Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Traitement automatique des évènements médiatiques : Détection, classification, segmentation et recherche sémantique

Abdelkrim Beloued
Institut national de l'audiovisuel, 4 avenue de l'Europe, 94360 Bry-sur-Marne, France
abeloued@ina.fr

RÉSUMÉ

Cet article présente une méthodologie pour l'analyse automatique des évènements rapportés par les médias. Elle s'appuie sur des techniques de traitement automatique des langues, notamment la représentation sémantique des contenus médiatiques, la classification thématique, l'extraction d'évènements à partir de flux d'information, ainsi que la détection d'évènements par regroupement de représentations vectorielles issues de modèles de plongement sémantique. L'approche combine des modèles supervisés et non supervisés ainsi que des architectures capables de prendre en compte un contexte large. Plusieurs corpus sont utilisés pour l'entraînement et l'évaluation de ces modèles. Les résultats obtenus montrent une efficacité élevée dans la détection, le regroupement, la classification thématique et la recherche sémantique des évènements médiatiques. Cette approche offre ainsi des perspectives significatives pour structurer les faits réels, analyser leur représentation médiatique et comprendre l'influence exercée par les médias sur le traitement de ces faits.

ABSTRACT

Automatic Processing of Media Events : Detection, classification, segmentation, and semantic search

This article presents a methodology for the automatic analysis of events as reported in the media. It relies on natural language processing techniques, including contextual semantic representation, thematic classification, event extraction via segmentation, and event detection through clustering of vector representations derived from semantic embedding models. The approach integrates both supervised and unsupervised models, as well as architectures specifically designed for processing long-range contexts. Various datasets are used for training and evaluating the models. Experimental results demonstrate high performance in event detection, clustering, thematic classification, and semantic retrieval. This methodology offers significant opportunities for structuring factual information, analyzing its representation in the media, and understanding the influence of media on how events are reported.

MOTS-CLÉS : Évènements médiatiques, Apprentissage contrastif, Détection d'évènements, Classification, Segmentation, Recherche sémantique.

KEYWORDS: Media events, Contrastive learning, Event detection, Clustering, Classification, Segmentation, Semantic search.

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux, pages 741–755.

1 Introduction

Dans un contexte de surabondance d'informations et de multiplication des canaux médiatiques, l'analyse des évènements médiatiques revêt une importance essentielle. Cette analyse permet de : (1) croiser et harmoniser les récits médiatiques afin d'obtenir une vision plus cohérente des faits ; (2) comprendre l'influence des médias sur l'opinion publique et la perception des évènements ; (3) étudier la divergence de traitement médiatique d'un même évènement selon les différentes sources médiatiques ; (4) reconstituer les faits réels en dépassant le prisme médiatique, permettant ainsi à chacun de se forger une opinion indépendante ; (5) faciliter l'extraction et la recherche d'évènements médiatiques au sein d'un flux d'informations continu ; (6) aligner les évènements médiatiques avec des référentiels d'évènements connus, comme l'AFP (Agence France-Presse).

Un évènement se définit comme étant un fait ou une situation d'ordre culturel, politique ou social susceptible d'intéresser le public. Sa médiatisation joue un rôle central dans sa diffusion, influençant ainsi son impact et la manière dont il est perçu. Le rôle des médias est central dans la construction d'un évènement médiatique. Ils sélectionnent, interprètent et mettent en récit les faits, jouant ainsi un rôle de filtre qui influence et façonne la perception collective. En fonction de leur ligne éditoriale, ils décident de la portée et de l'importance accordées à l'évènement, contribuant à sa transformation en "évènement médiatique". Ainsi, un évènement médiatique se caractérise par son degré de médiatisation, son impact sur le public, sa temporalité et sa granularité.

Le degré de médiatisation détermine la visibilité d'un évènement dans l'espace médiatique. Certains faits restent **isolés**, relayés brièvement par les médias et localisés dans le temps et dans l'espace, sans réelle amplification. C'est le cas, par exemple, d'un incendie dans un petit village sans conséquences majeures. D'autres évènements bénéficient d'une large couverture médiatique (télévision, presse, réseaux sociaux, etc.) comme la cérémonie d'ouverture des jeux olympiques, planifiée et largement diffusées. Ce type d'évènements est qualifié d'évènement **médiatisé**. Enfin, certains évènements connaissent une amplification médiatique intense en peu de temps, parfois déconnectée de l'importance réelle du fait et complètement imprévue, provoquant une **tempête médiatique**, comme, l'affaire des "Homards de François de Rugy". **L'impact** d'un évènement médiatique dépend de sa capacité à influencer l'opinion publique et à générer des débats ou des émotions collectives. Certains faits marquants s'inscrivent durablement dans la mémoire collective, comme les attentats. **La temporalité** médiatique des évènements peut varier : certains sont ponctuels, rapidement relayés par les médias puis oubliés, tandis que d'autres s'étendent sur une période prolongée, souvent marqués par des variations d'intensité médiatique selon les rebondissements. Par exemple, une crise politique qui évolue en plusieurs phases. Enfin **la granularité** d'un évènement médiatique peut être analysée sous différentes formes. Certains faits existent de manières isolées et uniques, tandis que d'autres s'imbriquent dans une série d'évènements ou des faits connexes traités comme un ensemble. Par exemple, une crise politique avec plusieurs rebondissements. Cette classification des évènements (isolé, médiatisé, viral et tempête médiatique), basée sur leur temporalité et leur degré de médiatisation, permet d'adapter les traitements automatiques et les approches d'analyse aux spécificités de chaque catégorie.

Un fait réel peut être relayé sous différentes formes (représentations) : un extrait audio/vidéo, un article de presse, une dépêche AFP, un document Wikipédia, un tweet, etc. Il peut être représenté par une spécification unique, par exemple sous la forme d'une entité nommée dans une base de connaissance comme Wikidata ou DBpedia. A l'INA (Institut national de l'audiovisuel), nous utilisons un référentiel interne appelé la **grille d'indexation**, associée aux notices évoquant un évènement donné. Cette appellation reviendra souvent dans la suite de cet article. Plusieurs tâches s'imposent

pour l'analyse des évènements médiatiques : l'extraction qui vise à repérer les segments pertinents dans les transcriptions textuelles ; la détection, réalisée par regroupement (*clustering*) de ces extraits ; l'identification qui consiste à les associer à des références connues permettant une reconstitution plus fidèle des faits réels à partir de leurs différentes représentations médiatiques ; enfin, la recherche sémantique, basée sur le calcul de similarités entre évènements, permet une exploration plus fine et pertinente de ces derniers.

