facilement identifiables.

4.4.1.5 Analyse de la mutation sémantique (labels)

Une fois le découpage temporel optimal en huit périodes ($T = 8$) validé, nous procédons à une analyse qualitative plus fine des évolutions sémantiques à travers les labels dynamiques générés par GPT-4o-mini (Table 6.6 en annexe).

4.4.2 Une forte stabilité globale du discours

L'examen des labels associés aux différentes périodes confirme les résultats déjà observés à partir de la divergence de Jensen-Shannon : le discours médiatique de la RTBF sur les catastrophes naturelles présente une forte continuité thématique sur l'ensemble de la période 2008–2021.

La majorité des topics conservent en effet un noyau sémantique relativement stable centré sur les catastrophes naturelles, la gestion des crises, les interventions d'urgence et les impacts environnementaux. Les évolutions observées relèvent davantage d'ajustements progressifs du vocabulaire que de véritables ruptures discursives.

1. Évolutions notables selon les topics.

Le **Topic 0** (*Gestion des catastrophes naturelles*) apparaît comme l'un des sujets les plus évolutifs du corpus. Durant les premières périodes, il reste principalement centré sur la gestion des risques urbains, les intempéries et les évacuations. À partir de 2018–2019, le vocabulaire évolue progressivement vers des enjeux plus explicitement liés au changement climatique. Cette dynamique aboutit, en 2020–2021, à une thématique plus large de *crise socio-économique et environnementale*, probablement influencée à la fois par la pandémie de COVID-19 et par la multiplication des événements climatiques extrêmes.

Le **Topic 2** (*Impact des catastrophes climatiques*) suit également une trajectoire d'élargissement progressif. Initialement focalisé sur les conséquences directes des intempéries entre 2008 et 2013, il intègre progressivement les enjeux liés au changement climatique à partir de 2016–2017. En fin de période, le topic se réorganise autour des risques d'incendie, d'inondation et de gestion environnementale, traduisant une montée en généralité du traitement médiatique des catastrophes naturelles.

Le **Topic 3** (*Gestion des catastrophes et secours*) demeure relativement stable tout au long de la période étudiée. On observe néanmoins une montée progressive du vocabulaire associé aux incendies à partir de 2018–2019, alors que les premières périodes étaient davantage dominées par les inondations et les interventions d'urgence classiques.

Le **Topic 4** (*Gestion des crises et impacts sociaux*) connaît une évolution comparable à celle du Topic 2. D'abord centré sur la gestion immédiate des crises et des perturbations sociales, il évolue progressivement vers des problématiques plus globales liées au changement climatique, à la biodiversité et aux conséquences environnementales de long terme. Cette transformation suggère une montée progressive en abstraction du discours médiatique autour des risques naturels.

Les **Topics 1, 5, 6 et 7** apparaissent quant à eux comme les plus stables sur l'ensemble de la période. Leur vocabulaire reste fortement ancré dans les dimensions locales des crises environnementales, des infrastructures, des services publics et des interventions opérationnelles. Cette faible variabilité confirme l'existence d'un socle discursif relativement permanent dans la couverture médiatique des risques naturels par la RTBF.

2. Une transition progressive du discours médiatique.

Dans l'ensemble, les labels dynamiques mettent en évidence une évolution lente mais réelle du traitement médiatique des catastrophes naturelles. Les premières périodes restent principalement centrées sur la réaction immédiate aux événements — secours, dégâts, évacuations, interventions — tandis que les périodes plus récentes accordent une place croissante aux causes structurelles et systémiques, notamment le changement climatique, la biodiversité ou les ressources environnementales.

Cette transition devient particulièrement visible à partir de 2016–2017, période qui correspond à la montée en puissance des événements climatiques extrêmes dans le corpus (Section 4.3.1), mais également à une médiatisation plus forte des enjeux climatiques globaux dans l'espace public et médiatique belge.

4.4.2.1 Analyse de la mutation sémantique (T_r)

Une fois le découpage temporel optimal en huit périodes ($T = 8$) validé, nous procédons à une analyse plus fine des évolutions lexicales internes à chaque thématique. Cette étape repose sur le calcul du taux de renouvellement sémantique (T_r), dont le seuil critique a été fixé à 25 % pour un voisinage de $n = 25$ termes les plus probables (voir Section 3.9). L'objectif est d'identifier les moments où la structure lexicale d'un topic se transforme suffisamment pour signaler une véritable mutation discursive.

Le Table 4.6 recense l'ensemble des transitions dont le taux de renouvellement dépasse ce seuil critique.

Plusieurs tendances importantes se dégagent de cette analyse.

1.Une rupture majeure isolée.

Le **Topic 0** présente la transition la plus marquée de l'ensemble du corpus, avec un taux de renouvellement atteignant $T_r = 0{,}48$ entre les périodes T_4 et T_5 (correspondant approximativement à 2016–2018). Il s'agit de la seule rupture réellement forte observée dans le modèle.

Les termes entrants — *rapport, pluie, zone, incendie* — traduisent une évolution vers une couverture plus analytique et plus diversifiée des aléas climatiques. À l'inverse, les mots sortants — *responsable, population, autorité, orage, température* — suggèrent un recul du registre institutionnel et météorologique immédiat.

Cette mutation coïncide avec une période d'intensification des événements climatiques extrêmes déjà identifiée dans l'analyse exploratoire (Section 4.3.1), ainsi qu'avec une montée progressive des discours liés au changement climatique dans l'espace médiatique.

2. Des mutations modérées mais régulières.

Les Topics 1, 4, 5, 6 et 7 présentent plusieurs transitions comprises entre $T_r = 0{,}28$ et $T_r = 0{,}36$. Ces valeurs restent supérieures au seuil de rupture fixé à 0,25, mais demeurent nettement inférieures à la rupture observée pour le Topic 0.

Ces variations correspondent davantage à des ajustements lexicaux progressifs qu'à de véritables reconfigurations thématiques.

Le **Topic 1** connaît ainsi trois transitions consécutives ($T_2 \to T_3$, $T_3 \to T_4$ et $T_4 \to T_5$), révélant une évolution graduelle du vocabulaire de gestion des crises vers des dimensions plus territoriales et institutionnelles.

Le **Topic 6** apparaît comme le topic le plus dynamique du corpus, avec quatre transitions successives entre T_3 et T_7. Cette trajectoire traduit une recomposition progressive du vocabulaire environnemental et climatique, particulièrement marquée dans le contexte belge de la seconde moitié de la période étudiée.

Le **Topic 7** présente également plusieurs transitions significatives. On y observe un déplacement progressif du registre des catastrophes internationales — notamment centré au départ sur les typhons et les événements extrêmes mondiaux — vers un vocabulaire plus localisé et institutionnel, où apparaissent des termes comme *pompiers*, *Belgique* ou *capital*.

3. Une forte inertie lexicale globale.

Malgré ces évolutions ponctuelles, la majorité des transitions inter-périodes reste en dessous du seuil de 0,25 et n'apparaît donc pas dans le Table 4.6. Ce résultat confirme que le discours médiatique de la RTBF sur les catastrophes naturelles présente une forte continuité lexicale sur l'ensemble de la période 2008–2021.

Autrement dit, les transformations discursives observées dans le corpus sont principalement progressives et cumulatives plutôt que brutales. À l'exception du Topic 0, aucun sujet ne connaît de rupture sémantique majeure.

Ces résultats apparaissent cohérents avec les labels dynamiques générés par GPT-4o-mini (Table 6.6 en annexe), qui mettent eux aussi en évidence une relative stabilité des grandes familles thématiques, accompagnée d'une montée progressive des enjeux climatiques globaux à partir de la période 2016–2018.

Table 4.6: Analyse des mutations thématiques du DTM (T=8, n=25, seuil Tr > 0,25)

Sujet	Transition	T_r	Stab.	Mots entrants	Mots sortants
0	$T_4 \to T_5$	0.48	13	rapport, pluie, zone, incendie, raison, village, américain, jusqu, voiture	responsable, exemple, population, nbr_million, côté, autorité, orage, température, agriculteur
1	$T_2 \to T_3$	0.32	17	train, afp, pouvoir_être, côté, situation, service, route, total	sud, bon, travail, raison, moyen, ouest, centre, cas
1	$T_3 \to T_4$	0.28	18	rapport, cas, sud, village, raison, ouest, jour	nbr_million, situation, moment, dernier, mois, lieu, route
1	$T_4 \to T_5$	0.28	18	zone, nbr_million, autorité, petit, mois, commun, route	responsable, rapport, tel, température, public, fois, bruxelle
3	$T_4 \to T_5$	0.32	17	zone, village, cause, gouvernement, seul, victime, ouest, centre	animal, raison, vie, température, arbre, bruxelle, projet, chaleur
4	$T_2 \to T_3$	0.28	18	neige, situation, part, voiture, victime, commun, jour	exemple, pouvoir_être, sud, gouvernement, service, important, centre
4	$T_5 \to T_6$	0.36	16	pouvoir_être, incendie, temps, raison, culture, nord, politique, local, suite	responsable, mesure, possible, moment, climat, fois, mois, rue, terrain
5	$T_1 \to T_2$	0.28	18	zone, président, site, partie, service, cours, total	rapport, typhon, pouvoir_être, village, fois, victime, centre
5	$T_2 \to T_3$	0.28	18	typhon, village, temps, situation, petit, fois, victime	mesure, train, zone, site, cours, lieu, maison
6	$T_3 \to T_4$	0.28	18	rapport, sud, petit, grand, météo, cas, enfant	village, forêt, agriculteur, politique, rue, moyen, centre

Suite page suivante…

Sujet	Transition	T_r	Stab.	Mots entrants	Mots sortants
6	$T_4 \to T_5$	0.32	17	zone, centre, village, raison, gouvernement, rue, moyen, maison	tel, côté, capital, vie, seul, public, grand, bruxelle
6	$T_5 \to T_6$	0.36	16	tel, incendie, vie, seul, agriculteur, partie, politique, bruxelle, important	etat, nbr_million, autorité, belge, température, bas, rue, victime, belgique
6	$T_6 \to T_7$	0.32	17	pluie, nbr_million, autorité, nord, précipitation, température, rue, victime	rapport, résultat, femme, culture, question, seul, politique, partie
7	$T_1 \to T_2$	0.28	18	rapport, afp, population, terre, orage, américain, route	zone, typhon, autorité, nombre, ouest, centre, vol
7	$T_4 \to T_5$	0.28	18	zone, nbr_million, village, niveau, partie, capital, route	nombreux, mesure, vie, risque, petit, politique, projet
7	$T_5 \to T_6$	0.28	18	nombreux, pluie, travail, vie, petit, politique, ouest	population, village, niveau, belge, bâtiment, capital, local
7	$T_6 \to T_7$	0.28	18	village, niveau, président, pompier, capital, local, important	nombreux, présent, etat, nbr_million, travail, étude, lieu

4.4.3 Modèle BERTopic

4.4.3.1 Optimisation des hyperparamètres

L'optimisation est conduite via Optuna sur 100 essais avec un budget temps de deux heures. Le pruner *MedianPruner* élimine les configurations sous-optimales dès les premiers essais, concentrant le budget sur les régions prometteuses de l'espace de recherche. Le modèle retenu maximise le score composite $\mathcal{S}$ (Équation 3.5) et atteint un score final de **0,4948**. Les hyperparamètres optimaux sont consignés dans le Table 4.7.

Table 4.7: Hyperparamètres optimaux du modèle BERTopic

Composant	Paramètre	Valeur optimale
UMAP	`n_neighbors`	5
UMAP	`n_components`	17
UMAP	`min_dist`	0.00609
UMAP	`metric`	cosine
HDBSCAN	`min_cluster_size`	60
HDBSCAN	`min_samples`	11
HDBSCAN	`cluster_selection_epsilon`	0.1308
HDBSCAN	`cluster_selection_method`	eom
Vectoriseur	`ngram_max`	3
Vectoriseur	`min_df`	3
Vectoriseur	`max_df`	0.9066

4.4.3.2 Traitement des outliers

Le modèle initial identifie **34 topics** avec un taux d'outliers de 21.92%. Trois stratégies de réduction sont comparées (Table 4.8) : par embeddings, par c-TF-IDF, et par distributions.

Table 4.8: Comparaison des stratégies de réduction des outliers

Stratégie	Topics	Outliers	Cohérence c_v	Δ cohérence
Modèle initial	34	21.92%	0.6432	—
Réduction par embeddings	34	0.01%	0.6366	−0.0066
Réduction par c-TF-IDF	34	14.69%	—	—
Réduction par distributions	34	3.92%	—	—

La stratégie par embeddings est retenue : elle réduit le taux d'outliers de 21,92,% à 0,01,% tout en conservant les 34 topics. La légère dégradation de cohérence ($\Delta C_v = -0{,}0066$) s'explique par la nature même de la mesure : la cohérence C_v évalue la qualité sémantique des distributions lexicales des topics (β_k), qui restent inchangées lors de la réaffectation des outliers. Ce faible écart confirme donc la robustesse sémantique des topics eux-mêmes, indépendamment de l'assignation documentaire.

Il convient néanmoins de noter une limite inhérente à cette stratégie : les outliers sont par construction des documents que HDBSCAN a jugés trop atypiques pour appartenir à un cluster. Leur réaffectation par proximité cosinus les place dans le topic *le plus proche*, non nécessairement le *plus pertinent*. Ce mécanisme introduit un biais d'assignation modéré, acceptable au regard du gain de couverture documentaire (+21,91 points de pourcentage), mais qui invite à interpréter avec prudence les résultats portant spécifiquement sur les documents réaffectés. Les analyses thématiques présentées dans la suite s'appuient sur l'ensemble du corpus réaffecté, en gardant à l'esprit cette réserve méthodologique.

4.4.3.3 Analyse de stabilité

4.4.3.3.1 Approche par sous-échantillonnage

En reprenant les hyperparamètres de Table 4.7, nous évaluons la stabilité du modèle sur **1 000 sous-échantillons aléatoires** constitués de 50% des documents (sans remise). Trois indicateurs sont observés par réplicat : cohérence c_v, nombre de topics, et taux d'outliers (Figure 4.14). Les résultats indiquent une **bonne robustesse globale** :

- **Cohérence** : moyenne de 0.65 ($std \approx .05$ écart-type du score de cohérence moyen), distribution quasi gaussienne sans effondrement, confirmant la stabilité sémantique du modèle sous perturbation.
- **Nombre de topics** : moyenne de 16.82 ($std \approx 5.29$ écart-type de la moyenne de topics), reflétant une sensibilité structurelle attendue du pipeline UMAP–HDBSCAN aux variations locales des données. La majorité des réplicats se concentre entre 15 et 22 topics.
- **Outliers** : taux moyen d'environ 21%, sans pics extrêmes ni asymétrie marquée, indiquant l'absence de dégradation importante sous perturbation du corpus.

