Détection de nouveauté au plus tôt dans des flux de données textuelles

Clément Christophe

N° d'ordre NNT : 2021LYSE2026

THESE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 512

Informatique et Mathématiques

Discipline : Informatique

Soutenue publiquement le 15 mars 2021, par :

Clément CHRISTOPHE

Détection de nouveauté au plus tôt dans des flux de données textuelles.

Devant le jury composé de :

Céline HUDELOT, Professeure des universités, Centrale Supélec, Présidente

Pascale KUNTZ-COSPEREC, Professeure des universités, Université de Nantes, Rapporteure

Josiane MOTHE, Professeure des universités, Université Toulouse 2 Jean Jaurès, Rapporteure

Alexandre ALLAUZEN, Professeur des universités, ESPCI Paris, Examinateur

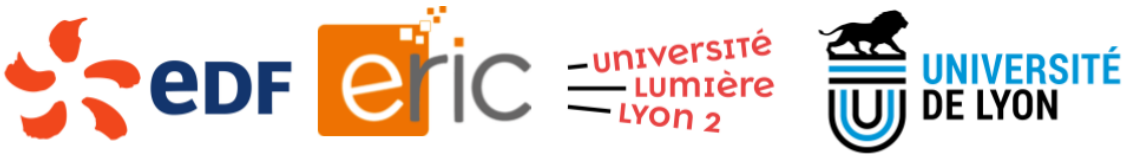
Philippe SUIGNARD, Encadrant entreprise, EDF LAB Paris Saclay, Examinateur

Manel BOUMGHAR, Encadrante entreprise, EDF LAB Paris Saclay, Examinatrice

Julien VELCIN, Professeur des universités, Université Lumière Lyon 2, Directeur de thèse

Jairo CUGLIARI, Maître de conférences, Université Lumière Lyon 2, co-Directeur de thèse

Contrat de diffusion



Thèse présentée pour obtenir le grade de Docteur de l'Université Lumière Lyon 2

École Doctorale Informatique et Mathématiques (ED 512)

Laboratoire ERIC (EA 3083)

Discipline: Informatique

Détection de nouveauté au plus tôt dans des flux de données textuelles.

Clément CHRISTOPHE

Présentée et soutenue publiquement le XXX, devant un jury composé de :

Josiane MOTHE, Professeur des Universités, Université de Toulouse 2	Rapportrice
Pascale KUNTZ, Professeur des Universités, Université de Nantes	Rapportrice
Alexandre ALLAUZENE, Professeur des Universités, ESPCI	Examinateur
Céline HUDELOT, Professeur des Universités, CentraleSupélec	Examinatrice
Jairo CUGLIARI, Maître de Conférences, Université Lyon 2	Co-directeur
Julien VELCIN, Professeur des Universités, Université Lyon 2	Directeur

UNIVERSITE LUMIERE LYON 2

Abstract

Early novelty detection in textual streaming data

by Clément CHRISTOPHE

The work presented in this thesis, made in collaboration with *Électricité de France* (EDF), aims to develop novelty detection models in textual data streams. For EDF, this is part of an approach to anticipate customer needs.

We present different novelty detection approaches that exist in the literature, which allows us to precisely define the tasks we want to solve. These definitions allow us to set up evaluation methods, based either on simulated data or on real data. Modifying real data allows us to simulate novelty arrival scenarios and therefore to measure the performance of existing methods.

We present two models of detection for new elements by first using topic probabilistic models. The second approach is CEND, an algorithm based on the movements of words in high dimensional representation spaces. This type of model allows us to distinguish words linked with abrupt events or slowly emerging themes.

We present a model for monitoring the dynamics of a classification plan. By linking methods of time series forecasting and sequential analysis, we estimate when the dynamic of a signal changes. We test these methods on public press data and on an EDF industrial dataset.

Keywords : novelty detection ; topic models ; temporal word embeddings ; forecasting

UNIVERSITE LUMIERE LYON 2

Résumé

Détection de nouveauté au plus tôt dans des flux de données textuelles

par Clément CHRISTOPHE

Les travaux présentés dans cette thèse, réalisés en partenariat avec l'entreprise Électricité de France (EDF), ont pour objectif de développer des modèles de détection de nouveauté dans des flux de données textuelles. Pour EDF, cela s'inscrit dans une démarche d'anticipation des besoins clients.

Nous présentons les différentes approches de détection de nouveauté existantes dans la littérature, ce qui nous permet de définir précisément les tâches que nous voulons résoudre. Ces définitions nous permettent de mettre en place des méthodes d'évaluations, basées soit sur des données simulées, soit sur des données réelles. La modification des données réelles nous permet de simuler des scénarios d'arrivées de la nouveauté et donc de mesurer l'efficacité des méthodes existantes.