2 État de l'art

La détection et l'analyse des évènements médiatiques ont fait l'objet de nombreuses recherches dans différents contextes, exploitant diverses approches en traitement du langage naturel et en apprentissage automatique. Mazoyer (Mazoyer, 2020) propose une approche de détection des évènements à partir de documents hétérogènes tels que les tweets et les articles de presse en ligne. Ce travail met en évidence la propagation de l'information entre les réseaux sociaux et les médias traditionnels, les relations entre la popularité d'un évènement sur Twitter et sa couverture par les médias, ainsi que l'impact des réseaux sociaux sur la production d'informations dans les médias. L'approche repose sur des algorithmes de clustering dynamique comme **First Story Detection** (Allan, 2002) et la **détection de communauté** (Traag *et al.*, 2015) ainsi que l'utilisation des représentations textuelles générées par **TF-IDF**, **Word2Vec** (Nguyen *et al.*, 2015) et **BERT** (Devlin *et al.*, 2019). L'auteur de (Bernard, 2022) s'intéresse à l'analyse et au suivi des évènements historiques à travers la presse. Son travail repose sur le regroupement des articles décrivant un même évènement à l'aide de représentations vectorielles et d'algorithmes de clustering (**K-means** (MacQueen, 1967), **Louvain** (Blondel *et al.*, 2008)) intégrant des fenêtres temporelles pour mieux capter les relations entre évènements. Des représentations vectorielles (**TF-IDF**, **BERT**, **Sentence-BERT** (Reimers & Gurevych, 2019)) sont utilisées pour le clustering. Une approche de recherche enrichie basée sur des requêtes structurées (Quoi ? Quand ? Où ? et Qui ?) est également proposée, en s'appuyant sur des outils comme **ElasticSearch** et **Okapi BM25** (Robertson *et al.*, 1994). Les auteurs de (Markus *et al.*, 2024) proposent une méthodologie combinant apprentissage automatique pour la détection d'anomalies et validation humaine pour détecter les tempêtes médiatiques à partir d'un large corpus d'articles de presse américaines couvrant la période 1996-2016. L'approche repose sur l'extraction d'entités nommées, l'analyse thématique (*Tropic Modeling*) et la détection d'éléments narratifs (complication, résolution, succès) à l'aide du modèle **NEAT** (Levi *et al.*, 2022), ainsi que d'embeddings textuels à l'aide du modèle **all-mpnet-base-v2** [1]. L'algorithme non supervisé **Prophet** (Taylor & Letham, 2018) est utilisé pour détecter les anomalies et les pics dans les signaux analytiques. Tarekegn (Tarekegn, 2024) utilise des **LLM** sur la base **GDELT** (Leetaru & Schrodt, 2013) pour détecter, regrouper et résumer les évènements. L'approche consiste à appliquer un pré-traitement des données avec **KeyBERT** (Grootendorst, 2020) pour l'extraction de mots-clés, suivi du calcul d'embeddings textuels avec le modèle **text-embedding-ada-002** [2]. En post-traitement, le clustering est réalisé avec **K-means** et **hdbscan** (McInnes *et al.*, 2017), et les évènements sont résumés automatiquement à l'aide du **LLM** utilisé, avec un étiquetage thématique utilisant **GPT-3.5-turbo-instruct** basé sur le référentiel **IPTC** [3]. Les auteurs de (Ishlach *et al.*, 2024) utilisent les embeddings pour structurer les évènements du corpus GDELT. Des modèles d'extraction d'entités nommées et de thématisation sont également utilisés pour extraire des informations clés et identifier les sujets des articles (ex. brutalité policière, élection présidentielle). Un regroupement

1. https://huggingface.co/sentence-transformers/all-mpnet-base-v2
2. https://huggingface.co/Xenova/text-embedding-ada-002
3. https://www.iptc.org/std/NewsCodes/treeview/mediatopic/mediatopic-en-GB.html

temporel (ex. par mois) est effectué pour garantir que les embeddings capturent le contexte pertinent. Deux approches d'embeddings sont comparées : **SIF**(Smooth Inverse Frequency) (Arora *et al.*, 2017) et un réseau siamois. Cette revue de l'état de l'art met en évidence la diversité des approches explorées pour la détection et l'analyse des évènements médiatiques. Toutefois, plusieurs limites communes émergent, notamment l'absence d'évaluation sur des corpus audiovisuels et l'exploitation partielle des avancées récentes en traitement du langage naturel. Ces observations soulignent la nécessité de recherches complémentaires pour améliorer les performances des modèles sur ces tâches.

3 Ensembles de données

3.1 Corpus de référence

Notre approche pour l'entraînement des modèles dédiés à l'analyse des événements médiatiques repose sur divers corpus couvrant différents formats et sources médiatiques. Nous avons constitué nos datasets à partir de quatre sources aux natures différentes : les bases spécialisées en évènements (comme l'AFP), les articles issus du scraping web (presse écrite et en ligne), les collections audiovisuelles (notamment celles de l'INA), et les bases de connaissances génériques (comme DBpedia). L'exploitation conjointe de ces données permet de construire des modèles robustes assurant une analyse fine et contextualisée des évènements médiatiques.

AFP La quasi-totalité des évènements médiatiques nationaux et internationaux relayés par les médias audiovisuels, la presse écrite et la presse en ligne trouvent leur origine dans une dépêche AFP. Ce corpus, issu du projet OTMedia (Hervé *et al.*, 2013) et aligné sur le référentiel **IPTC**, constitue une source de données fiable, essentielle pour l'analyse du traitement médiatique des évènements.

Presse écrite et presse en ligne Ce corpus, hétérogène et non rattaché à un référentiel spécifique, regroupe des documents et articles issus des chaînes de radio (RFI, France Bleu, ect.), des journaux (Le Monde, 20-MINUTES, etc.) et des chaînes de télévision (France-Info, BFM-TV, etc.). Il est particulièrement adapté à l'apprentissage non-supervisés.

Collections audiovisuelles de l'INA qui sont utilisées pour l'évaluation des modèles proposés. Deux types de collections nous intéressent particulièrement dans ce travail : (1) les collections de contenus liés à un ou plusieurs évènements médiatiques, parmi lesquelles nous avons utilisé **ARTE_INFO** (sélection de journaux télévisés diffusés sur Arte) et **GRILLE** (notices documentaires rattachées à une grille d'indexation). Une grille d'indexation représente une entité nommées de type évènement dans le référentiel interne de l'INA, pouvant être associée à plusieurs notices documentaires. En revanche, une notice documentaire ne peut être associée qu'à une seule grille. (2) Les **collections NON_EVENT** qui Regroupe des contenus sans lien avec des évènements médiatiques, axés notamment sur le documentaire et le reportage ("Échappées belles", "Invitation au voyage", etc).

DBpedia Les données issues de DBpedia enrichissent l'analyse en apportant un corpus conséquent d'entités nommées de type évènement, obtenu via des requêtes SPARQL.

3.2 Similarité sémantique

La similarité entre textes permet de constituer des Datasets destinés au fine-tuning de modèles contrastifs, sous forme de paires de phrases associées à un score de similarité. Trois types peuvent être distingués : (1) la similarité textuelle qui est basée uniquement sur la distance entre deux textes, et sert à entraîner des modèles génériques ; (2) la similarité thématique, calculée à partir des catégories attribuées aux textes, est adaptée à l'apprentissage de modèles orientés thématiques ; (3) la similarité évènementielle qui combine les critères de similarité textuelle, thématique et entités nommées, convient particulièrement aux modèles dédiés à la détection et l'analyse d'évènements médiatiques.