4.4.3.3.2 Importance des hyperparamètres

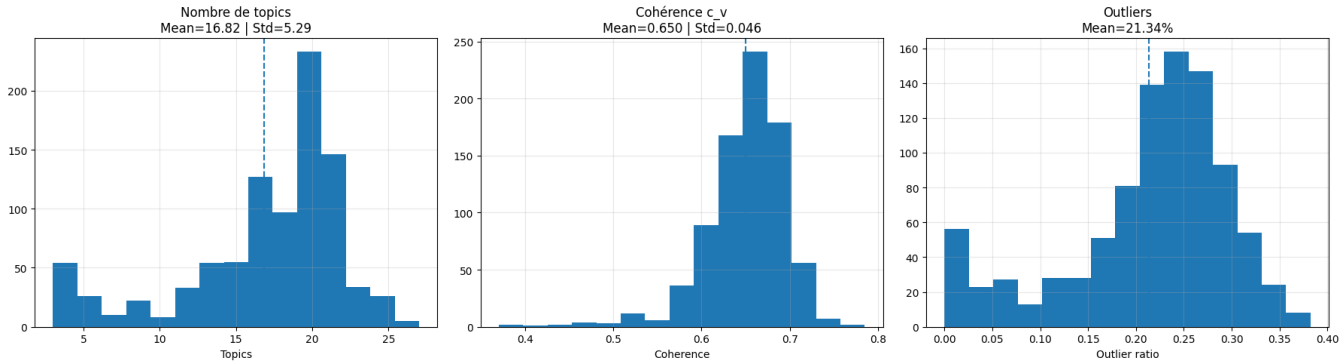
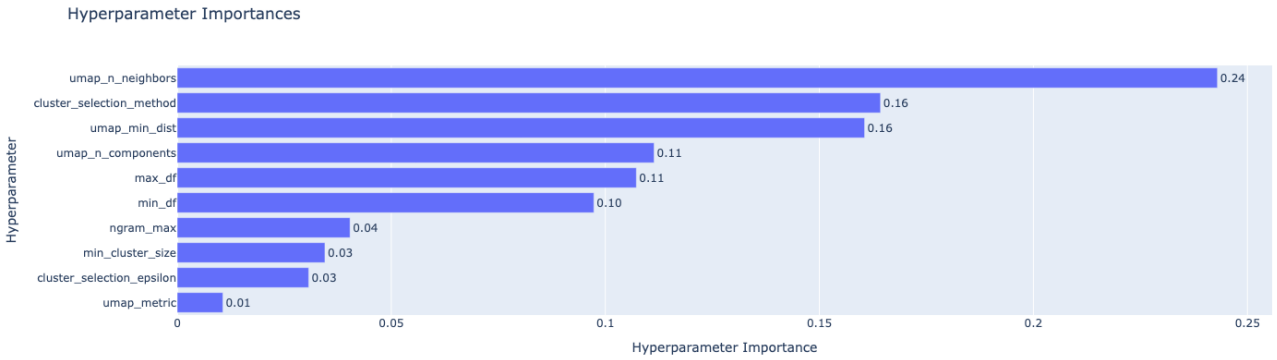
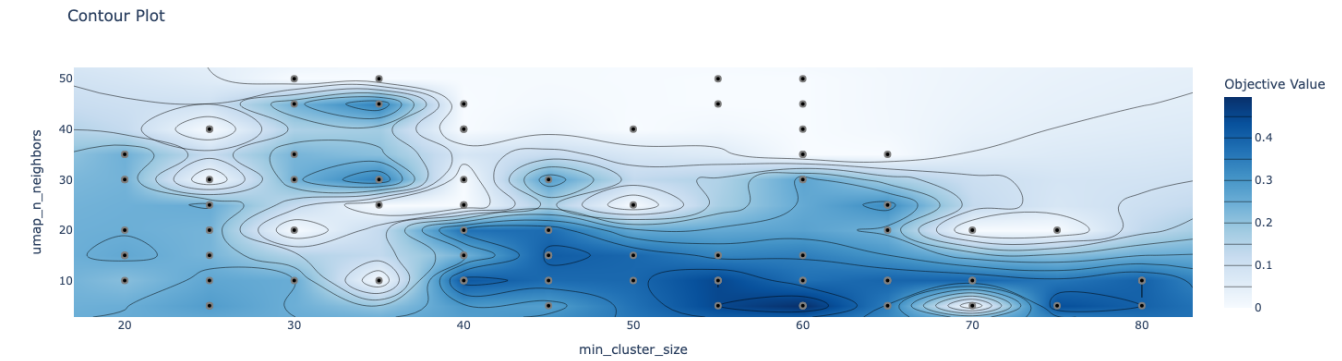


Figure 4.14: Analyse de stabilité du modèle BERTopic sur 1000 sous-échantillons



(a) Importance des hyperparamètres



(b) Graphique de contour

Figure 4.15: Analyse de l'importance des hyperparamètres et interactions.

L'analyse d'importance (Figure 4.15a) révèle que les paramètres UMAP dominent les performances du modèle. Le paramètre `umap_n_neighbors` est le facteur le plus influent ($\approx$ 0.24), suivi de `cluster_selection_method` et `umap_min_dist` ($\approx$ 0.16 chacun). Les paramètres `umap_n_components`, `max_df` et `min_df` présentent une influence intermédiaire, tandis que `umap_metric`, `cluster_selection_epsilon` et `min_cluster_size` montrent une contribution plus faible.

Le graphique de contour (Figure 4.15b) met en évidence une interaction notable entre `min_cluster_size` et `umap_n_neighbors` : les meilleures performances se concentrent pour de **faibles valeurs de `n_neighbors`**

(5–15) associées à des **valeurs intermédiaires à élevées de `min_cluster_size`** (45–80), favorisant des clusters denses et bien séparés. Des valeurs élevées de `n_neighbors` produisent des représentations trop globales qui réduisent la séparation entre thèmes.

4.4.3.3.3 Similarité de Jaccard

L'évaluation de la stabilité des partitions à l'aide de l'indice de Jaccard (Ben-Hur, Elisseeff, et Biermann (2002) ; Jaccard (1901)) met en évidence une stabilité modérée du modèle BERTopic entre les différentes réplications réalisées avec des graines aléatoires distinctes. Les scores observés varient entre 0.157 et 0.425, avec une moyenne de $\bar{J} = 0.236$ et un écart-type de $\sigma = 0.076$. Ces valeurs restent inférieures au seuil généralement considéré comme satisfaisant pour une forte stabilité des partitions ($J > 0.5$).

Cette sensibilité apparaît néanmoins cohérente avec le fonctionnement attendu du pipeline UMAP–HDBSCAN. La faible valeur du paramètre `n_neighbors` retenue lors de l'optimisation ($n\_neighbors = 5$) privilégie une représentation très locale de l'espace sémantique, ce qui rend la projection UMAP particulièrement sensible aux conditions d'initialisation (McInnes et al. (2018)). Ce phénomène est renforcé par la dimension relativement élevée de l'espace réduit (`n_components = 17`), qui augmente le nombre de configurations potentielles exploitables par HDBSCAN lors de la phase de clustering.

Plusieurs éléments permettent toutefois de relativiser cette instabilité structurelle.

Premièrement, la cohérence sémantique des topics reste remarquablement stable entre les différentes réplications, avec un score moyen de $C_v = 0.65 \pm 0.05$. Cela suggère que, malgré certaines variations dans la répartition exacte des documents, les thématiques extraites conservent une cohérence lexicale et sémantique globalement robuste.

Deuxièmement, certaines paires de réplications atteignent des niveaux de similarité relativement élevés, notamment entre `seed1` et `seed2` ($J = 0.425$). Ce résultat indique l'existence d'une structure thématique récurrente partiellement stable à travers plusieurs exécutions du modèle.

Troisièmement, le modèle final utilisé dans ce mémoire est entraîné avec une graine fixe (`random_state = 42`), ce qui garantit la reproductibilité exacte des résultats présentés dans l'analyse.

Dans l'ensemble, ces résultats suggèrent que BERTopic parvient à capturer des tendances thématiques globalement robustes sur le plan sémantique, même si l'affectation précise des documents aux clusters demeure sensible aux conditions d'initialisation. Cette variabilité constitue une limite connue des approches neuronales de *topic modeling*, en particulier lorsqu'elles reposent sur des étapes successives de réduction de dimensionnalité et de clustering non supervisé (Grootendorst (2022)).

4.4.3.3.4 Analyse du modèle finale.

Le graphique (Figure 4.16) met en évidence une structuration thématique relativement nette du corpus dans l'espace bidimensionnel produit par la réduction de dimensionnalité. Trois ensembles principaux peuvent être distingués.

Le premier cluster, situé dans la partie inférieure gauche de la projection, apparaît comme le plus dense et le plus volumineux. Il regroupe plusieurs topics importants, notamment le Topic 0, représenté par le cercle le plus large et donc dominant en nombre de documents. Autour de ce noyau central gravitent plusieurs topics secondaires de taille plus réduite mais sémantiquement proches.

Le second ensemble, localisé dans la zone centrale du graphique, présente une structure plus diffuse et moins compacte. Il rassemble principalement des topics de taille intermédiaire, suggérant des relations sémantiques plus variées et moins fortement polarisées.

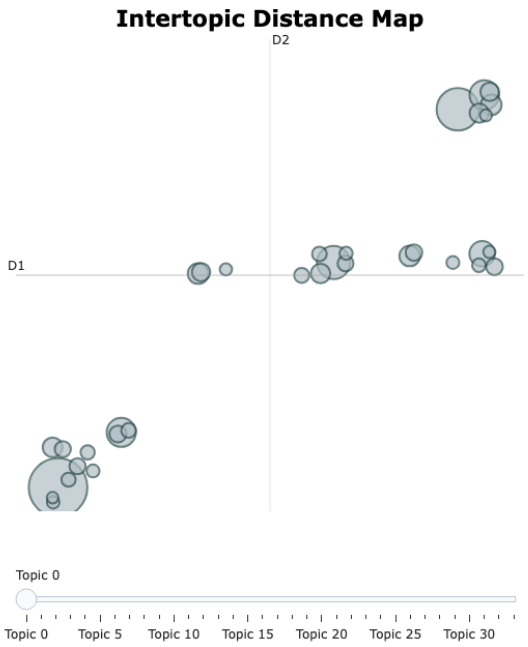


Figure 4.16: Visualisation du modèle BERTopic

Enfin, le troisième cluster, situé dans la partie supérieure droite, est le plus restreint en termes de densité. Il regroupe des topics très proches les uns des autres sur le plan sémantique, formant un ensemble relativement homogène.

La procédure de labellisation automatique réalisée via l'API OpenAI conduit aux résultats présentés dans le tableau 7.1 de la Section 6. L'analyse de ces labels permet de regrouper les 34 topics identifiés par BERTopic en sept grandes catégories thématiques.

Les trois catégories dominantes concentrent à elles seules 25 des 34 topics extraits par le modèle.

La catégorie **Climat et phénomènes météorologiques** ($K = 8$ topics) couvre un spectre géographique particulièrement large. Elle rassemble aussi bien des événements météorologiques européens liés à la chaleur, aux sécheresses ou aux orages, que des catastrophes tropicales majeures à l'échelle internationale, comme les ouragans en Haïti et au Mexique, les typhons aux Philippines ou au Japon, ou encore les cyclones au Bangladesh et à Vanuatu.

La catégorie **Inondations et gestion hydrique** ($K = 9$ topics) apparaît comme la plus fragmentée du corpus. Elle distingue plusieurs niveaux de traitement médiatique : les interventions locales des pompiers belges (Topics 4 et 22), la gestion communale wallonne (Topic 5), mais aussi des crises humanitaires internationales de grande ampleur, comme les inondations au Pakistan (Topic 20) ou les risques hydrogéologiques en Indonésie (Topic 27).

La catégorie **Société, économie et territoires** ($K = 8$ topics) révèle quant à elle la dimension systémique des aléas climatiques. Les thèmes identifiés concernent notamment les coûts assurantiels des catastrophes naturelles (Topic 26), les conséquences sur l'agriculture et les rendements agricoles (Topic 1), ou encore les mesures de sécurisation des espaces publics bruxellois lors d'épisodes de vents violents (Topic 29).

Les quatre catégories restantes regroupent les neuf topics résiduels : - **Incendies et catastrophes naturelles** ($K = 4$) ; - **Santé et crises humanitaires** ($K = 2$) ; - **Transport et infrastructures** ($K = 2$) ; - **Politique et gouvernance** ($K = 1$).

La présence de certains topics plus périphériques, comme le Topic 16 centré sur la COVID-19 ou le Topic 2 consacré aux négociations gouvernementales, témoigne d'une certaine contamination thématique du corpus. Ces sujets apparaissent généralement dans des articles associant simultanément crises sanitaires, enjeux politiques et événements climatiques.

Plus largement, l'ensemble des labels obtenus met surtout en évidence une forte présence des conséquences des aléas dans le discours médiatique. Les dimensions humaines, économiques, sociales ou environnementales des catastrophes occupent une place centrale dans les topics identifiés. À l'inverse, les thématiques liées aux causes structurelles des phénomènes climatiques restent beaucoup plus marginales. Ce constat rejoint les observations formulées par Rouquette et Bihay (2022), selon lesquelles les médias tendent davantage à privilégier le récit de l'événement et de ses impacts immédiats qu'une explication approfondie des mécanismes structurels à l'origine des catastrophes.

4.4.4 BERTopic dynamique

4.4.4.1 Découpage temporel et comparabilité

Conformément aux principes de comparabilité établis dans la section Section 3.7, le même découpage en **huit périodes** ($T = 8$) que celui retenu pour le DTM est appliqué au BERTopic dynamique. Ce choix garantit que les évolutions thématiques observées dans les deux modèles sont directement comparables sans biais lié à la segmentation chronologique.