Notre approche pour calculer la similarité thématique est hybride et combine graphe de connaissances et embeddings. Elle tient compte de la hiérarchie des concepts associés aux texte pour construire des embeddings. Ainsi, les concepts associés à chaque texte sont organisés en hiérarchie. Seuls les concepts terminaux sont pris en compte dans un premier temps, chacun étant initialement pondéré avec un poids de 1. Les poids sont ensuite diffusés à travers les relations hiérarchiques selon une fonction décroissante. Chaque texte est alors représenté sous forme d'un vecteur dans l'espace conceptuel hiérarchisé. Une fois ces représentations vectorielles obtenues, la similarité cosinus est utilisée pour mesurer la distance entre deux textes. Cette approche permet de capturer (1) les relations hiérarchiques où les concepts parent-enfant sont associés à des représentations vectorielles proches dans l'espace latent, (2) les relations de proximité topologique où deux concepts proches dans la structure du graphe auront des vecteurs similaires grâce à l'intégration des pondérations de leurs ancêtres, (3) les relations analogiques en exploitant les liens de similarité qui unissent certains concepts. Les modèles supervisés sont ensuite entraînés à reproduire ces schémas de similarité sur des corpus non annotés.

3.3 Datasets utilisés

Les dépêches AFP, avec environ 263 000 publications en 2018, constituent une ressource essentielle pour l'analyse des évènements médiatiques. Bien qu'il soit théoriquement possible de générer jusqu'à 34 milliards de paires potentielles, la majorité de ces paires concerne des dépêches différentes. Pour équilibrer cette disparité et rendre le corpus exploitable, nous avons réduit le volume initial en privilégiant les paires les plus pertinentes. Les dépêches ont été regroupées par périodes (mois, semaines, jours, heures) afin de limiter le calcul de similarité aux combinaisons internes. Le choix final s'est porté sur le regroupement par période de deux semaines qui présente le score de similarité le plus faible (0.07) parmi toutes les options. Après échantillonnage, nous avons obtenu un ensemble de 242 millions de paires, puis extrait des ensembles équilibrés pour l'entraînement (76 645 paires) et l'évaluation (13 765 paires).

Les évènements extraits de DBpedia apportent un corpus riche de plus de 38 209 entités, générant jusqu'à 730 millions de paires potentielles. Un échantillonnage par périodes, basé sur les scores de similarité, a permis de réduire ce volume. Le regroupement par tranches de 20 ans a été retenu, ramenant le total à environ 55 millions de paires. Un équilibrage complémentaire a ensuite permis de former deux ensembles homogènes pour l'entraînement (22 340 paires) et l'évaluation (6 160 paires).

Les données INA apportent une dimension audiovisuelle essentielle pour l'analyse des évènements et sont utilisées à des fins d'évaluation. La grille d'indexation de l'INA comprend 4 649 entrées, générant près de 11 millions de paires exploitables. En complément, les notices liées à une grille d'indexation totalisent 9 207 entrées et plus de 42 millions de paires associées. Trois ensembles sont constitués pour l'évaluation de nos modèles : **GRILLE**, **ARTE_INFO** et Collection **NON_EVENT**.

4 Architectures mises en œuvre

L'entraînement des modèles pour l'analyse des événements médiatiques s'appuient sur différentes architectures adaptées aux besoins spécifiques des tâches supervisées, non supervisées et au contexte large. Ces modèles visent à constituer un espace d'embeddings adapté aux évènements médiatiques. Pour cela, nous avons porté une attention particulière aux modèles contrastifs qui permettent d'apprendre des représentations plus discriminantes en mettant en contraste des paires de phrases.

4.1 Modèles supervisés et non-supervisés

Les modèles supervisés utilisés dans nos expérimentations reposent sur des architectures de type Sentence-Transformer (Reimers & Gurevych, 2019), conçues pour apprendre des représentations textuelles optimisées à partir de paires de phrases. Nous avons opté pour l'architecture Bi-Encoder, qui encode séparément chacune des phrases et évalue leur similarité. Deux modèles ont été entraînés sur le dataset AFP (3.3) : Le premier, spécialisé en évènement médiatique, est entraîné sur des paires de dépêches AFP avec un score calculé à partir des trois dimensions de similarité (3.2) : thématique (concepts IPTC), entités nommées et contenu textuel des dépêches. Le second modèle, axé sur la classification thématique, s'appuie uniquement sur la similarité thématique entre concepts IPTC. Il est présenté en détail dans la section 4.3.

Concernant les modèles non-supervisés, nous avons développé une implémentation simplifiée du modèle SimCSE(Gao *et al.*, 2021) en utilisant camembert-base comme modèle de base. Une couche linéaire a été ajoutée afin de projeter les représentations extraites du modèle pré-entraîné dans un nouvel espace vectoriel de même dimension, permettant ainsi d'affiner les embeddings et de les adapter à l'objectif contrastif. Cette couche est désactivée à l'inférence, permettant une extraction directe des représentations vectorielles optimisées. Le modèle a été entraîné sur les dépêches AFP ainsi que les articles de la presse écrite et en ligne, aboutissant à un modèle non-supervisé spécialisé en évènements médiatiques.

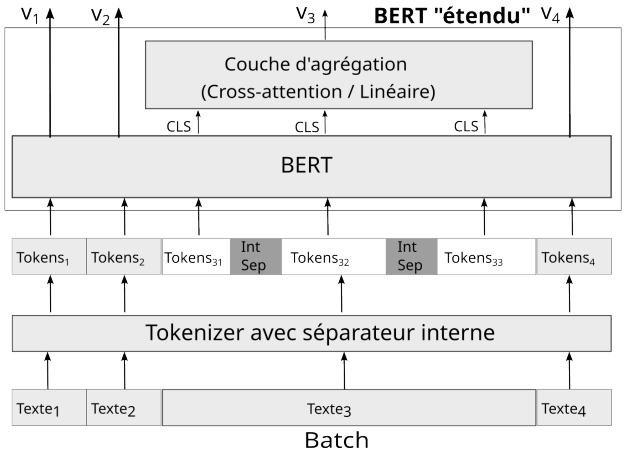
4.2 Modèles pour contexte large

Les modèles précédents présentent une limitation liée à la gestion du contexte large. Ils ne prennent pas en compte l'intégralité du texte, ce qui peut s'avérer problématique pour certaines tâches nécessitant une compréhension approfondie du contenu global, notamment la segmentation thématique des transcriptions, où une représentation complète du contexte est essentielle. Bien que dans la plupart des documents d'actualités, l'essentiel de l'information soit concentré au début du texte, rendant ces modèles capables d'extraire les éléments les plus pertinents, certaines tâches requièrent néanmoins une prise en compte plus large du contexte pour capturer la sémantique d'un texte.