4.4.4.2 Évolution des topics au fil du temps

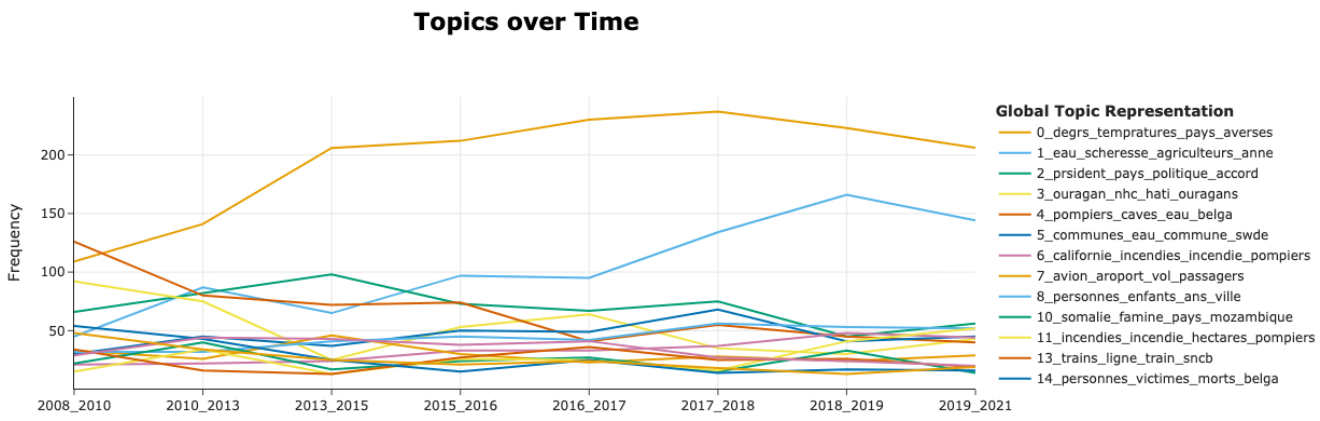


Figure 4.17: Topics over Time — BERTopic dynamique ($T = 8$)

L'analyse de la Figure 4.17 révèle plusieurs dynamiques distinctes parmi les quinze premiers topics représentés.

Topic dominant et stable. Le Topic 0 (*degrés, températures, pays, averses*) constitue de loin la thématique la plus fréquente sur l'ensemble de la période, avec une fréquence oscillant entre 120 et 230 articles par période. Sa trajectoire ascendante de 2008 à 2017–2018, suivie d'une légère stabilisation, reflète la montée en puissance progressive de la couverture météorologique et climatique dans le corpus RTBF.

Émergence tardive. Le Topic 1 (*eau, sécheresse, agriculteurs*) présente la dynamique la plus spectaculaire : quasi inexistant avant 2015, il connaît une croissance exponentielle à partir de 2016–2017 pour atteindre

un pic de ≈ 165 articles en 2018–2019, avant de se stabiliser autour de 145 en 2019–2021. Cette trajectoire coïncide précisément avec la succession de sécheresses et de vagues de chaleur qui ont frappé la Belgique et l'Europe à partir de 2017, confirmant la sensibilité du modèle aux événements climatiques réels.

Topic en déclin. Le Topic 4 (*pompiers, caves, eau, belga*) présente la trajectoire inverse : très présent en 2008–2010 (≈ 120 articles), il décroît progressivement pour atteindre un plancher en 2013–2015, avant de se stabiliser autour de 50 articles par période. Ce recul relatif ne signifie pas une disparition du sujet mais une dilution dans un corpus en forte croissance.

Topics stables à faible variance. La majorité des topics restants — Topics 2, 3, 5, 6, 7, 8, 10, 11, 13 et 14 — présentent des trajectoires relativement stables, oscillant entre 25 et 100 articles par période sans tendance marquée. Cette stabilité confirme l'existence d'un socle thématique permanent dans la couverture médiatique des risques naturels par la RTBF.

4.4.4.3 Analyse des labels dynamiques

L'examen du Table 6.2 en annexe, qui recense les labels produits par GPT-4o-mini pour chacun des 34 topics sur les 8 périodes, appelle plusieurs observations.

Contamination thématique persistante. Une fraction significative des topics — notamment les Topics 0, 2, 7, 8, 9, 10, 13–17, 21, 24–26, 28–30 — produit des labels dominés par les *violences urbaines* sur l'ensemble des périodes. Ce phénomène s'explique par la nature généraliste du corpus RTBFINFO, qui couvre l'actualité belge et internationale dans son ensemble. Ces topics captent des articles de faits divers, de politique intérieure et de sécurité publique qui partagent un vocabulaire partiellement commun avec les articles sur les catastrophes naturelles (urgence, intervention, autorités, victimes).

Topics à ancrage climatique clair. Plusieurs topics maintiennent un ancrage thématique cohérent et pertinent tout au long de la période :

- Le **Topic 4** (*Inondations et interventions des pompiers*) reste centré sur les inondations belges et les interventions d'urgence de 2008 à 2021, avec une évolution sémantique vers les *interventions d'urgence en contexte climatique* en 2019–2021, témoignant d'une montée en généralité du discours.
- Le **Topic 6** (*Incendies*) suit une trajectoire géographique intéressante : centré sur les incendies urbains belges en début de période, il évolue progressivement vers les incendies de forêt en Californie à partir de 2015–2016, reflétant l'internationalisation de la couverture des feux de forêt dans le corpus RTBF.
- Le **Topic 5** (*Gestion des crises hydriques*) présente une évolution cohérente, passant des inondations et crises sociales en Wallonie (2008–2015) vers la *gestion des crises liées à l'eau* et la *crise de l'eau et gestion communale* en fin de période, reflétant une institutionnalisation progressive du discours sur la gestion hydrique.

Évolutions géographiques notables. Le Topic 20 (*Pakistan*) illustre une dynamique particulière : centré sur les inondations catastrophiques au Pakistan en 2008–2013, il dérive ensuite vers des conflits sociaux et politiques plus généraux, suggérant que les mots-clés initialement liés aux inondations pakistanaises ont été progressivement remplacés par un vocabulaire de crise géopolitique.

Finalement. Les labels dynamiques du BERTopic confirment le constat établi par le DTM : le discours médiatique de la RTBF sur les catastrophes naturelles présente une forte inertie thématique, avec néanmoins des signaux d'évolution clairs autour de la sécheresse et des incendies de forêt à partir de 2016–2018. La présence importante de topics *hors périmètre* (violences urbaines, politique) constitue une limite inhérente à l'approche non supervisée sur un corpus généraliste, et plaide pour une comparaison systématique avec les résultats du DTM, dont le périmètre lexical est mieux contrôlé par la procédure de tri présentée en Section 3.3. Aucune mention du taux de renouvellement T_r n'est faite ici du au fait que pour toutes les thématiques analysées, aucune n'a dépasser le seuil fixé ($T_r > 0,25$).

5 Conclusion

5.1 Synthèse de la démarche

Ce mémoire s'est donné pour ambition de répondre à une question à la fois méthodologique et substantielle : comment choisir et mettre en œuvre une méthode de modélisation de thématiques adaptée à l'étude de la sensibilisation aux risques de catastrophes naturelles dans un corpus journalistique de grande taille? Pour y répondre, nous avons adopté une démarche en deux temps : d'abord construire un corpus ciblé à partir des archives de la RTBF (2008–2021), ensuite comparer quatre modèles de *topic modelling* — LDA, DTM, BERTopic et BERTopic dynamique — selon des critères de cohérence sémantique, de stabilité lexicale et de capacité à capturer des évolutions diachroniques.

5.2 Atteinte des objectifs

Les quatre objectifs définis dans le cahier des charges ont été atteints.

Objectif 1 : Comprendre les spécificités de la modélisation dynamique probabiliste. Le travail d'optimisation du LDA et du DTM a permis d'en explorer les mécanismes en profondeur. Les analyses de sensibilité par sous-échantillonnage et variation de graine aléatoire ont mis en évidence une stabilité modérée ($\bar{J} = 0{,}487$ après filtrage), jugée satisfaisante pour un corpus journalistique généraliste. L'optimisation du découpage temporel ($T = 8$, retenu sur la base du pic de divergence JS à $\overline{JS} = 0{,}053$) et du paramètre `rm_top` ($n = 10$) constituent des contributions méthodologiques concrètes, directement transférables à d'autres corpus.

Objectif 2 : Fournir un panorama des méthodologies de l'état de l'art. La revue de littérature couvre de manière exhaustive les approches classiques (LSA, NMF, pLSA, LDA, DTM) et neuronales (BERTopic, Top2Vec, BERTopic dynamique), en illustrant certains modèles sur un mini-corpus pédagogique commun. Cette progression logique — de l'algèbre linéaire vers les Transformers — constitue un outil de référence utile pour tout chercheur abordant ce domaine.

Objectif 3 : Synthétiser l'évolution de la structure des topics liés aux catastrophes. L'ensemble des modèles convergent vers un même constat : le discours médiatique de la RTBF présente une forte inertie thématique sur la période 2008–2021, avec une inflexion progressive amorcée autour de 2016–2018. Cette transition, bien que lente, est réelle et cohérente entre les approches probabilistes et neuronales.

Objectif 4 : Comparer les algorithmes et illustrer leurs capacités. L'évaluation croisée (C_v, $C_{U_{\text{mass}}}$, indice de Jaccard, divergence JS, taux de renouvellement T_r) permet une comparaison rigoureuse et multidimensionnelle des quatre modèles.

Pour un corpus journalistique francophone traitant de thématiques environnementales complexes, les modèles probabilistes interprétables conservent un avantage décisif sur les approches neuronales récentes, comme le soulignent Tanguy et al. (2025) : la transparence algorithmique et la possibilité d'un dialogue avec des experts du domaine priment sur la performance brute.

5.3 Comparaison des modèles

Le tableau suivant synthétise les performances et caractéristiques des quatre modèles sur notre corpus :

Table 5.1: Comparaison synthétique des quatre modèles sur le corpus RTBF

Critère	LDA	DTM	BERTopic	BERTopic dyn.
Cohérence C_v	0,4196	—	**0,6366**	**0,6366**
Stabilité $\bar{J}$	0,487	—	0,236	—
Outliers	—	—	0,01 %	—
Topics identifiés	8	8	**34**	**34**
Interprétabilité	Forte	Forte	Modérée	Modérée
Contamination thématique	Modérée	Modérée	Élevée	Élevée
Dimension temporelle	Non	Native	Non	Native
Reproductibilité	Forte	Forte	Modérée	Modérée

Trois enseignements principaux ressortent de cette comparaison.

BERTopic produit la cohérence sémantique la plus élevée, mais au prix d'une granularité thématique beaucoup plus fine (34 topics contre 8 pour le LDA et le DTM) et d'une contamination thématique plus importante — notamment par des sujets de faits divers et de politique intérieure qui partagent le vocabulaire de l'urgence avec les catastrophes naturelles (victimes, autorités, intervention). Cette contamination est inhérente à l'approche non supervisée appliquée à un corpus généraliste : BERTopic ne dispose pas d'un mécanisme de tri lexical a priori comme celui mis en place lors du tri de niveau 2. En d'autres termes, là où le LDA et le DTM reçoivent en entrée un corpus déjà filtré par un dictionnaire d'aléas climatiques — et ne modélisent donc que des articles thématiquement ciblés —, BERTopic opère directement sur les similarités sémantiques brutes entre documents : tout article partageant un vocabulaire contextuel proche de celui des catastrophes naturelles (victimes, autorités, intervention, zone) peut se retrouver regroupé dans un même cluster, indépendamment de son sujet réel.

Les modèles probabilistes (LDA, DTM) offrent une interprétabilité supérieure et un périmètre thématique mieux contrôlé, grâce précisément au tri lexical préalable qui garantit que les topics émergent d'un sous-corpus sémantiquement ciblé. En contrepartie, ils dépendent d'une hypothèse de sac de mots qui ignore les relations sémantiques fines, et leur stabilité reste modérée sur un corpus journalistique à vocabulaire riche et varié.

Le DTM et le BERTopic dynamique sont les seuls à capturer l'évolution temporelle nativement. Le DTM le fait par un processus gaussien de dérive lexicale, garantissant une continuité entre périodes ; le BERTopic dynamique recalcule les représentations c-TF-IDF par période sans modifier les clusters, ce qui le rend plus sensible aux variations locales. Les deux approches convergent néanmoins vers les mêmes constats : forte inertie globale et inflexion progressive à partir de 2016–2018.

5.4 Discussion thématique

5.4.1 Une couverture structurellement centrée sur les conséquences

L'un des résultats les plus saillants de cette analyse est la prédominance écrasante des thématiques relevant de la phase d'urgence et de réponse immédiate dans le discours médiatique de la RTBF, au détriment des thématiques relevant de la prévention et des facteurs de risque structurels. La (Figure 5.1) illustre cette asymétrie pour les quatre modèles sur l'ensemble de la période 2008–2021.

Par phase d'urgence (conséquences, réactions immédiates), nous entendons l'ensemble des thématiques centrées sur la gestion immédiate de la catastrophe : interventions des pompiers, évacuations, bilans humains et matériels, opérations de secours. Ces registres réactifs et événementiels dominent structurellement la couverture médiatique de la RTBF sur l'ensemble la période étudiée. Pour ce qui est des facteurs de risque structurels et de prévention (causes structurelles), nous regroupons les thématiques portant sur les mécanismes physiques et climatiques qui expliquent l'intensification des aléas (dérèglement climatique, artificialisation des sols, déficit hydrique chronique), ainsi que sur les politiques d'anticipation, de gestion préventive et d'adaptation à long terme. Ces dimensions restent marginales et instables dans le corpus, confirmant que la RTBF traite davantage les catastrophes comme des événements ponctuels que comme des manifestations d'un risque structurel croissant.

Ce résultat confirme les études sur la manière dont les médias traitent les risques environnementaux. Vasterman (2005) montre que les catastrophes provoquent un pic d'attention soudain, centré sur l'événement et ses conséquences immédiates. De plus, Boykoff et Boykoff (2007) soulignent que les journalistes privilégient le spectaculaire, l'émotion et la nouveauté. Les médias mettent donc plus facilement en avant les crises soudaines et visuelles que les problèmes lents et structurels.

Nos résultats suggèrent que le manque de visibilité des causes climatiques à la RTBF ne dépend pas seulement de ses propres choix éditoriaux. Il s'explique plus largement par le fonctionnement global des médias modernes. Cette idée rejoint le constat de Rouquette et Bihay (2022) : les risques qui évoluent lentement peinent à être médiatisés, car leur progression lente manque du spectaculaire nécessaire pour faire les gros titres.

Enfin, la lente évolution observée dans notre corpus correspond aux conclusions de Meier et Eskjaer (2024). Leurs travaux montrent que l'intégration des causes climatiques et des stratégies d'adaptation dans les médias prend du temps : elle ne s'installe que très progressivement, après plusieurs décennies de couverture centrée uniquement sur les catastrophes immédiates.