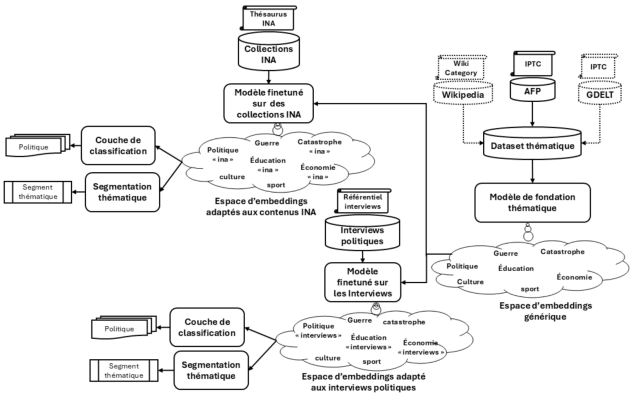
Afin de pallier cette limitation, nous avons étendu le contexte de ces modèles afin d'analyser leur capacité à capturer des relations sémantiques sur de longues séquences. Les modèles existants comme Longformer(Beltagy *et al.*, 2020) et BigBird(Zaheer *et al.*, 2020) ont été conçus pour traiter des séquences étendues en combinant plusieurs mécanismes d'attention : locale, globale et aléatoire. Cependant, ces modèles ne s'intègrent pas facilement aux architectures de type sentence-transformer, limitant ainsi leur utilisation pour l'apprentissage contrastif. Pour remédier à cette contrainte, nous avons proposé une nouvelle architecture, capable de gérer les longues séquences tout en s'intégrant

facilement aux architectures contrastives. L'approche repose sur un découpage des séquences en blocs de taille fixe (128, 256, 512 tokens), avec ou sans fenêtre glissante. Chaque bloc est ensuite traité indépendamment par une instance de BERT qui en génère une représentation vectorielle. Une couche d'agrégation est ajoutée pour capturer les interactions entre les blocs et fusionner leurs représentations en un vecteur unique représentant l'ensemble de la séquence.

Nous avons encapsulé cette architecture dans un modèle, compatible avec BERT, que nous appelons LCLM : *Large-Context Language Model* (voir la figure 1a). Pour cela, nous avons modifié le tokenizer pour introduire un nouveau token (INTERNAL_SEP), servant de séparateur interne entre les différents blocs d'une même séquence. Chaque bloc est traité indépendamment par BERT et les représentations vectorielles obtenues sont transmises à une couche d'agrégation qui fusionne ces informations en un vecteur unique de représentation globale. La Couche d'agrégation de contextes peut être soit linéaire, soit basée sur une couche cross-attention. La couche linéaire projette les représentations et réduit les distances dans l'espace latent, la rendant particulièrement adaptée à la segmentation des textes. En revanche, la cross-attention permet de capturer des relations complexes entre différentes parties du texte, ce qui la rend plus efficace pour les tâches de détection et de recherche sémantique. Dans nos expérimentations, nous avons priviligié une couche linéaire, et avons ainsi fine-tuné deux modèles : LCLM-Bi-Encoder (spécialisé en évènements et thématisation) et LCLM-SimCSE.



(a) Architecture du modèle LCLM



(b) Adaptation d'un modèle thématique générique à des corpus spécifiques

FIGURE 1 – Architecture étendue des modèles proposés

4.3 Modèles de classification thématique

Les émissions de télévision et de radio offrent généralement un contexte limité des évènements, en particulier dans certains formats comme les interviews politiques, où les événements sont souvent évoqués sans être détaillés. Cette caractéristique constitue un défi lorsqu'il s'agit d'affiner un modèle pour l'analyse d'événements à partir de ces données, car la représentation des événements peut manquer de précision contextuelle. Pour surmonter cette limitation, nous avons proposé d'entraîner un modèle générique de catégorisation, que l'on peut qualifier de modèle de fondation thématique, en s'appuyant sur des bases d'événements riches et validées, telles que l'AFP, DBpedia et GDELT comme illustré dans la figure 1b. Ce modèle générique peut ensuite être affiné sur des domaines

particuliers. Par exemple, sur des collections de l'INA annotées à l'aide d'un thésaurus internes des noms communs ou encore des corpus d'interviews politiques, annotés avec un référentiel spécialisé. Cette approche permet ainsi de transférer les connaissances acquises à partir de bases factuelles étendues vers des domaines plus spécifiques. Le modèle envisagé est un modèle supervisé, entraîné sur des paires de textes décrivant des événements médiatiques. La similarité entre deux textes est évaluée exclusivement sur la base de leur proximité thématique, c'est-à-dire en comparant les concepts qui leur sont associés comme présenté dans la section 3.2. Ce modèle vise à construire un espace d'embeddings thématique générique, où les événements sont représentés et organisés selon les concepts qui leur sont reliés. Ce modèle intègre des connaissances issues de bases de données spécialisées et du Linked Open Data et peut être affiné sur des corpus spécialisés propres à des domaines spécifiques.

5 Fine-tuning et évaluation des modèles

5.1 Fine-tuning

Nous avons fine-tuné ces modèles sur nos datasets. Les résultats du finetuning montrent des performances variables selon les architectures utilisées. Le Bi-Encoder obtient des scores élevés avec un coefficient de Spearman et Pearson de 0.95, témoignant d'une excellente corrélation entre les similarités prédictives et les annotations manuelles. L'extension LCLM-Bi-Encoder atteint des performances comparables avec un Spearman de 0.93. Pour les modèles non supervisés, SimCSE atteint un score de Spearman de 0.76, tandis que son extension adaptée aux longues séquences, LCLM-SimCSE, obtient un score légèrement inférieur à 0.74. Ces résultats démontrent que les modèles supervisés surpassent généralement les modèles non supervisés en termes de corrélation avec les annotations manuelles. Cependant, les modèles non supervisés restent une alternative viable lorsqu'aucune annotation de similarité n'est disponible.

5.2 Évaluation

L'évaluation des modèles fine-tunés permet de mesurer leur efficacité à détecter, extraire, identifier et classifier les événements médiatiques. Les jeux de données utilisées pour cette évaluation incluent principalement des données INA tels que présentés dans la section 3.3. Le modèle de référence choisi pour comparer nos résultats est le modèle Dangvantuan [4].

5.2.1 Corrélation des prédictions

L'évaluation de la corrélation des prédictions repose sur la mesure de la similarité entre les annotations manuelles et les prédictions des modèles. Les jeux de données utilisés pour cette évaluation sont les datasets **ARTE_INFO** et **GRILLE**. La similarité sémantique est calculée pour chaque paire de notices documentaires en combinant les trois types de similarités selon la méthode présentée dans la section 3.2. Par la suite, la similarité entre les embeddings des transcriptions de chaque paire de notices est calculée. Enfin, les métriques de corrélation de Pearson et de spearman (voir le tableau 1) sont utilisées afin d'évaluer la corrélation entre les distributions des deux scores de similarités. Les résultats

4. https://huggingface.co/dangvantuan/sentence-camembert-large

montrent que le modèle SimCSE affiche de meilleures performances sur les deux datasets, indiquant une forte capacité à aligner les prédictions avec les annotations manuelles. Le Bi-encoder, quant à lui, présente une bonne corrélation, bien que légèrement inférieure à celle de SimCSE. Le modèle LCLM-SimCSE obtient une corrélation jugée acceptable, témoignant d'une performance modérée dans la capture des similarités sémantiques. Enfin, les modèles LCLM-Bi-Encoder et dangvantuan affichent des corrélations moyennes, suggérant une capacité réduite à reproduire avec précision les relations sémantiques présentes dans les annotations manuelles.