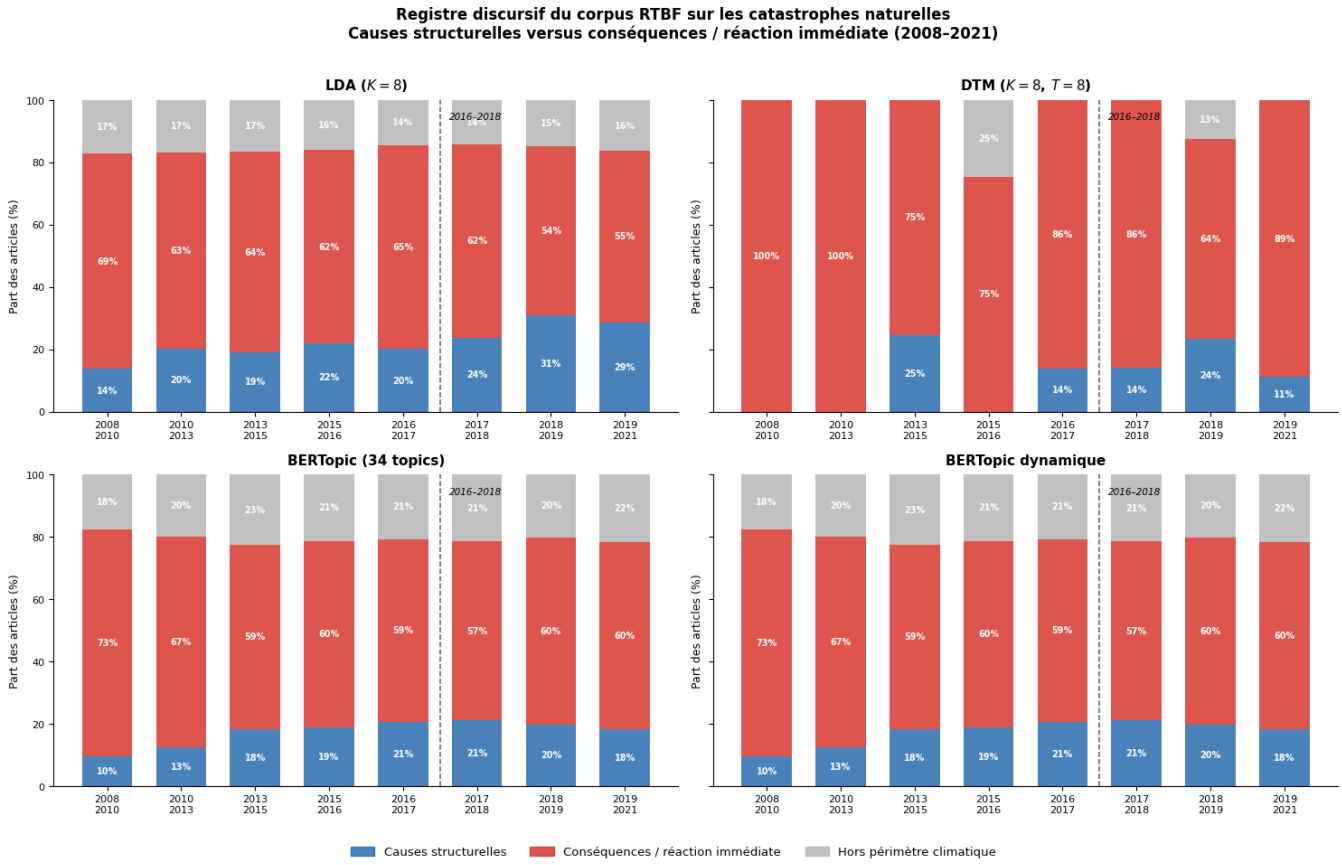


Figure 5.1: Évolution de la proportion d'articles classés selon leur registre discursif (causes structurelles versus conséquences / réaction immédiate) pour les quatre modèles (2008–2021). La ligne pointillée marque l'inflexion de 2016–2018.

5.4.2 Une dimension internationale dominante

L'analyse du contexte de diffusion (Section 4.3.3) montre que la catégorie « MONDE » occupe une place prépondérante dans le corpus sélectionné. Cette domination de l'international se retrouve également dans les thématiques identifiées par les modèles (Figure 5.2) : typhons au Japon (2011), inondations au Pakistan (2010), incendies en Californie (2019) figurent parmi les thématiques les mieux identifiées par les deux familles de modèles. La dimension belge et wallonne, bien que présente — notamment dans les topics consacrés aux interventions des pompiers et à la gestion communale des inondations — reste quantitativement minoritaire.

Ce constat rejoint les analyses de Joye (2010) sur le phénomène de souffrance à distance (**distant suffering**). L'attention médiatique tend en effet à se concentrer sur des catastrophes spectaculaires et émotionnellement marquantes, même lorsqu'elles se produisent loin du public concerné. Les événements tropicaux majeurs — cyclones de forte intensité, tsunamis ou typhons meurtriers — bénéficient ainsi d'une forte visibilité. Ce traitement asymétrique se fait au détriment de risques plus diffus ou progressifs affectant directement le territoire belge, tels que les sécheresses agricoles, les tassements de terrain ou les inondations récurrentes.

Ces observations mettent en évidence une distribution inégale de l'attention médiatique selon les espaces géographiques concernés. Cette asymétrie soulève une question importante pour la sensibilisation et la prévention : dans quelle mesure une exposition majoritairement centrée sur des drames lointains favorise-t-elle la préparation des populations face aux risques locaux ? Les travaux de Schmidt, Ivanova, et Schäfer

(2013) montrent à cet égard que les événements extrêmes de proximité restent le principal déclencheur d'une hausse de l'attention médiatique nationale. Cette dynamique se vérifie dans notre corpus, où les inflexions thématiques les plus nettes coïncident avec des crises directement vécues en Belgique : c'est le cas des inondations à Liège en 2018 (avec un pic d'articles visible à la Figure 4.1b) et des épisodes de sécheresse observés entre 2017 et 2020, qui marquent une augmentation progressive du nombre d'articles liés à ces aléas (Figure 4.6).

Dans ce graphique (Figure 5.2), la catégorie non classée en gris représente toutes les autres catégories (Figure 4.7).

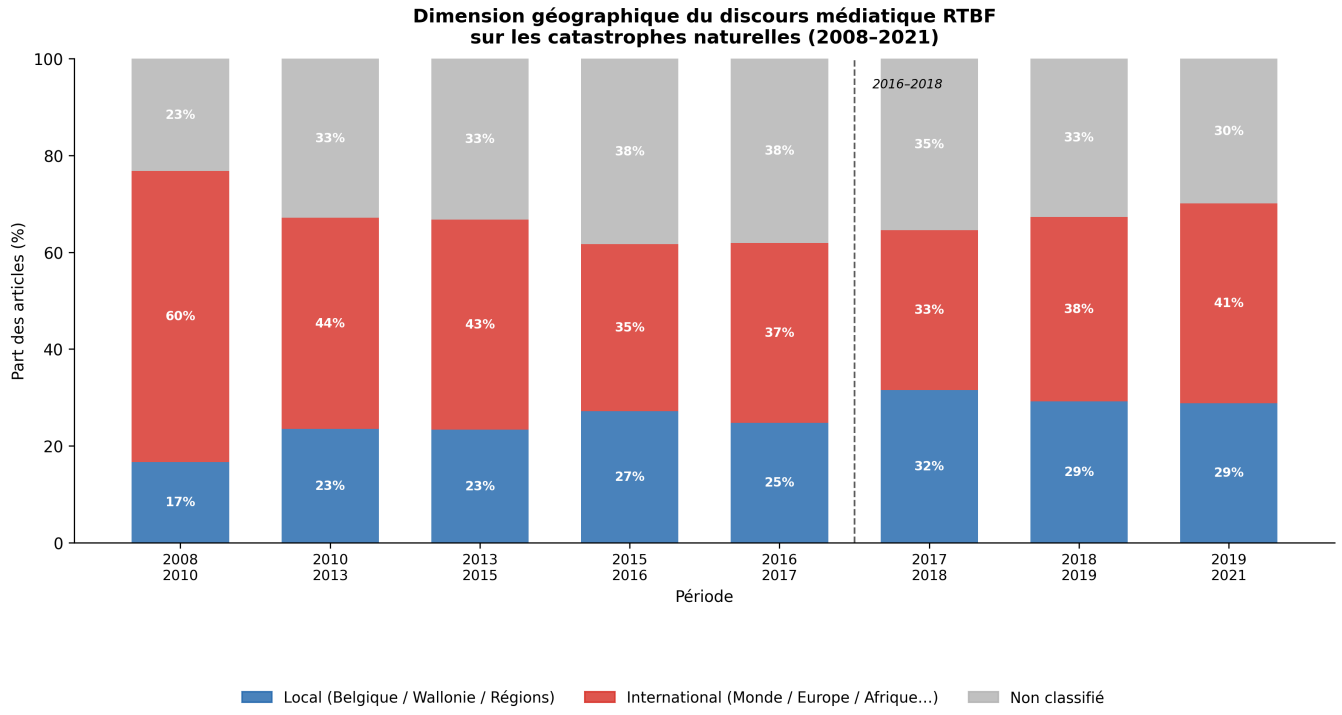


Figure 5.2: Répartition géographique des articles du corpus RTBF consacrés aux catastrophes naturelles, entre couverture locale (Belgique, Wallonie, Régions) et internationale (Monde, Europe, Afrique…), par période temporelle (2008–2021).

5.4.3 Contextualisation chronologique des événements majeurs

Afin de mieux interpréter les évolutions thématiques extraites, il convient de replacer le corpus dans sa chronologie historique. Le tableau Table 5.2 synthétise les principaux événements climatiques extrêmes survenus entre 2008 et 2021 susceptibles d'avoir influencé la couverture journalistique de la RTBF.

Cette contextualisation chronologique permet notamment de mieux comprendre les ruptures discursives identifiées par les modèles dynamiques. Les périodes de forte évolution lexicale ou de montée rapide de certains topics coïncident fréquemment avec la survenue d'événements climatiques extrêmes bénéficiant d'une forte visibilité médiatique, qu'ils soient internationaux ou directement observés en Belgique. Les sections suivantes analysent plus en détail la manière dont ces événements semblent avoir influencé la structuration progressive du discours médiatique sur les risques naturels et climatiques.

Table 5.2: Principales catastrophes naturelles susceptibles d'avoir influencé le corpus RTBF (2008–2021)

Année	Événement	Localisation	Impact et pertinence corpus
2008	**Cyclone Nargis**	Myanmar	138 000 morts, couverture internationale majeure
2010	**Séisme et inondations**	Haïti / Pakistan	Double catastrophe, pics d'articles dans le corpus
2011	**Tsunami et Fukushima**	Japon	Couverture exceptionnelle, topics infrastructures
2013	**Typhon Haiyan**	Philippines	Topic 9 BERTopic dominant, 6 000 morts
2015	**Cyclone Pam**	Vanuatu	Topic LDA 1 — catastrophes et impacts humains
2017	**Sécheresse Belgique**	Belgique	Début de la montée du Topic 1 BERTopic
2018	**Inondations Liège**	Belgique	Pic d'articles 2018, rupture DTM $T_r = 0{,}48$
2018	**Typhon Mangkhut**	Philippines	Topic LDA 1 — couverture internationale
2019	**Canicule européenne**	Europe / Belgique	Montée des aléas thermiques dans le corpus
2019	**Incendies Californie**	États-Unis	Topic 6 BERTopic — incendies forestiers
2020	**Incendies Australie**	Australie	Topic 11 BERTopic — Nouvelle-Galles-du-Sud
2021	**Inondations Liège**	Belgique	Événement le plus meurtrier, fin de période

5.4.4 Une inflexion discursive amorcée autour de 2016–2018

L'ensemble des modèles dynamiques mobilisés dans cette étude — DTM et BERTopic évolutif — converge vers un même constat : une inflexion progressive du discours médiatique apparaît à partir de la période 2016–2018. Cette évolution coïncide avec la succession de plusieurs événements climatiques extrêmes marquants, parmi lesquels les sécheresses répétées en Belgique (2017, 2018 et 2020), les inondations dans la province de Liège ou encore les canicules records observées en Europe à partir de 2019. Ces événements semblent avoir contribué à un élargissement progressif des thématiques abordées dans la couverture médiatique.

Dans le DTM, cette évolution se traduit par une rupture lexicale importante au sein du Topic 0 (T_r = 0,48, période 2016–2018), caractérisée par un déplacement progressif du vocabulaire utilisé : les références essentiellement institutionnelles et météorologiques laissent davantage place à des termes associés aux impacts environnementaux et à la gestion des risques, tels que *rapport, zone, incendie ou pluie*. Du côté de BERTopic évolutif, cette transformation apparaît notamment à travers la montée progressive du Topic 1 (*eau, sécheresse, agriculteurs*) à partir de 2016–2017.

Ce signal rejoint le cadre théorique établi en Section 2.6 : Vasterman (2005) montre que les **media hypes** climatiques sont généralement déclenchés par des événements extrêmes locaux d'une intensité suffisante pour rompre le seuil d'indifférence éditoriale. Les sécheresses belges de 2017–2018, en affectant directement l'agriculture wallonne et les ressources en eau potable, ont potentiellement joué ce rôle de déclencheur. Schmidt, Ivanova, et Schäfer (2013) confirment que de tels événements locaux produisent des effets durables sur la structure de la couverture médiatique, bien au-delà de l'épisode immédiat. Ce que la trajectoire progressive du Topic 1 BERTopic évolutif illustre parfaitement, puisque la montée en puissance de la thématique sécheresse se maintient et se consolide jusqu'en 2021 (Figure 4.17).

La Figure 5.3 combine la trajectoire du Topic 1 BERTopic (*sécheresse, eau, agriculteurs*) et le taux de renouvellement T_r du Topic 0 DTM sur l'axe temporel commun, superposés aux dates des événements climatiques extrêmes majeurs. La concomitance entre la rupture lexicale du Topic 0 DTM (T_r = 0,48, période 2017–2018) et l'émergence exponentielle du Topic 1 BERTopic confirme la réactivité du discours médiatique de la RTBF aux événements climatiques réels, et illustre la bifurcation discursive amorcée autour de 2016–2018.

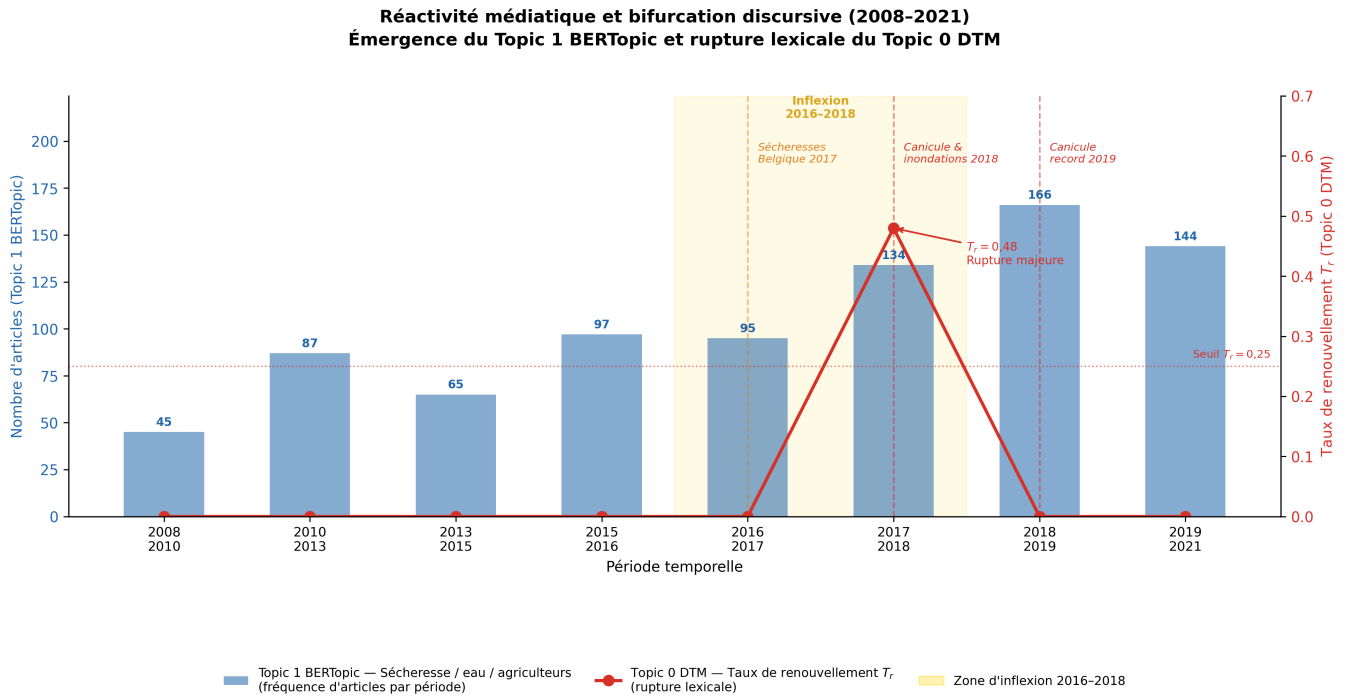


Figure 5.3: Réactivité médiatique et bifurcation discursive (2008–2021) : émergence du Topic 1 BERTopic (*sécheresse / eau / agriculteurs*) et rupture lexicale du Topic 0 DTM ($T_r = 0{,}48$). La zone dorée marque la période d'inflexion 2016–2018 ; les repères verticaux correspondent aux événements climatiques extrêmes majeurs survenus en Belgique.

5.4.5 La montée des aléas thermiques : émergence tardive et consolidation progressive

La progression des aléas thermiques — sécheresses et vagues de chaleur — constitue l'un des signaux les plus robustes observés dans l'ensemble des modèles mobilisés. Quasi absents avant 2013, ces enjeux gagnent progressivement en visibilité jusqu'en 2021, où ils occupent désormais une place significative dans la couverture médiatique. Cette évolution apparaît de manière cohérente dans l'analyse exploratoire (Section 4.3.1, Figure 4.5a), dans les labels dynamiques du DTM (Table 6.6) ainsi que dans la trajectoire temporelle du Topic 1 du BERTopic dynamique (Figure 4.17).

Ces résultats apportent un appui direct aux travaux de Rouquette et Bihay (2022) sur la faible visibilité médiatique des risques climatiques lents et structurels. Nos données montrent en effet que cette invisibilité caractérise largement la période 2008–2013, avant de se réduire progressivement à partir de 2016–2018 sous l'effet de plusieurs événements climatiques extrêmes affectant directement le territoire belge. La sécheresse agricole et les enjeux liés à la gestion de l'eau deviennent alors des objets médiatiques plus autonomes et plus durablement intégrés à l'espace journalistique, notamment dans le contexte wallon. L'un des apports de nos modèles réside précisément dans leur capacité à dater et à quantifier cette transition discursive de manière systématique.

Cette transformation graduelle semble par ailleurs cohérente avec les travaux de Boykoff (2010) sur les cycles d'attention médiatique liés au changement climatique. Notre corpus suggère l'existence d'une phase de transition amorcée autour de 2016–2018, marquée par une montée progressive des enjeux thermiques dans l'agenda médiatique. Toutefois, la phase de consolidation complète (caractérisée par une intégration durable des dimensions causales, structurelles et préventives dans les discours journalistiques) demeure encore incomplète sur la période étudiée. L'extension du corpus au-delà de 2021, notamment à la suite

des inondations catastrophiques de juillet 2021, permettrait d'évaluer si cette évolution s'est effectivement consolidée dans les années suivantes.

5.5 Limites et biais de l'étude

Plusieurs limites méritent d'être explicitées.

La généralité du corpus. Le corpus RTBFINFO couvre l'ensemble de l'actualité belge et internationale. Cette généralité, si elle offre une richesse documentaire indéniable, génère inévitablement une contamination thématique dans les modèles non supervisés — en particulier dans BERTopic, où de nombreux topics capturent des violences urbaines, des faits divers ou des crises politiques qui partagent un vocabulaire partiellement commun avec les catastrophes naturelles (urgence, victimes, autorités, intervention).

La labellisation automatique. Le recours à GPT-4o-mini pour la labellisation des topics est reproductible et scalable, mais il introduit un biais de formulation : le modèle de langage tend à produire des labels génériques et formellement similaires, ce qui peut masquer des nuances sémantiques fines entre topics proches. La validation humaine, bien qu'abandonnée pour les modèles dynamiques, resterait souhaitable pour le modèle LDA statique.

La couverture temporelle. Le corpus s'arrête en 2021. Les inondations catastrophiques de juillet 2021 en province de Liège — l'un des événements climatiques les plus meurtriers de l'histoire belge récente — sont donc partiellement couvertes. Il est probable que leur inclusion aurait amplifié les signaux d'évolution thématique déjà observés, en particulier autour de la gestion des risques hydrologiques locaux.

La dimension causale absente. Ce mémoire analyse principalement l'évolution du discours médiatique relatif aux catastrophes naturelles et aux risques climatiques, sans mesurer directement ses effets sur la perception publique, les comportements individuels ou les décisions politiques. Il serait dès lors réducteur de considérer que la montée en visibilité des aléas thermiques dans les médias entraîne mécaniquement une amélioration de la sensibilisation citoyenne au changement climatique. Les dynamiques observées peuvent en réalité résulter de plusieurs facteurs entremêlés : la survenue d'événements climatiques extrêmes locaux, les évolutions des lignes éditoriales et des priorités journalistiques, les transformations du contexte politique international, ou encore une sensibilisation croissante de la société aux enjeux environnementaux. Les signaux mis en évidence dans le corpus apparaissent ainsi comme des effets composites, dont les contributions respectives restent difficiles à isoler avec les seules méthodes de modélisation thématique mobilisées dans ce travail.

Cette limite rend également délicate la construction d'un véritable indicateur de sensibilisation médiatique aux risques climatiques. Les modèles utilisés permettent d'identifier des évolutions discursives et des transformations thématiques, mais ils ne suffisent pas à distinguer clairement ce qui relève d'une réaction ponctuelle à l'actualité, d'un changement structurel des représentations médiatiques ou d'une évolution plus profonde de la conscience collective face aux risques environnementaux.

5.6 Perspectives de recherche

Ce travail ouvre plusieurs pistes pour des recherches futures.

Extension temporelle et corpus post-2021. L'intégration des archives 2021–2024, en particulier des inondations de juillet 2021 et des vagues de chaleur de 2022, permettrait de tester la robustesse de l'inflexion observée et d'en mesurer l'ampleur réelle dans les années qui ont suivi.

Comparaison multi-source. Une extension à d'autres médias belges francophones (RTL, Le Soir, La Libre Belgique) ou à la presse flamande (VRT, De Standaard) permettrait de déterminer si les tendances observées sont propres à la RTBF ou caractéristiques du paysage médiatique belge dans son ensemble.

Analyse de sentiment et de cadrage (*framing*). Le *topic modelling* identifie *ce dont* les médias parlent, mais pas *comment* ils en parlent. L'intégration d'une analyse de sentiment ou d'une analyse du cadrage narratif (*frame analysis*) permettrait de compléter les résultats en mesurant l'évolution du registre émotionnel — alarmiste, neutre, solutionniste — associé aux différentes catégories d'aléas.

Amélioration du modèle BERTopic. L'instabilité structurelle de BERTopic ($\bar{J} = 0{,}236$) suggère qu'un travail plus fin sur les hyperparamètres UMAP ou l'utilisation d'embeddings spécialisés pour le domaine climatique (par exemple, des modèles pré-entraînés sur des corpus environnementaux ou journalistiques) pourrait significativement améliorer la cohérence et la reproductibilité du modèle.

Vers une amélioration méthodologique. Comme indiqué dans le sujet de mémoire, une piste de recherche naturelle consiste à exploiter les limites identifiées des méthodes existantes pour proposer une amélioration d'un modèle existant capable de mieux traiter des corpus journalistiques généralistes multi-thématiques. Le principal défi identifié — la contamination thématique de BERTopic — pourrait être abordé par un filtrage semi-supervisé intégrant le dictionnaire d'aléas directement dans la phase de clustering, ou par une contrainte sur les représentations c-TF-IDF forçant les topics à s'ancrer dans le vocabulaire climatique.

Compléter l'analyse thématique par des approches d'extraction d'information appliquées aux contenus journalistiques. Alors que le topic modeling permet principalement d'identifier les grands thèmes et leurs évolutions discursives, des méthodes d'extraction d'entités, d'événements ou de relations permettraient d'étudier plus directement les dimensions opérationnelles de la gestion des risques. Une telle approche pourrait notamment servir à détecter automatiquement la présence d'alertes précoces, de mesures d'évacuation, d'initiatives de prévention ou encore de dispositifs institutionnels mis en place avant ou après les catastrophes. Elle permettrait également d'évaluer si la fréquence des discours liés à l'anticipation, à la préparation et à l'adaptation tend à augmenter au fil du temps dans la couverture médiatique des risques naturels et climatiques. Cette complémentarité entre modélisation thématique et extraction d'information ouvrirait ainsi la voie à une analyse plus fine du rôle des médias dans la diffusion des stratégies de prévention et de résilience face aux catastrophes environnementales.

5.7 Conclusion générale

Pour un corpus journalistique francophone traitant de thématiques environnementales complexes, les modèles probabilistes interprétables conservent un avantage décisif sur les approches neuronales récentes, comme le soulignent Tanguy et al. (2025) : la transparence algorithmique et la possibilité d'un dialogue avec des experts du domaine priment sur la performance brute de cohérence sémantique. Le LDA et le DTM, grâce au tri lexical préalable qui délimite leur périmètre thématique, produisent des topics plus directement interprétables et moins contaminés par le bruit du corpus généraliste. BERTopic, de son côté, confirme sa capacité à identifier des structures sémantiques plus fines, mais au prix d'une sensibilité accrue aux conditions d'initialisation et d'une contamination thématique plus difficile à maîtriser.

Sur le fond, ce mémoire confirme et précise la conjecture initiale : un glissement thématique s'est bien produit dans le discours médiatique de la RTBF sur les catastrophes naturelles entre 2008 et 2021. Ce glissement n'est pas une rupture franche — l'inertie lexicale domine — mais une dérive progressive et cohérente, amorcée autour de 2016–2018, qui traduit une adaptation de l'agenda médiatique à la réalité climatique belge et européenne. La sécheresse et les vagues de chaleur, longtemps absentes du corpus, y occupent désormais une place significative ; et les aléas, longtemps traités comme des événements lointains

et spectaculaires, font progressivement l'objet d'un traitement éditorial plus ancré dans les réalités locales et dans les enjeux de long terme.

Cette évolution demeure néanmoins lente et partielle. La couverture des causes structurelles du changement climatique — politiques d'atténuation, trajectoires d'émissions, mécanismes physiques des aléas — reste très minoritaire face à la prédominance du registre événementiel et réactif. À l'heure où les États membres de l'Union européenne renforcent leurs politiques d'adaptation climatique, la question de savoir si et comment les médias peuvent contribuer à une meilleure compréhension des risques naturels et non seulement à leur narration reste entière.

5.8 Reproductibilité et accès au code source

Dans une démarche de reproductibilité scientifique, l'ensemble des scripts, notebooks et fichiers utilisés dans le cadre de ce mémoire a été regroupé dans un dépôt GitHub dédié :

https://github.com/Belanov17/memoire-topic-modeling

L'analyse exploratoire des données ainsi que les modèles probabilistes de modélisation de thématiques (LDA et DTM) sont principalement implémentés dans le notebook lda_model.ipynb. Les approches neuronales, notamment BERTopic et BERTopic dynamique, sont regroupées dans le notebook bert_model.ipynb.

Le document principal du mémoire est contenu dans le fichier redaction_memoire.qmd, accompagné des fichiers bibliographiques (reference.bib), figures et annexes nécessaires à la compilation du rapport final.

Bibliographie

Aho, Alfred V., et Margaret J. Corasick. 1975. « Efficient String Matching: An Aid to Bibliographic Search ». *Communications of the ACM* 18 (6): 333-40.

Akiba, Takuya, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, et Masanori Koyama. 2019. « Optuna: A Next-generation Hyperparameter Optimization Framework ». In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2623-31. https://doi.org/10.1145/3292500.3330701.

Angelov, Dimo. 2020. « Top2Vec: Distributed Representations of Topics ». *arXiv preprint* arXiv:2008.09470. https://arxiv.org/abs/2008.09470.

Belazzougui, Djamal. 2010. « Succinct Dictionary Matching With No Slowdown ». *Algorithms* 3 (4): 414-28.

Ben-Hur, Asa, Andre Elisseeff, et Isabelle Biermann. 2002. « A stability based method for discovering structure in clustered data ». *Pacific Symposium on Biocomputing* 7: 6-17.

Berelson, Bernard. 1952. *Content Analysis in Communication Research*. Glencoe, Illinois: Free Press.

Bianchi, Federico, Silvia Terragni, et Dirk Hovy. 2021. « Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence ». In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 759-66. https://doi.org/10.18653/v1/2021.acl-short.96.

Blei, David M., et John D. Lafferty. 2006. « Dynamic Topic Models ». *Proceedings of the International Conference on Machine Learning*.

———. 2007. « A Correlated Topic Model of Science ». *The Annals of Applied Statistics* 1 (1): 17-35.

Blei, David M., Andrew Y. Ng, et Michael I. Jordan. 2003. « Latent Dirichlet Allocation ». *Journal of Machine Learning Research* 3: 993-1022.