Dataset	Algorithm	Pearson	Spearman
GRILLE	dangvantuan	59.2867	57.8999
GRILLE	bi-encoder	75.2592	76.7257
GRILLE	lclm-bi-encoder	65.8061	61.6149
GRILLE	simcse	**75.9466**	**79.8427**
GRILLE	lclm-simcse	71.5427	75.7982
ARTE_INFO	dangvantuan	56.6811	54.6432
ARTE_INFO	bi-encoder	71.2997	69.7191
ARTE_INFO	lclm-bi-encoder	51.894	47.9386
ARTE_INFO	simcse	**73.5886**	**73.3609**
ARTE_INFO	lclm-simcse	67.7815	65.2361

TABLE 1 – corrélation des prédictions

modèle	community_detection	dbscan	hierarchical
Accuracy			
dangvantuan	0.9972	0.9963	0.9965
bi-encoder	0.9985	0.998	0.9992
simcse	0.9986	0.9983	0.9965
Precision			
dangvantuan	1.0	1.0	1.0
bi-encoder	0.7064	0.618	0.8599
simcse	0.7722	0.8153	1.0
Recall			
dangvantuan	0.2505	0.1765	0.1818
bi-encoder	0.6539	0.619	0.7398
simcse	0.6316	0.5396	0.1818
F1-score			
dangvantuan	0.4007	0.3	0.3077
bi-encoder	0.6791	0.6185	**0.7953**
simcse	0.6948	0.6494	0.3077
Notices Clusterisées (%)			
dangvantuan	4.7188%	0.2359%	0.3146%
bi-encoder	52.2218%	44.475%	**67.0861%**
simcse	47.1844%	33.5037%	0.3146%
Grilles reconstituées (%)			
dangvantuan	10.303%	0.6061%	0.8081%
bi-encoder	64.4444%	54.3434%	**82.2222%**
simcse	57.7778%	43.4343%	0.8081%

TABLE 2 – Clustering

Dataset	Algorithm	Mean Similarity
GRILLE	dangvantuan	0.5165
GRILLE	bi-encoder	0.6312
GRILLE	lclm-bi-encoder	0.5182
GRILLE	simcse	0.7206
GRILLE	lclm-simcse	**0.8480**
ARTE_INFO	dangvantuan	0.5879
ARTE_INFO	bi-encoder	0.6601
ARTE_INFO	lclm-bi-encoder	0.5511
ARTE_INFO	simcse	0.7830
ARTE_INFO	lclm-simcse	**0.8818**

TABLE 3 – Proximité vectorielle

Nous avons également comparé les performances des modèles en fonction des datasets, les résultats indiquent que les performances sont meilleures sur les notices associées à une grille d'indexation que sur celles issues du corpus ARTE_INFO. Cette différence peut s'expliquer par la présence, en quantité faible, de notices documentaires ne correspondant pas à des évènements dans le corpus ARTE_INFO, alors que les notices du second corpus sont toutes des évènements et ont été classées manuellement. Par conséquent, les modèles ont tendance à se spécialiser davantage dans la détection des évènements.

5.2.2 Détection d'évènements médiatisés : Clustering

L'expérimentation vise à reconstituer les grilles d'indexation de l'INA à partir des résumés des notices documentaires, afin d'évaluer la capacité des modèles à regrouper correctement ces notices selon des similarités sémantiques. Le jeu de données utilisé est composé de notices documentaires associées aux grilles d'indexation (le dataset **GRILLE**). L'approche consiste à encoder le champ "résumé" à l'aide des modèles fine-tunés, et à appliquer ensuite différents algorithmes de clustering afin d'identifier des regroupements pertinents. Plusieurs algorithmes ont été testés pour le clustering des embeddings : Dbscan (Ester *et al.*, 1996), détection de communautés(Traag *et al.*, 2015) et clustering hiérarchique(Miyamoto, 2012), tandis que K-means s'est révélé inadapté à cette tâche à cause du nombre fixe de clusters (k). Les clusters obtenus ont été comparés aux grilles d'indexation de référence afin d'évaluer leur pertinence. Les performances des modèles ont été mesurées à l'aide des métriques suivantes : Prcision, Recall, F1-score, nombre de grilles reconstituées et nombre de notices regroupées comme présenté dans le tableau 2. Les résultats indiquent que l'approche Bi-encoder, combinée à la méthode de clustering hiérarchique, offre de meilleures performances en termes de cohérence des regroupements. SimCSE, associé à la méthode de détection de communauté, obtient des performances moyennes. En revanche, le modèle Dangvantuan affiche des résultats inférieurs, montrant une faible capacité à structurer efficacement les notices en clusters pertinents.

5.2.3 Détection d'évènements médiatisés : Proximité vectorielle

L'objectif de cette expérimentation est d'évaluer la proximité des vecteurs d'embeddings de plusieurs textes décrivant le même événement dans l'espace latent. Cette analyse permet d'examiner la capacité des modèles à rapprocher efficacement les évènements similaires sur la base de leur représentation vectorielle. Les jeux de données utilisés pour cette évaluation sont ARTE_INFO, GRILLE et NON_EVENT. Ils couvrent différents formats de contenus médiatiques, permettant ainsi d'évaluer la robustesse des modèles dans des contextes variés. L'évaluation repose sur la moyenne de similarité entre les vecteurs d'embeddings des transcriptions et ceux des résumés correspondants dans les notices documentaires. Plus cette similarité est élevée, plus le modèle est jugé efficace dans sa capacité à représenter l'information évènementielle. Afin de quantifier cette similarité, nous avons suivi les étapes suivantes : dans un premier temps, les embeddings des champs "Résumé" ont été générés à l'aide des modèles fine-tunés, puis les embeddings des transcriptions correspondantes ont été calculés. Ensuite, nous avons calculé la similarité moyenne entre ces deux ensembles de vecteurs afin d'évaluer leur proximité sémantique. Les résultats obtenus (Tableau 3) révèlent des écarts de performance notables entre les modèles testés. LCLM-SimCSE affiche les meilleures similarités sur les deux datasets, démontrant une capacité optimisée à aligner les représentations textuelles. SimCSE suit de près avec des performances très satisfaisantes. Bi-Encoder, bien que performant, reste légèrement en retrait par rapport aux modèles précédents. Enfin, Dangvantuan et LCLM-Bi-Encoder se situent à un niveau intermédiaire, avec des performances moyennes par rapport aux autres approches.

5.2.4 Détection d'évènements isolés : Classification binaire

Cette expérimentation vise à classifier les événements selon une approche binaire, distinguant les événements des non-événements. Les datasets utilisés pour cette évaluation sont ARTE_INFO, GRILLE et NON_EVENT. L'objectif est de tester plusieurs architectures et d'évaluer leur capacité à discriminer efficacement ces deux catégories. Trois architectures ont été évaluées : modèles sans couche de classification, modèles avec couche de classification, et modèle SetFit(Tunstall *et al.*, 2022). L'évaluation repose sur les métriques suivantes : Precision, Recall, F1-score.

modèle	accuracy	precision	recall	f1_score
Résumé				
dangvantuan	0.5237	0.7344	0.5223	0.3833
bi-encoder	0.6220	0.6545	0.6213	0.6002
lclm-bi-encoder	0.5581	0.6071	0.5571	0.4992
simcse	0.7267	0.7395	0.7270	**0.7232**
lclm-simcse	0.5014	0.2507	0.5000	0.3340
Transcription				
dangvantuan	0.5291	0.7291	0.5277	0.3954
bi-encoder	0.4695	0.4694	0.4695	0.4693
lclm-bi-encoder	0.5742	0.6130	0.5734	**0.5330**
simcse	0.5481	0.7393	0.5493	0.4348
lclm-simcse	0.5014	0.2507	0.5000	0.3340

TABLE 4 – Sans couche de classification

modèle	accuracy	precision	recall	f1_score
Résumé				
dangvantuan	0.9283	0.9287	0.9280	**0.9282**
bi-encoder	0.9178	0.9178	0.9178	0.9178
lclm-bi-encoder	0.9196	0.9203	0.9192	0.9195
simcse	0.9161	0.9161	0.9160	0.9160
lclm-simcse	0.9196	0.9195	0.9196	0.9196
Transcription				
dangvantuan	0.8986	0.8987	0.8989	0.8986
bi-encoder	0.9231	0.9234	0.9228	0.9230
lclm-bi-encoder	0.9441	0.9442	0.9439	**0.9440**
simcse	0.9266	0.9266	0.9265	0.9265
lclm-simcse	0.9318	0.9321	0.9316	0.9318

TABLE 5 – Avec couche de classification

modèle	accuracy	precision	recall	f1_score
Résumé				
dangvantuan	0.9391	0.9424	0.9391	**0.9390**
bi-encoder	0.6559	0.7965	0.6559	0.6103
lclm-bi-encoder	0.8925	0.8977	0.8925	0.8921
simcse	-	-	-	-
lclm-simcse	-	-	-	-
Transcription				
dangvantuan	0.9839	0.9839	0.9839	0.9839
bi-encoder	0.9892	0.9893	0.9892	0.9892
lclm-bi-encoder	0.9910	0.9911	0.9910	**0.9910**
simcse	-	-	-	-
lclm-simcse	-	-	-	-

TABLE 6 – Classification SETFIT

Les résultats montrent que l'architecture sans couche de classification (Tableau 4) affiche des performances relativement faibles. Toutefois, elle donne de meilleurs résultats sur les résumés, avec SimCSE se distinguant comme le meilleur modèle dans ce contexte. L'ajout d'une couche de classification (Tableau 5) permet une amélioration significative des performances pour tous les modèles testés. Les résultats indiquent que les performances sont nettement meilleures sur les transcriptions, en particulier pour les modèles fine-tunés. LCLM-Bi-Encoder et LCLM-SimCSE émergent comme les meilleurs modèles sur cette configuration. Enfin, l'architecture SetFit (Tableau 6) présente les

meilleures performances globales, bien que SimCSE n'ait pas été évalué dans ce cadre. Les résultats montrent que LCLM-Bi-Encoder est le modèle le plus performant, tandis que Dangvantuan excelle sur les résumés. Quant aux transcriptions, LCLM-Bi-Encoder conserve son avantage, confirmant ainsi sa robustesse pour la classification binaire des évènements.

5.2.5 Catégorisation d'évènements

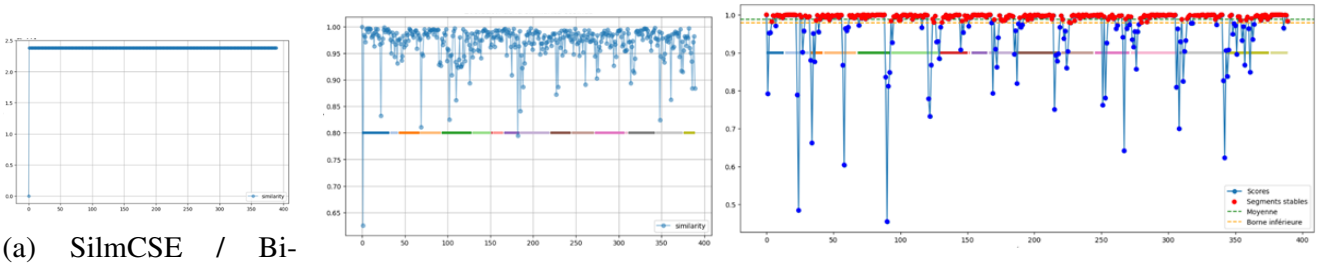
L'évaluation de la classification thématique a été réalisée à l'aide de modèles entraînés sur les jeux de données GRILLE et DBpedia. Le dataset GRILLE se base sur un référentiel composé de 1229 Noms communs de l'INA, utilisant une classification multi-label qui permet d'attribuer plusieurs catégories à un même évènement. En revanche, le dataset DBPedia annoté en mono-label avec 1253 catégories issues de Wikipedia. L'objectif de cette expérimentation est d'évaluer la capacité des modèles à catégoriser correctement les évènements selon des référentiels distincts et de comparer leurs performances sur ces ensembles de données. Pour mener cette classification thématique, le modèle Bi-Encoder thématique présenté dans la section 4.3 a été utilisé. Ce modèle a été affiné séparément sur les deux datasets GRILLE et DBpedia afin d'optimiser sa spécialisation sur leurs référentiels respectifs. Le fine-tuning a été effectué selon l'approche illustré dans la figure 1b. L'évaluation des modèles a été réalisée en s'appuyant sur des métriques standards de classification : Precision, Recall et F1-score. Les résultats obtenus montrent des performances excellentes sur le dataset DBpedia avec un F1-score de 0.97, indiquant une bonne structuration des catégories dans cette base. Concernant le dataset GRILLE, les performances sont acceptables avec un F1-score de 0.65 pour une tâche de classification multi-label impliquant 1229 classes.