Boykoff, Maxwell T. 2010. « Indian media representations of climate change in a threatened journalistic ecosystem ». *Climatic Change* 99 (1): 17-25. https://doi.org/10.1007/s10584-010-9807-8.

Boykoff, Maxwell T., et Jules M. Boykoff. 2007. « Climate change and journalistic norms: A case-study of US mass-media coverage ». *Geoforum* 38 (6): 1190-1204. https://doi.org/10.1016/j.geoforum.2007.01.008.

Bozic, Bojan, et Matjaz Murko. 2021. « Multilingual COVID-19 Fake News Detection using BERTopic ». *International Journal of Advances in Intelligent Systems* 14 (3 & 4): 224-35. https://iariajournals.org.

Campello, Ricardo J. G. B., Davoud Moulavi, et Jörg Sander. 2013. « Density-Based Clustering Based on Hierarchical Density Estimates ». In *Advances in Knowledge Discovery and Data Mining (PAKDD)*, 7819:160-72. Lecture Notes in Computer Science. Springer. https://doi.org/10.1007/978-3-642-37456-2_14.

Chib, Siddhartha. 1995. « Marginal Likelihood from the Gibbs Output ». *Journal of the American Statistical Association* 90 (432): 1313-21. https://doi.org/10.1080/01621459.1995.10476635.

Cohen, Jacob. 1960. « A coefficient of agreement for nominal scales ». *Educational and Psychological Measurement* 20 (1): 37-46.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, et Richard Harshman. 1990. « Indexing by Latent Semantic Analysis ». *Journal of the American Society for Information Science* 41 (6): 391-407.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, et Kristina Toutanova. 2019. « BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding ». In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171-86. Association for Computational Linguistics. https://aclanthology.org/N19-1423.

Eckart, Carl, et Gale Young. 1936. « The approximation of one matrix by another of lower rank ». *Psychometrika* 1 (3): 211-18. https://doi.org/10.1007/BF02288367.

Efron, Bradley. 1979. « Bootstrap Methods: Another Look at the Jackknife ». *The Annals of Statistics* 7 (1): 1-26. https://doi.org/10.1214/aos/1176344552.

Egger, Roman, et Joanne Yu. 2022. « Identifying Topics in Scientific Articles: An Optimization of BERTopic's Hyperparameters ». *Frontiers in Sociology* 7: 886498. https://doi.org/10.3389/fsoc.2022.886498.

Escouflaire, Louis, Romain Fonteneau, Benoît Descy, et Cédrick Fairon. 2023. « The RTBF Corpus: a Dataset of 750,000 Belgian French News Articles Published Between 2008 and 2021 ». In *Proceedings of the International Conference on Corpus Linguistics (JLC23)*. Dataverse. https://doi.org/10.14428/DVN/PEVSSI.

Fleiss, Joseph L. 1971. « Measuring nominal scale agreement among many raters ». *Psychological Bulletin* 76 (5): 378-82.

Greene, Derek, Derek O'Callaghan, et Pádraig Cunningham. 2014. « How Many Topics? Stability Analysis for Non-Negative Matrix Factorization ». In *Proceedings of the 15th International Conference on Computing (ECML PKDD)*. Springer.

Griffiths, Thomas L., et Mark Steyvers. 2004. « Finding scientific topics ». *Proceedings of the National Academy of Sciences* 101 (suppl_1): 5228-35. https://doi.org/10.1073/pnas.0307752101.

Groot, Muriël de, Mohammad Aliannejadi, et Marcel R. Haas. 2022. « Experiments on Generalizability of BERTopic on Multi-Domain Short Text ». *arXiv preprint arXiv:2212.08459*. https://arxiv.org/abs/2212.08459.

Grootendorst, Maarten. 2022. « BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure ». *arXiv preprint* arXiv:2203.05794. https://arxiv.org/abs/2203.05794.

Groupe d'experts intergouvernemental sur l'évolution du climat. 2014. « Glossaire : Changements climatiques 2014, Conséquences, adaptation et vulnérabilité ». In *Contribution du Groupe de travail II au cinquième Rapport d'évaluation du GIEC*, édité par J. Agard et E. L. F. Schipper. Organisation météorologique mondiale et Programme des Nations Unies pour l'environnement. https://www.ipcc.ch/site/assets/uploads/2018/02/AR5_WGII_glossary_FR.pdf.

Hofmann, Thomas. 1999b. « Probabilistic latent semantic indexing ». In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50-57. ACM. https://doi.org/10.1145/312624.312649.

———. 1999a. « Probabilistic Latent Semantic Indexing ». In *Proceedings of SIGIR*, 50-57.

Huang, Qunying, et Yu Xiao. 2015. « Geographic Situational Awareness: Mining Tweets for Disaster Preparedness, Emergency Response, Impact, and Recovery ». *ISPRS International Journal of Geo-Information* 4 (3): 1549-68. https://doi.org/10.3390/ijgi4031549.

Hubert, Lawrence, et Phipps Arabie. 1985. « Comparing Partitions ». *Journal of Classification* 2 (1): 193-218. https://doi.org/10.1007/BF01908075.

IPCC. 2021. « Climate Change 2021: The Physical Science Basis ». Intergovernmental Panel on Climate Change.

Islam, Md. Rezwanul et al. 2023. « Adaptation of BERTopic for Short Text Corpora and Multi-domain Contexts ». *Journal of Data Science and Intelligent Systems*.

Jaccard, Paul. 1901. « Étude comparative de la distribution florale dans une portion des Alpes et des Jura ». *Bulletin de la Société Vaudoise des Sciences Naturelles* 37: 547-79. https://doi.org/10.5169/seals-266440.

Jacobi, Carina, Wouter van Atteveldt, et Kasper Welbers. 2016. « Quantitative analysis of large amounts of journalistic texts using topic modelling ». *Digital Journalism* 4 (1): 89-106. https://doi.org/10.1080/21670811.2015.1093271.

Joye, Stijn. 2010. « The hierarchy of global suffering: A critical discourse analysis of television news reporting on foreign natural disasters ». *Journal of International Communication* 16 (2): 45-61. https://doi.org/10.1080/13216597.2010.9751882.

Krippendorff, Klaus. 2004. *Content Analysis: An Introduction to Its Methodology*. 2 éd. Thousand Oaks,

California: Sage Publications.
Kullback, Solomon, et Richard A. Leibler. 1951. « On Information and Sufficiency ». *The Annals of Mathematical Statistics* 22 (1): 79-86. https://doi.org/10.1214/aoms/1177729694.
Lalande, Justine, et Stéphanie Yates. 2025. « Communication climatique : Synthèse de litérature ». In *Actes de CORIA-TALN-RECITAL*.
Landis, J. Richard, et Gary G. Koch. 1977. « The measurement of observer agreement for categorical data ». *Biometrics* 33 (1): 159-74.
Lee, Daniel D., et H. Sebastian Seung. 1999. « Learning the Parts of Objects by Non-Negative Matrix Factorization ». *Nature* 401: 788-91.
———. 2001. « Algorithms for Non-negative Matrix Factorization ». In *Advances in Neural Information Processing Systems 13 (NIPS 2000)*, 556-62. MIT Press.
Lin, Jianhua. 1991. « Divergence Measures Based on the Shannon Entropy ». *IEEE Transactions on Information Theory* 37 (1): 145-51.
McInnes, Leland, John Healy, et Steve Astels. 2017. « hdbscan: Hierarchical density based clustering ». *Journal of Open Source Software* 2 (11).
McInnes, Leland, John Healy, Nathaniel Saul, et Lukas Großberger. 2018. « UMAP: Uniform Manifold Approximation and Projection ». *Journal of Open Source Software* 3 (29): 861. https://doi.org/10.21105/joss.00861.
Meier, Hanne, et Mikkel Eskjaer. 2024. « The Evolution of Climate Change Coverage: A Structural Topic Model Approach ». *Journal of Environmental Communication*.
Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, et Jeffrey Dean. 2013. « Distributed Representations of Words and Phrases and their Compositionality ». In *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc. https://papers.nips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.
Mimno, David, Hanna Wallach, Edmund Talley, Miriam Leenders, et Andrew McCallum. 2011. « Optimizing Semantic Coherence in Topic Models ». In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 262-72. https://aclanthology.org.
Nations Unies. 2016. « Rapport du groupe de travail intergouvernemental d'experts à composition non limitée chargé des indicateurs et de la terminologie relatifs à la réduction des risques de catastrophe (A/71/644) ». Assemblée générale des Nations Unies. https://www.undrr.org/publication/report-open-ended-intergovernmental-expert-working-group-indicators-and-terminology.
Office québécois de la langue française. s.d. « Grand dictionnaire terminologique ». Gouvernement du Québec. s.d. https://www.oqlf.gouv.qc.ca/ressources/bibliotheque/dictionnaires/VocabulairesPDF/vocabulaire-changements-climatiques.pdf.
Pennington, Jeffrey, Richard Socher, et Christopher D. Manning. 2014. « GloVe: Global Vectors for Word Representation ». In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-43. Association for Computational Linguistics. https://aclanthology.org/D14-1162.
Rand, William M. 1971. « Objective Criteria for the Evaluation of Clustering Methods ». *Journal of the American Statistical Association* 66 (336): 846-50. https://doi.org/10.1080/01621459.1971.10482356.
Reimers, Nils, et Iryna Gurevych. 2019. « Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks ». In *Proceedings of EMNLP 2019*.
Roder, Michael, Andreas Both, et Alexander Hinneburg. 2015. « Exploring the Space of Topic Coherence Measures ». In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399-408. https://doi.org/10.1145/2684822.2685324.
Rosen-Zvi, Michal, Thomas Griffiths, Mark Steyvers, et Padhraic Smyth. 2004. « The Author-Topic Model for Authors and Documents ». In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 487-94. AUAI Press.
Rouquette, Sébastien, et Thomas Bihay. 2022. « Les risques naturels médiatiquement invisibles ». *études de communication* 48: 117-36. https://doi.org/10.4000/edc.6914.

Schmidt, Andreas, Ana Ivanova, et Mike S. Schäfer. 2013. « Media attention for climate change around the world: A comparative analysis of newspaper coverage in 27 countries ». *Global Environmental Change* 23 (5): 1233-48. https://doi.org/10.1016/j.gloenvcha.2013.07.020.

Sebastiani, Fabrizio. 2002. « Machine Learning in Automated Text Categorization ». *ACM Computing Surveys* 34 (1): 1-47.

Shannon, Claude E. 1948. « A Mathematical Theory of Communication ». *Bell System Technical Journal* 27 (3): 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Sievert, Carson, et Kenneth Shirley. 2014. « LDAvis: A method for visualizing and interpreting topics ». In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63-70. Association for Computational Linguistics. https://doi.org/10.3115/v1/w14-3110.

Sjöbergh, Jonas, et Niko Nanas. 2012. « The Effect of Time Granularity on Topic Models ». *Proceedings of the Workshop on Computational Linguistics for Literature*.

Stevens, Keith, Philip Anderson, et Mark Liberman. 2012. « Exploring topic coherence over many models and datasets ». In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 952-61. Association for Computational Linguistics.

Student. 1908. « The Probable Error of a Mean ». *Biometrika* 6 (1): 1-25. https://doi.org/10.1093/biomet/6.1.1.

Tanguy, Ludovic, Cécile Fabre, Nabil Hathout, et Lydia-Mai Ho-Dac. 2025. « Embeddings, Topic Models, LLM : un air de famille ». In *Actes de CORIA-TALN-RECITAL*.

Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, et David M. Blei. 2006. « Hierarchical Dirichlet Processes ». In *Journal of the American Statistical Association*, 101:1566-81. 476.

Tumeo, Antonino, Oreste Villa, et Daniel Chavarria-Miranda. 2012. « Aho–Corasick String Matching on Shared and Distributed Memory Parallel Architectures ». In *Proceedings of the International Parallel and Distributed Processing Symposium*.

United Nations Office for Disaster Risk Reduction. s.d. « Hazard Information Profiles (HIPs) ». PreventionWeb.net. s.d. https://www.preventionweb.net/drr-glossary/hips.

Vasterman, Peter L. M. 2005. « Media-Hype: Self-Reinforcing News Waves, Journalistic Standards and the Construction of Social Problems ». *European Journal of Communication* 20 (4): 508-30. https://doi.org/10.1177/0267323105058254.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, et Illia Polosukhin. 2017. « Attention Is All You Need ». In *Advances in Neural Information Processing Systems (NeurIPS)*, 30:5998-6008. https://neurips.cc.