5.2.6 Extraction d'évènements dans les flux d'information médiatique

Le corpus utilisé est constitué des journaux télévisés, représentant un volume total de 242 éditions. L'expérimentation menée vise à segmenter ces JT afin d'en extraire les évènements potentiels. L'approche repose sur l'utilisation des modèles LCLM-Bi-Encoder et LCLM-SimCSE. La segmentation est basée sur la détection des changements de représentation dans l'espace latent, où les variations significatives indiquent un passage à un nouvel évènement potentiel. Un seuil de rupture proche de 1 a été défini pour garantir une segmentation fine et cohérente. Ce paramétrage permet d'identifier avec précision les transitions tout en limitant la fragmentation excessive des segments. Le texte de transcription de chaque journal est d'abord découpé en phrases distinctes. Le processus débute avec les deux premières phrases, dont les représentations sont générées à l'aide du modèle utilisé. Si leur score de similarité dépasse le seuil, indiquant que leurs vecteurs de représentation appartiennent à la même région de l'espace latent, elles sont alors agrégées. Cette agrégation est ensuite comparée à la phrase suivante, et tant que le score reste supérieur au seuil, le processus d'agrégation se poursuit en intégrant progressivement les phrases suivantes. Lorsque le score de similarité descend en dessous du seuil, cela signifie qu'un changement de zone dans l'espace latent est détecté, suggérant ainsi une transition entre évènements. A ce stade un segment est créé à partir des phrases agrégées, et un nouveau processus d'agrégation démarre à partir de la phrase suivante, qui devient alors le point de départ de la nouvelle agrégation.

La figure 2a illustre l'application de ce processus en utilisant le modèle SimCSE ou Bi-Encoder. Dans ce cas, toutes les représentations générées sont équivalentes car ce modèle ne peut traiter un contexte dépassant 512 tokens. L'ajout progressif de phrases n'altère pas la position du vecteur

agrégé dans l'espace latent. Cette approche s'avère donc inadaptée aux modèles ne prenant pas en charge les contextes étendus. Une alternative à cette approche, en utilisant les modèles SimCSE ou Bi-Encoder, consisterait à appliquer une fenêtre glissante de 512 tokens pour analyser les variations des représentations dans l'espace latent. Cependant, nous avons observé un ensemble important de ruptures irrégulières (Figure 2b), rendant difficile l'exploitation des résultats.



(a) SilmCSE / Bi-Encoder (b) Fenêtre glissante (c) LCLM

FIGURE 2 – Segmentation en évènements

Enfin, nous avons appliqué notre méthode d'agrégation de contexte en utilisant les modèles de type LCLM (LCLM-Bi-Encoder et LCLM-SimCSE). Ces modèles ont permis d'obtenir des ruptures nettes (Figure 2c), facilement identifiables à l'aide d'un simple post-traitement des résultats. L'évaluation menée sur le corpus des journaux télévisés a montré de meilleures performances, avec un score Windowdiff de 0.15 pour LCLM-Bi-Encoder et 0.14 pour LCLM-SimCSE.

5.2.7 Recherche sémantique d'évènements

L'expérimentation menée vise à évaluer la capacité des modèles à rechercher, trouver et identifier un évènement à partir d'un autre en analysant leur similarité sémantique. Les évaluations ont été réalisées sur les datasets ARTE_INFO et GRILLE. Dans un premier temps, la similarité sémantique entre les notices documentaires de l'INA a été calculée selon la méthode décrite dans la section 3.2. Chaque notice documentaire est considérée comme une requête, et les réponses pertinentes sont définies par les notices présentant une similarité sémantique élevée avec cette requête. L'approche est ensuite reproduite en utilisant les embeddings afin de comparer leurs résultats respectifs. Cette évaluation se déroule selon les étapes suivantes : (1) Calcul des embeddings à partir des transcriptions des notices documentaires. (2) Recherche des résumés de notices dans ces embeddings. (3) Comparaison des K meilleurs résultats de la recherche par embeddings avec les réponses pertinentes obtenues selon les mesures de similarité décrites dans la section 3.2.

L'évaluation des performances des modèles repose sur plusieurs métriques : (1) MRR@k : évalue la capacité d'un modèle à positionner le premier élément pertinent en tête de liste. (2) NDCG@k : évalue la qualité du classement en tenant compte de la position des éléments pertinents (évaluer si les documents pertinents apparaissent en haut de la liste). (3) MAP@k : évalue la précision moyenne des résultats jusqu'à la position k.

Les modèles ont été testés sur les notices associées aux grilles d'indexation INA. Les résultats (Tableau 7) montrent que SimCSE offre les performances les plus élevées sur toutes les métriques. Il positionne correctement le premier élément pertinent dans 90% des cas parmi les 10 premiers résultats. Il atteint 74% de la qualité du classement attendu dans les 10 premiers résultats. Il atteint également 66%

de précision moyenne parmi les 100 premiers résultats. Le Bi-Encoder affiche également de bonnes performances, tandis que LCLM-Bi-Encoder présente des performances moyennes. En revanche, les modèles Dangvantuan et LCLM-SimCSE enregistrent des performances faibles. Les performances globales sur ARTE_INFO sont inférieures à celles obtenues sur GRILLE d'indexation. Toutefois, SimCSE reste le modèle le plus performant. Les modèles Dangvsantuan, Bi-encoder et LCLM affichent des performances moyennes.

Dataset	Algorithm	cos_sim-MRR@10	cos_sim-NDCG@10	cos_sim-MAP@100	dot_score-MRR@10	dot_score-NDCG@10	dot_score-MAP@100
GRILLE	dangvantuan	0.8606	0.6165	0.5004	0.8337	0.5879	0.4733
GRILLE	bi-encoder	0.8639	0.7087	0.6373	0.8050	0.6450	0.5808
GRILLE	lclm-bi-encoder	0.8342	0.6485	0.5596	0.8039	0.6135	0.5244
GRILLE	simcse	**0.8998**	**0.7353**	**0.6594**	**0.8998**	**0.7352**	**0.6594**
GRILLE	lclm-simcse	0.7938	0.6116	0.5310	0.7938	0.6116	0.5310
ARTE_INFO	dangvantuan	0.7558	0.5409	0.4492	0.7054	0.5027	0.4157
ARTE_INFO	bi-encoder	0.6801	0.5340	0.4555	0.6286	0.4876	0.4140
ARTE_INFO	lclm-bi-encoder	0.6278	0.4771	0.3999	0.6028	0.4620	0.3841
ARTE_INFO	simcse	**0.7715**	**0.6060**	**0.5247**	**0.7715**	**0.6060**	**0.5247**
ARTE_INFO	lclm-simcse	0.6561	0.5103	0.4350	0.6561	0.5103	0.4350

TABLE 7 – Évaluation de la recherche sémantique d'évènements

6 Conclusion

Cet article met en lumière le potentiel des modèles contrastifs tels que SimCSE et Sentence-BERT ainsi que des architectures permettant l'extension du contexte, pour l'analyse automatique des évènements médiatiques. Ces modèles, affinés sur des corpus riches et diversifiés (AFP, DBpedia et INA), améliorent la représentation vectorielle des récits médiatiques pour une analyser approfondie des évènements. Les évaluations réalisées indiquent que ces modèles ont des performances globales relativement équivalentes, bien que certaines spécialisations par tâche soient observées. Par exemple, le Bi-Encoder thématique excelle dans la catégorisation des évènements, tandis que LLCM se distingue par ses performances en segmentation.