6 Annexes

Table 6.1: Regroupement thématique des topics BERTopic

Catégorie thématique	T.	Label du topic	Mots-clés représentatifs
Climat et phénomènes météorologiques	0	Chaleur, sécheresse et instabilité orageuse	degrés, températures, pays, averses, ouest, vent, chaleur, temps, température, soleil
	3	Ouragans tropicaux et risques côtiers	ouragan, nhc, Haïti, ouragans, tempête tropicale, vents, tempête, sud, Mexique, Floride
	9	Impacts et réponses au typhon Haiyan	Philippines, typhon, Manille, personnes, archipel, morts, Haiyan, Hong, autorités, typhons
	12	Cyclones et vulnérabilité des populations côtières	cyclone, Bangladesh, archipel, île, personnes, vents, habitants, autorités, Vanuatu, côtes
	17	Impacts socio-environnementaux des typhons à Tokyo	Tokyo, typhon, préfecture, NHK, personnes, région, agence météorologique, japonaise, glissements de terrain, archipel
	18	Impacts hydrométéorologiques et socio-économiques des typhons	province, personnes, typhon, inondations, autorités, provinces, morts, pluies torrentielles, pays, maisons
	30	Tempêtes hivernales : vents, submersions et perturbations	tempête, neige, matin, départements, nord, Angleterre, pays, nuit, météo, vent
	33	Vents violents et cyclones tropicaux	vents, vents violents, tempête, belga, tornade, région, Mexique, nhc, sud, rafales
Inondations et gestion hydrique	4	Interventions d'urgence des pompiers en inondations	pompiers, caves, eau, belga, inondations, rue, commune, soir, matin, maisons
	5	Gestion communale de l'eau et inondations	communes, eau, commune, Suède, projet, travaux, situation, inondations, région wallonne, province
	14	Inondations, glissements et mortalité	personnes, victimes, morts, belga, ville, victime, inondations, maisons, capitale, état
	19	Inondations, mortalité et aide humanitaire	personnes, ville, inondations, morts, habitants, eau, autorités, région, pays, capitale
	20	Crise humanitaire des inondations au Pakistan	Pakistan, pakistanais, inondations, Afghanistan, Sind, pays, province, millions de personnes, talibans, millions

Suite page suivante…

Catégorie thématique	T.	Label du topic	Mots-clés représentatifs
	22	Interventions des pompiers lors d'intempéries	pompiers, interventions, caves, matin, eau, dégâts, Bruxelles, région, nuit, routes
	23	Inondations, impacts humains et réponses publiques	inondations, personnes, pays, ville, morts, belga, habitants, état, autorités, région
	27	Risques hydrogéologiques en Indonésie	personnes, glissements de terrain, inondations, habitants, région, corps, maisons, Indonésie, terrain, secours
	31	Gestion des niveaux d'eau et barrages	barrage, eau, niveau, barrages, kayaks, cours d'eau, Ourthe, bas, débit, lac
Incendies et catastrophes naturelles	6	Incendies forestiers en Californie	californie, incendies, incendie, pompiers, feu, flammes, personnes, angeles, nord, feux
	11	Incendies forestiers en Nouvelle-Galles-du-Sud	incendies, incendie, hectares, pompiers, feux, feu, flammes, galles-sud, région, sud
	24	Catastrophes naturelles et gestion russe	neige, russes, russe, région, nord, centre, capitale, incendies, poutine, ministère-russe
	32	Incendies forestiers et protection civile portugaise	portugal, pompiers, incendie, incendies, protection civile, flammes, feu, portugais, feux, centre
Santé et crises humanitaires	10	Famine, sécheresse et déplacements en Somalie	somalie, famine, pays, mozambique, réfugiés, millions personnes, enfants, personnes, sécheresse, kenya
	16	COVID-19 : épidémiologie, mortalité et interventions sanitaires	coronavirus, patients, virus, cas, hôpital, épidémie, médecins, personnes, traitement
Politique et gouvernance	2	Présidence, partis et négociations gouvernementales	président, pays, politique, accord, gouvernement, parti, ans, tous, ministre, opposition
Transport et infrastructures	7	Accidentologie et sécurité aérienne	avion, aéroport, vol, passagers, vols, compagnie, appareil, pilote, navire, accident
	13	Fiabilité et perturbations du service ferroviaire	trains, ligne, train, sncb, bruxelles, lignes, bus, voyageurs, trafic, voies
Société, économie et territoires	1	Sécheresse, eau, rendements et prix	eau, sécheresse, agriculteurs, année, production, prix, espèces, récolte, arbres, plantes
	8	Vécu collectif des habitants urbains	personnes, enfants, ans, ville, eau, habitants, temps, place, gens, tous
	15	Intempéries extrêmes et sécurité des festivals	festival, films, festivaliers, albums, cinéma, concert, scène, musique, épisodes
	21	Accidents, victimes et procédures judiciaires	police, accident, bâtiment, victimes, policiers, justice, personnes, pompiers, matin, tribunal
	25	Fréquentation touristique des musées et attractions	musée, tourisme, visiteurs, année, ans, Venise, statue, parc, fréquentation

Suite page suivante…

Catégorie thématique	T.	Label du topic	Mots-clés représentatifs
	26	Coûts assurantiels des catastrophes naturelles	milliards de dollars, milliard, milliard d'euros, catastrophes, assureurs, dommages, assurances, dégâts
	28	Revitalisation commerciale des centres-villes	ville, rue, travaux, centre, commerces, projets, commune, euro, place
	29	Fermetures et sécurisation des parcs	Bruxelles Environnement, parcs, Bruxelles, rafales de vent, inspection, nettoyage, parcs bois, bruxellois, public, bois cambre

Table 6.2: Évolution thématique des topics BERTopic dynamique par période

Topic	Période	Label dynamique
0	2008_2010	Violences urbaines et crise sociale
0	2010_2013	Violences urbaines et crise sociétale
0	2013_2015	Violences urbaines et crises sociales
0	2015_2016	Violences urbaines et inondations en Belgique
0	2016_2017	Violences urbaines et enjeux sociaux
0	2017_2018	Violences urbaines et crises climatiques
0	2018_2019	Violences urbaines et crise sociale
0	2019_2021	Violences urbaines et crise sociale
1	2008_2010	Violences urbaines et crise sociale
1	2010_2013	Violences urbaines et crise sociale
1	2013_2015	Émeutes et enjeux sociaux contemporains
1	2015_2016	Violences urbaines et crise sanitaire
1	2016_2017	Crise de l'eau et tensions sociales
1	2017_2018	Gestion des crises et conséquences environnementales
1	2018_2019	Crise climatique et tensions sociales
1	2019_2021	Violences urbaines et enjeux sociaux
2	2008_2010	Violences urbaines et enjeux sociopolitiques
2	2010_2013	Violences urbaines et crise politique
2	2013_2015	Violences urbaines et réponses politiques
2	2015_2016	Violences urbaines et politiques de gestion
2	2016_2017	Violences urbaines et enjeux sociaux
2	2017_2018	Violences urbaines et crise sociale
2	2018_2019	Violences urbaines et crises sociopolitiques
2	2019_2021	Violences urbaines et perceptions politiques
3	2008_2010	Violences urbaines et catastrophes naturelles
3	2010_2013	Violences urbaines et catastrophes naturelles
3	2013_2015	Violences urbaines et catastrophes naturelles
3	2015_2016	Violences urbaines et enjeux sociaux
3	2016_2017	Incidences des catastrophes climatiques sur la société
3	2017_2018	Violences urbaines et crises sociétales
3	2018_2019	Violences urbaines et gestion des crises
3	2019_2021	Événements climatiques et tensions sociales

Suite page suivante...

Topic	Période	Label dynamique
4	2008_2010	Violences urbaines et gestion des crises
4	2010_2013	Inondations et réponses des autorités locales
4	2013_2015	Inondations et enjeux socio-politiques en Belgique
4	2015_2016	Gestion des crises et violences urbaines
4	2016_2017	Interventions d'urgence en contextes émergents
4	2017_2018	Inondations et interventions des pompiers
4	2018_2019	Crise des services d'urgence et inondations
4	2019_2021	Interventions d'urgence et gestion des crises
5	2008_2010	Gestion des crises hydriques et sociales
5	2010_2013	Inondations et dérives sociales en Belgique
5	2013_2015	Gestion des inondations et crises urbaines
5	2015_2016	Gestion des inondations et crise sociale
5	2016_2017	Violences urbaines et gestion des crises
5	2017_2018	Violences urbaines et gestion de crises
5	2018_2019	Violences urbaines et crises sociales
5	2019_2021	Crise de l'eau et gestion communale
6	2008_2010	Incendies et gestion des crises urbaines
6	2010_2013	Violences urbaines et crises sociales
6	2013_2015	Incendies et interventions des pompiers
6	2015_2016	Incendies en Californie et réponse policière
6	2016_2017	Incendies et violences urbaines en Belgique
6	2017_2018	Incendies en Californie et tensions sociales
6	2018_2019	Incendies et crises sociales contemporaines
6	2019_2021	Incendies de forêt en Californie
7	2008_2010	Violences urbaines et crise sanitaire
7	2010_2013	Violences urbaines et crise sociale
7	2013_2015	Violences urbaines et tensions sociales
7	2015_2016	Violences urbaines et crises sociales
7	2016_2017	Violences urbaines et inondations en Belgique
7	2017_2018	Violences urbaines et répercussions sociales
7	2018_2019	Conflits sociaux et gestion des crises
7	2019_2021	Violences urbaines et tensions sociales
8	2008_2010	Violences urbaines et enjeux sociétaux
8	2010_2013	Crise sociale et tensions urbaines
8	2013_2015	Violences urbaines et crises sociales
8	2015_2016	Violences urbaines et enjeux sociaux
8	2016_2017	Violences urbaines et crise sociale
8	2017_2018	Violences urbaines et crise sociale
8	2018_2019	Violences urbaines et enjeux sociétaux
8	2019_2021	Violences urbaines et impact social
9	2008_2010	Violences urbaines et réponses institutionnelles
9	2010_2013	Violences urbaines et crises sociales
9	2013_2015	Entre catastrophes naturelles et violences urbaines
9	2015_2016	Violences urbaines et crises sociales
9	2016_2017	Violences urbaines et crise sociale
9	2017_2018	Violences urbaines et crise climatique

Suite page suivante...

Topic	Période	Label dynamique
9	2018_2019	Violences policières et crises sociales
9	2019_2021	Violences urbaines et impacts sociaux
10	2008_2010	Violences urbaines et crise sociale
10	2010_2013	Violences urbaines et crise sociale
10	2013_2015	Conflits sociaux et crises sanitaires
10	2015_2016	Violences urbaines et crise sanitaire
10	2016_2017	Violences urbaines et crise sociale
10	2017_2018	Violences urbaines et inégalités sociales
10	2018_2019	Violences sociales et crise sanitaire
10	2019_2021	Violences urbaines et crises sociales
11	2008_2010	Incidents urbains et réponses policières
11	2010_2013	Incendies et interventions des pompiers
11	2013_2015	Incendies et interventions des pompiers
11	2015_2016	Incendies et gestion des crises urbaines
11	2016_2017	Violences urbaines et interventions policières
11	2017_2018	Violences urbaines et crises sociales
11	2018_2019	Incendies et crises sociales en Belgique
11	2019_2021	Gestion des crises et violences urbaines
12	2008_2010	Gestion des crises climatiques et sociales
12	2010_2013	Violences urbaines et crises sociales
12	2013_2015	Violences urbaines et crises sociales
12	2015_2016	Crise climatique et réactions sociétales
12	2016_2017	Violences urbaines et crises sanitaires
12	2017_2018	Violences urbaines et crises sociales
12	2018_2019	Violence urbaine et crise sociale
12	2019_2021	Violences urbaines et crise climatique
13	2008_2010	Violences urbaines et impact social
13	2010_2013	Conflits urbains et crises sociales
13	2013_2015	Violences urbaines et gestion des crises
13	2015_2016	Violences urbaines et crise sanitaire
13	2016_2017	Violences urbaines et crise sociale
13	2017_2018	Violences urbaines et tensions sociales à Bruxelles
13	2018_2019	Violences urbaines et tensions sociales
13	2019_2021	Mobilité, sécurité, et tensions urbaines
14	2008_2010	Violences urbaines et réactions institutionnelles
14	2010_2013	Violences urbaines et crise sociétale
14	2013_2015	Violences urbaines et tensions sociales
14	2015_2016	Violences urbaines et crises sociales
14	2016_2017	Violences urbaines et tensions sociales
14	2017_2018	Violences urbaines et réponses policières
14	2018_2019	Violences urbaines et crise sanitaire
14	2019_2021	Violences urbaines et réponses institutionnelles
15	2008_2010	Violences urbaines et crise sanitaire
15	2010_2013	Violences urbaines et enjeux sociaux
15	2013_2015	Violences urbaines et enjeux sociopolitiques
15	2015_2016	Violences urbaines et crises sociales

Suite page suivante…

Topic	Période	Label dynamique
15	2016_2017	Violences urbaines et interventions policières
15	2017_2018	Violences urbaines et crises sociales
15	2018_2019	Violences urbaines et enjeux sociaux
15	2019_2021	Violences urbaines et crise sociale
16	2008_2010	Violences urbaines et interventions policières
16	2010_2013	Violences urbaines et contextes socio-économiques
16	2013_2015	Violences urbaines et crise sociale
16	2015_2016	Violences urbaines et crise sanitaire
16	2016_2017	Violences urbaines et crise sanitaire
16	2017_2018	Violences urbaines et enjeux sociaux
16	2018_2019	Violences urbaines et perceptions sociales
16	2019_2021	Crise sanitaire et tensions sociales
17	2008_2010	Violences urbaines et gestion de crise
17	2010_2013	Violences urbaines et impacts sociétaux
17	2013_2015	Violences urbaines et crises sociales
17	2015_2016	Violences urbaines et gestion des crises
17	2016_2017	Gestion des crises et violences urbaines
17	2017_2018	Violences urbaines et crises sociales
17	2018_2019	Violences urbaines et crise sanitaire
17	2019_2021	Violences urbaines et réponses policières
18	2008_2010	Gestion des catastrophes naturelles en Belgique
18	2010_2013	Crise climatique et gestion des inondations
18	2013_2015	Violences urbaines et crise sociale
18	2015_2016	Violences urbaines et crises climatologiques
18	2016_2017	Violences urbaines et gestion des inondations
18	2017_2018	Crise des inondations en province
18	2018_2019	Violences sociales et gestion des crises
18	2019_2021	Violences urbaines et crise sociale
19	2008_2010	Violences urbaines et tensions sociopolitiques
19	2010_2013	Violences urbaines et crise sociale
19	2013_2015	Violence et inondations en Belgique 2013–2015
19	2015_2016	Gestion des inondations urbaines et crises sociales
19	2016_2017	Inondations et gestion de crise
19	2017_2018	Violences urbaines et enjeux sociaux
19	2018_2019	Violences urbaines et gestion de crise
19	2019_2021	Violences urbaines et crise sanitaire
20	2008_2010	Inondations et crises sociales au Pakistan
20	2010_2013	Crise sociale et féminisation des métiers
20	2013_2015	Conflits et crises sociales au Pakistan
20	2015_2016	Violences urbaines et enjeux sociaux
20	2016_2017	Violences urbaines et crise sociale
20	2017_2018	Conflits sociaux et enjeux politiques contemporains
20	2018_2019	Conflits sociaux et violences urbaines
20	2019_2021	Conflits régionaux et crises sociopolitiques
21	2008_2010	Violences urbaines et interventions policières
21	2010_2013	Violences urbaines et interventions policières