Dans une perspective d'amélioration, plusieurs axes de recherche méritent d'être approfondis. En premier lieu, une évaluation sur des benchmarks de référence permettrait une comparaison avec d'autres méthodes sur différentes tâches, notamment pour des tâches complexes comme la segmentation en évènements. Toutefois, les ressources d'évaluation en Français pour ce type de tâche restent limitées, ce qui constitue un obstacle à l'évaluation de nos modèles. L'intégration et l'évaluation de nouveaux modèles dédiés aux longues séquences comme LLM2Vec(BehnamGhader *et al.*, 2024) permettrait de comparer les modèles proposés avec les grands modèles de langue, notamment en terme de gestion du contexte étendu. Par ailleurs, l'ajout d'une couche de cross-attention aux sein des modèles LCLM pourrait améliorer leur performances sur les tâches de clustering et de recherche sémantique. Enfin, l'utilisation d'algorithme de détection d'anomalies pour repérer et identifier des phénomènes médiatiques atypiques comme les tempêtes médiatiques, ainsi que l'utilisation des LLMs pour générer des résumés automatiques d'évènements et la construction d'une base de connaissance spécialisée, constituent des perspectives prometteuses. Ces perspectives ouvrent la voie à une amélioration continue des modèles, en affinant leur spécialisation et en renforçant leur robustesse pour mieux capturer la complexité du traitement médiatique des évènements.

Remerciements

Cette recherche a été financée en tout ou partie, par l'Agence Nationale de la Recherche (ANR) au titre du projet "ANR-23-IAS1-0001".

Références

ALLAN J. (2002). *Introduction to Topic Detection and Tracking*, In *Topic Detection and Tracking : Event-based Information Organization*, p. 1–16. Springer US : Boston, MA. DOI : 10.1007/978-1-4615-0933-2_1.

ARORA S., LIANG Y. & MA T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*.

BEHNAMGHADER P., ADLAKHA V., MOSBACH M., BAHDANAU D., CHAPADOS N. & REDDY S. (2024). Llm2vec : Large language models are secretly powerful text encoders. *ArXiv*, **abs/2404.05961**.

BELTAGY I., PETERS M. E. & COHAN A. (2020). Longformer : The long-document transformer. *ArXiv*, **abs/2004.05150**.

BERNARD G. (2022). *Détection et suivi d'événements dans des documents de presse historique*. Theses, Université de La Rochelle (ULR), La Rochelle. HAL : tel-04115986.

BLONDEL V. D., GUILLAUME J.-L., LAMBIOTTE R. & LEFEBVRE E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment*, **2008**(10), P10008.

DEVLIN J., CHANG M., LEE K. & TOUTANOVA K. (2019). BERT : pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, p. 4171–4186 : Association for Computational Linguistics.

ESTER M., KRIEGEL H.-P., SANDER J. & XU X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery and Data Mining*.

GAO T., YAO X. & CHEN D. (2021). Simcse : Simple contrastive learning of sentence embeddings. *ArXiv*, **abs/2104.08821**.

GROOTENDORST M. (2020). Keybert : Minimal keyword extraction with bert. DOI : 10.5281/zenodo.4461265.

HERVÉ N., VIAUD M.-L., THIÈVRE J., SAULNIER A., CHAMP J., LETESSIER P., BUISSON O. & JOLY A. (2013). Otmedia : the french transmedia news observatory. *Proceedings of the 21st ACM international conference on Multimedia*.

ISHLACH K., BEN-DAVID I., FIRE M. & ROKACH L. (2024). A novel method for news article event-based embedding. *ArXiv*, **abs/2405.13071**.

LEETARU K. & SCHRODT P. A. (2013). Gdelt : Global data on events, location, and tone. *ISA Annual Convention*.

LEVI E., MOR G., SHEAFER T. & SHENHAV S. (2022). Detecting narrative elements in informational text. In *Findings of the Association for Computational Linguistics : NAACL 2022*, p. 1755–1765, Seattle, United States : Association for Computational Linguistics.

MACQUEEN J. B. (1967). Some methods for classification and analysis of multivariate observations. In L. M. L. CAM & J. NEYMAN, Éds., *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, p. 281–297 : University of California Press.

MARKUS D. K., LEVI E., SHEAFER T. & SHENHAV S. R. (2024). Reap the wild wind : Detecting media storms in large-scale news corpora. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2024*, p. 4786–4797, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-emnlp.275.

MAZOYER B. (2020). *Social Media Stories. Event detection in heterogeneous streams of documents applied to the study of information spreading across social and news media*. Theses, Université Paris-Saclay. HAL : tel-02987720.

MCINNES L., HEALY J. & ASTELS S. (2017). hdbscan : Hierarchical density based clustering. *J. Open Source Softw.*, **2**, 205.

MIYAMOTO S. (2012). An overview of hierarchical and non-hierarchical algorithms of clustering for semi-supervised classification. In V. TORRA, Y. NARUKAWA, B. LÓPEZ & M. VILLARET, Éds., *MDAI*, volume 7647 de *Lecture Notes in Computer Science*, p. 1–10 : Springer.

MOUTIDIS I. & WILLIAMS H. T. P. (2020). Complex networks for event detection in heterogeneous high volume news streams. *ArXiv*, **abs/2005.13751**.

NGUYEN D. Q., BILLINGSLEY R., DU L. & JOHNSON M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, **3**, 299–313.

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT : Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3982–3992, Hong Kong, China : Association for Computational Linguistics.

ROBERTSON S. E., WALKER S., JONES S., HANCOCK-BEAULIEU M. & GATFORD M. (1994). Okapi at trec-3. In D. K. HARMAN, Éd., *TREC*, volume 500-225 de *NIST Special Publication*, p. 109–126 : National Institute of Standards and Technology (NIST).

TAREKEGN A. N. (2024). Large language model enhanced clustering for news event detection. *ArXiv*, **abs/2406.10552**.

TAYLOR S. J. & LETHAM B. (2018). Forecasting at scale. *The American Statistician*, **72**, 37 – 45.

TRAAG V. A., ALDECOA R. & DELVENNE J.-C. (2015). Detecting communities using asymptotical surprise. *Physical review. E, Statistical, nonlinear, and soft matter physics*, **92 2**, 022816.

TUNSTALL L., REIMERS N., JO U. E. S., BATES L., KORAT D., WASSERBLAT M. & PEREG O. (2022). Efficient few-shot learning without prompts. *ArXiv*, **abs/2209.11055**.

ZAHEER M., GURUGANESH G., DUBEY K. A., AINSLIE J., ALBERTI C., ONTAÑÓN S., PHAM P., RAVULA A., WANG Q., YANG L. & AHMED A. (2020). Big bird : Transformers for longer sequences. *ArXiv*, **abs/2007.14062**.