Suite page suivante…

Topic	Période	Label dynamique
21	2013_2015	Violences policières et tensions sociales
21	2015_2016	Violences urbaines et interventions policières
21	2016_2017	Violences urbaines et interventions policières
21	2017_2018	Violences urbaines et tensions sociales
21	2018_2019	Violences urbaines et interventions policières
21	2019_2021	Violences policières et tensions urbaines
22	2008_2010	Violences urbaines et crises sociales
22	2010_2013	Gestion des interventions d'urgence en crise
22	2013_2015	Violences urbaines et interventions d'urgence
22	2015_2016	Violences urbaines et interventions policières
22	2016_2017	Inondations et réponse des services d'urgence
22	2017_2018	Violences urbaines et interventions policières
22	2018_2019	Interventions d'urgence et violences policières
22	2019_2021	Interventions d'urgence des pompiers et inondations
23	2008_2010	Violences urbaines et inondations en Belgique
23	2010_2013	Violences urbaines et inondations en Belgique
23	2013_2015	Violences urbaines et crise sociale
23	2015_2016	Inondations et interventions d'urgence
23	2016_2017	Violences urbaines et gestion des crises
23	2017_2018	Inondations et crises sociales en 2017–2018
23	2018_2019	Gestion des inondations et conflits socio-urbains
23	2019_2021	Violences urbaines et crise sociétale
24	2008_2010	Violences policières et crises sociales
24	2010_2013	Violences urbaines et enjeux sociopolitiques
24	2013_2015	Violences urbaines et crise sociale
24	2015_2016	Conflits sociaux et crises environnementales
24	2016_2017	Violences urbaines et enjeux sociaux
24	2017_2018	Violences urbaines et crise sociale
24	2018_2019	Violences urbaines et crise sanitaire
24	2019_2021	Violences, tensions sociales et crises politiques
25	2008_2010	Violences urbaines et crise sociale
25	2010_2013	Violences urbaines et enjeux sociaux
25	2013_2015	Violences urbaines et réponses policières
25	2015_2016	Violences urbaines et mobilisation sociale
25	2016_2017	Conflits sociaux et enjeux de santé
25	2017_2018	Violences urbaines et enjeux sociaux
25	2018_2019	Violences urbaines et contextes sociaux
25	2019_2021	Violences policières et tensions sociales
26	2008_2010	Gestion des crises et catastrophes naturelles
26	2010_2013	Violences urbaines et enjeux sociaux
26	2013_2015	Violences urbaines et tensions sociales
26	2015_2016	Violences urbaines et catastrophes naturelles
26	2016_2017	Crise sociale et gestion des catastrophes
26	2017_2018	Catastrophes naturelles et réponses sociopolitiques
26	2018_2019	Violences urbaines et crise sociale
26	2019_2021	Violences urbaines et crise sanitaire

Suite page suivante…

Topic	Période	Label dynamique
27	2008_2010	Violences urbaines et crises sociales
27	2010_2013	Violences urbaines et contexte socio-économique
27	2013_2015	Violences urbaines et gestion des crises
27	2015_2016	Violences urbaines et crises sociales
27	2016_2017	Violences urbaines et inondations belges
27	2017_2018	Violences urbaines et crises climatiques
27	2018_2019	Inondations et crise sociale en Belgique
27	2019_2021	Violences urbaines et inondations en Belgique
28	2008_2010	Émeutes et tensions sociales urbaines
28	2010_2013	Violences urbaines et gestion de crise
28	2013_2015	Violences urbaines et gestion des crises
28	2015_2016	Violences urbaines et crise sociétale
28	2016_2017	Violences urbaines et crises sociales
28	2017_2018	Violences urbaines et gestion des crises
28	2018_2019	Violences urbaines et gestion municipale
28	2019_2021	Violences urbaines et enjeux communautaires
29	2008_2010	Violences urbaines et gestion de crise
29	2010_2013	Violences urbaines et tensions sociales à Bruxelles
29	2013_2015	Violences urbaines et tensions sociales
29	2015_2016	Violences urbaines et tensions policières
29	2016_2017	Violences urbaines et gestion de crise
29	2017_2018	Violences urbaines et crises sociales à Bruxelles
29	2018_2019	Violences urbaines et tensions sociales
29	2019_2021	Violence urbaine et enjeux sociaux à Bruxelles
30	2008_2010	Affrontements sociaux et intempéries
30	2010_2013	Gestion des crises et intempéries urbaines
30	2013_2015	Violences urbaines et gestion de crise
30	2015_2016	Violences urbaines et gestion policière
30	2016_2017	Violences urbaines et crise sociale
30	2017_2018	Violences urbaines et réponse policière
30	2018_2019	Violences urbaines et enjeux sociaux
30	2019_2021	Violences urbaines et réponses institutionnelles
31	2008_2010	Gestion des crises et conflits urbains
31	2010_2013	Gestion des crises hydrauliques et sociales
31	2013_2015	Violences urbaines et gestion des crises
31	2015_2016	Violence urbaine et gestion des crises
31	2016_2017	Gestion des crises et événements climatiques
31	2017_2018	Violences urbaines et gestion des crises
31	2018_2019	Crises hydrauliques et tensions sociales
31	2019_2021	Gestion des crises hydriques et sociales
32	2008_2010	Violences urbaines et gestion des crises
32	2010_2013	Violences urbaines et gestion des crises
32	2013_2015	Incendies et gestion des crises socio-environnementales
32	2015_2016	Incendies et émeutes en milieu urbain
32	2016_2017	Violences urbaines et gestion des crises
32	2017_2018	Violences urbaines et gestion des crises

Suite page suivante…

Topic	Période	Label dynamique
32	2018_2019	Violences urbaines et gestion des crises
32	2019_2021	Violences urbaines et gestion de crise
33	2008_2010	Violences urbaines et crise sanitaire
33	2010_2013	Violences urbaines et tensions sociales
33	2013_2015	Violences urbaines et crises sociales
33	2015_2016	Violences urbaines et crises climatiques
33	2016_2017	Violences urbaines et crises sociales
33	2017_2018	Violences urbaines et crises sociales
33	2018_2019	Violences urbaines et catastrophes climatiques
33	2019_2021	Violences urbaines et crise sociale

Table 6.5: Documents représentatifs par thématique — Modèle LDA ($K = 8$)

Topic	Label	Résumé du document représentatif
1	Catastrophes naturelles et impacts humains	Passage du cyclone Pam sur Vanuatu (catégorie 5) : destructions massives, villages entiers dévastés, communications interrompues, opérations humanitaires d'urgence depuis Port-Vila.
		Super typhon Mangkhut frappe Hong Kong et les Philippines : évacuation de millions de personnes, arbres arrachés, inondations graves, grand nettoyage urbain au lendemain du passage.
2	Gestion des inondations urbaines et projets	Programme de majorité communale : projets de rénovation de voirie, solutions aux inondations du centre-ville, bassins d'orage et canalisations prévus à moyen terme.
		Rénovation de la collégiale Saint-Croix : travaux de toiture et maçonnerie pour mettre l'édifice à l'abri des intempéries, première phase estimée à plusieurs millions d'euros.
3	Politique, médias et liberté d'expression	Cinquième anniversaire de l'attentat de Charlie Hebdo : réflexion sur la liberté de la presse, la censure croissante et la pression des réseaux sociaux sur les médias traditionnels.
		Processus de paix israélo-palestinien après la décision américaine sur Jérusalem : analyse des réactions diplomatiques et du rôle des États-Unis comme médiateur.
4	Santé publique et impacts environnementaux	Journée mondiale de la maladie cœliaque : diagnostic difficile, diversité des symptômes, carences nutritionnelles et prise en charge médicale en Belgique.
		Surmortalité en Belgique lors de la vague de chaleur de juin-juillet : analyse de Sciensano sur les décès supplémentaires par tranche d'âge, liés aux pics d'ozone.
5	Changements climatiques et agriculture durable	Vendanges précoces dans les Pyrénées-Orientales : effet du changement climatique sur les cépages, hausse des températures caniculaires, inquiétudes des viticulteurs sur les rendements.
		Bilan pluviométrique hivernal en Belgique : recharge des nappes phréatiques après déficit hydrique, situation proche de la normale, suivi via piézomètres du SPW.

Suite page suivante…

Topic	Label	Résumé du document représentatif
6	Catastrophes naturelles et pertes humaines	Crash d'un ATR taïwanais lors d'un vol intérieur : appareil s'écrase sur une autoroute, opérations de sauvetage d'urgence, bilan humain lourd.
		Naufrage meurtrier au Brésil sur le fleuve Xingu : dizaines de morts dans l'État du Pará, opérations de recherche et sauvetage, réaction du président brésilien.
7	Changements climatiques et impacts forestiers	Malgré la hausse des températures, la croissance des forêts ralentit en raison des maladies, incendies et sécheresses. Appel à diversifier les monocultures en Belgique et dans le monde.
		Les forêts tropicales, autrefois puits de carbone, deviennent émettrices nettes en raison de la déforestation et des sécheresses, selon plusieurs études publiées dans Nature.
8	Interventions des pompiers en intempéries	Violents orages en Hainaut : grêlons, arbres abattus, caves inondées, dizaines d'interventions des pompiers, perturbations ferroviaires et évacuation du parc Pairi Daiza.
		Inondations à Ittre et dans le Brabant wallon : plan communal d'urgence activé, pompiers mobilisés toute la nuit, routes fermées, dizaines de caves inondées dans plusieurs communes.

Table 6.6: Évolution thématique des topics DTM par période

Topic	Période	Label dynamique
0	2008–2010	Gestion des catastrophes naturelles et urbanisme
0	2010–2013	Gestion des catastrophes naturelles locales
0	2013–2015	Gestion des catastrophes et évacuations
0	2015–2016	Gestion des risques environnementaux et urbains
0	2016–2017	Incendies et impacts socio-économiques
0	2017–2018	Gestion des catastrophes et impacts climatiques
0	2018–2019	Changements climatiques et impacts environnementaux
0	2020–2021	Crise socio-économique et environnementale actuelle
1	2008–2010	Gestion des crises humanitaires régionales
1	2010–2013	Gestion des crises et développement territorial
1	2013–2015	Gestion des crises et ressources locales
1	2015–2016	Dynamique sociale et territoriale américaine
1	2016–2017	Circulation et gestion des infrastructures urbaines
1	2017–2018	Gestion des crises environnementales et humaines
1	2018–2019	Gestion des crises et accidents locaux
1	2020–2021	Gestion des crises climatiques et économiques
2	2008–2010	Impact des catastrophes climatiques
2	2010–2013	Impact des intempéries sur la population
2	2013–2015	Impact des catastrophes naturelles locales
2	2015–2016	Gestion des catastrophes météorologiques
2	2016–2017	Changements climatiques et catastrophes naturelles
2	2017–2018	Changements climatiques et impacts environnementaux

Suite page suivante…

Topic	Période	Label dynamique
2	2018–2019	Incidences climatiques et gestion des crises
2	2020–2021	Gestion des risques d'incendie et inondation
3	2008–2010	Gestion des catastrophes et secours
3	2010–2013	Gestion des crises et catastrophes naturelles
3	2013–2015	Gestion des crises environnementales locales
3	2015–2016	Gestion des crises et incendies urbains
3	2016–2017	Gestion des crises et interventions d'urgence
3	2017–2018	Gestion des crises environnementales et urbaines
3	2018–2019	Incendies et impacts environnementaux locaux
3	2020–2021	Incendies et impacts environnementaux locaux
4	2008–2010	Gestion des catastrophes naturelles et impacts
4	2010–2013	Gestion des crises et impacts sociaux
4	2013–2015	Changements climatiques et impacts sociaux
4	2015–2016	Gestion des risques environnementaux et économiques
4	2016–2017	Gestion des crises et populations vulnérables
4	2017–2018	Gestion des crises climatiques et sociales
4	2018–2019	Impact du changement climatique sur la biodiversité
4	2020–2021	Gestion des crises climatiques et services
5	2008–2010	Gestion des urgences climatiques régionales
5	2010–2013	Gestion des crises et politiques locales
5	2013–2015	Changements climatiques et interventions humanitaires
5	2015–2016	Dynamiques territoriales et projets de développement
5	2016–2017	Gestion des infrastructures face aux crises
5	2017–2018	Impact des événements climatiques sur la population
5	2018–2019	Mobilité et gestion des services publics
5	2020–2021	Gestion des crises environnementales urbaines
6	2008–2010	Gestion des urgences et catastrophes naturelles
6	2010–2013	Gestion des crises environnementales et politiques
6	2013–2015	Gestion des catastrophes naturelles locales
6	2015–2016	Gestion des crises environnementales et sociales
6	2016–2017	Gestion des crises environnementales locales
6	2017–2018	Gestion des crises environnementales et sociales
6	2018–2019	Gestion des crises environnementales en Belgique
6	2020–2021	Gestion des ressources et enjeux sociaux
7	2008–2010	Gestion des crises environnementales et sociales
7	2010–2013	Incendies et impacts socio-économiques
7	2013–2015	Gestion des crises et catastrophes naturelles
7	2015–2016	Gestion des crises environnementales locales
7	2016–2017	Gestion des crises liées aux intempéries
7	2017–2018	Gestion des risques environnementaux et sociaux
7	2018–2019	Gestion des risques météorologiques en Belgique
7	2020–2021	Gestion des risques environnementaux et sociaux

Table 6.3: Aperçu des articles extraits par catégorie (hors RTBFINFO)

Catégorie	Article	Aperçu du contenu
OTHER	A1	Cérémonie des Magritte, interview de Marie Gillain et sorties cinéma (Eric & Ramzy, Deadpool, Zootopie).
	A2	Documentaire BBC sur Harvey Weinstein : silence des victimes, accords de non-divulgation, complicités.
RADIO	A1	Portrait du compositeur Francis Poulenc, CD de la semaine proposé à l'écoute et au concours.
	A2	Projet pharaonique de centre européen de sports de glisse à Antoing, porté par le Prince de Ligne.
RTBFCULTURE	A1	Résultats d'un concours de violon : programme joué par un candidat de la première épreuve à la finale.
	A2	Présentation de la série *Unbreakable Kimmy Schmidt* (Netflix), produite par Tina Fey.
RTBFSPORT	A1	Essais libres du Grand Prix du Japon F1 : Hamilton le plus rapide, Vandoorne de retour en piste.
	A2	Portrait de Felice Mazzu, entraîneur du Sporting de Charleroi, et ses ambitions futures.
RTBFTENDANCE	A1	Chronologie historique des droits LGBT, de la dépénalisation à la Gay Pride de Paris.
	A2	Sélection de produits cosmétiques de fin d'année (sérums, baumes, bases matifiantes).
TV	A1	Résumé d'épisode de soap opera : intrigues autour de Phyllis, Victoria et une liaison secrète.
	A2	Lancement de l'émission estivale *Stoemp, Pèkèt…et des rawettes!* avec Fanny Jandrain.

Table 6.4: Labels thématiques du modèle LDA ($K = 8$) générés via GPT-4o-mini

Topic	Label	Mots-clés représentatifs
1	Catastrophes naturelles et impacts humains	inondation, mort, personne, ville, état
2	Gestion des inondations urbaines et projets	commun, projet, certain, cas, ville
3	Politique, médias et liberté d'expression	pays, politique, gouvernement, président, grand
4	Santé publique et impacts environnementaux	cas, personne, hôpital, enfant, maladie
5	Changements climatiques et agriculture durable	année, température, jour, agriculteur, chaleur
6	Catastrophes naturelles et pertes humaines	mort, ville, région, personne, inondation
7	Changements climatiques et impacts forestiers	année, rapport, scientifique, nbr_million, pays
8	Interventions des pompiers en intempéries	pompier, incendie, orage, région, heure

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des sciences
Place des sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/sc**