Nous présentons deux modèles de détections d'éléments nouveaux en utilisant tout d'abord les modèles thématiques probabilistes. Le deuxième modèle est CEND, un algorithme se basant sur les mouvements des mots dans des espaces de représentations en grandes dimensions. Ce type de modèle nous permet de faire la différence entre des mots liés à des évènements abrupts et des thématiques émergents doucement.

Nous présentons un modèle de surveillance des dynamiques des plans de classements. En liant des méthodes de prévision de série temporelle et d'analyse séquentielle, nous arrivons à estimer quand est ce qu'un signal temporel change de dynamique. Nous testons ces méthodes sur des données d'articles de presse et sur des données industrielles d'EDF.

Mots-clés : détection de nouveauté ; modèles thématique ; modèles de plongements temporels ; prévision de séries temporelles

Remerciements

Je remercie tout d'abord mes encadrants, Julien Velcin et Jairo Cugliari pour m'avoir accompagné durant cette thèse. Vos conseils, votre patience et votre soutien m'ont permis de réaliser ce travail dans les meilleures conditions. Je n'oublierai jamais toutes ces discussions très riches où nous avions toujours de nouvelles idées à développer.

Je tiens à remercier les membres du jury, Pascale Kuntz et Josiane Mothe en tant que rapporteur, ainsi que Céline Hudelot et Alexandre Allauzène pour avoir accepté d'évaluer ce travail.

Je tiens à remercier Manel Boumghar et Philippe Suignard qui m'ont permis d'effectuer cette thèse CIFRE au sein d'EDF. Je vous remercie pour votre encadrement sans faille et votre accompagnement au jour le jour. Merci de m'avoir soutenu.

Je tiens à témoigner toute ma reconnaissance aux personnes suivantes, pour leur aide et leur soutien dans la réalisation de ce travail :

Aux membres du laboratoire ERIC que j'ai pu côtoyer lors de mes séjours à Lyon, tout particulièrement Antoine, Margot, Robin, Adrien et Jean. Merci pour la bonne ambiance de travail et toutes ces discussions au laboratoire et en dehors.

À mes responsables, présents et anciens, au sein d'EDF, notamment Maud Imberty, Marie Kennis, Meryl Bothua et Delphine Lagarde, sans qui ce projet de collaboration n'aurait pu avoir lieu.

À l'ensemble des membres de l'équipe SOAD d'EDF. Merci pour votre bienveillance, votre ouverture d'esprit et votre professionnalisme. J'ai énormément appris grâce à chacun d'entre vous. J'ai une pensée particulière pour mes co-bureaux, Alexandra et Guillaume. Merci pour cette excellente ambiance de travail.

À mes amis les plus proches : Matthieu, Victor, Amine, François, Walid, Sofiane et Alexis. Merci pour votre soutien, votre curiosité, votre ouverture d'esprit et pour tous ces débats sans fin.

À mes frères et soeurs : Emmanuel, Sébastien, Sophie et Solenne. Merci de m'avoir transmis votre curiosité, votre culture, votre recherche de l'excellence. Votre exemple n'a jamais cessé de me guider.

À mes parents, j'espère que ce travail vous rendra fiers, il est d'abord pour vous. Merci d'avoir fait de moi la personne que je suis aujourd'hui.

À Inas. Merci pour ton soutien sans faille durant ces trois ans. Ce travail, c'est aussi le tien, je n'y serai jamais arrivé sans toi.

Table des matières

Table des figures

Liste des tableaux

Abréviations

ARIMA	**A**uto**R**egressive **I**ntegrated **M**oving **A**verage
AUC	**A**rea **U**nder **C**urve
CUSUM	**Cu**mulative **Sum** control chart
DTM	**D**ynamic **T**opic **M**odel
EDF	**É**lectricité **D**e **F**rance
KL-Div	**K**ullback-**L**eibler **Div**ergence
KNN	**K**-**N**earest **Neighbors**
LDA	**L**atent **D**irichlet **A**llocation
NMF	**N**on Negative **M**atrix **F**actorization
NYTAC	**N**ew **Y**ork **T**imes **A**nnotated **C**orpus
OLDA	**O**n-line **L**atent **D**irichlet **A**llocation
PPMI	**P**ositive **P**ointwise **M**utual **I**nformation
RMSE	**R**oot **M**ean **S**quare **E**rror
ROC	**R**eceiver **O**perating **C**haracteristic
SCI	Jeu de données de résumés d'articles **SCI**entifiques
SGNS	**S**kip-**G**ram with **N**egative **S**ampling
SVD	**S**ingular **V**alue **D**ecomposition
TAL	**T**raitement **A**utomatique des **L**angues
TDT	**T**opic **D**etection and **T**racking
TFIDF	**T**erm **F**requency-**I**nverse **D**ocument **F**requency

Symboles

$\mathcal{D}$	Ensemble de documents
d_i	i-ème document de l'ensemble
$\mathcal{V}$	Vocabulaire, ensemble de mots
w_j	j-ème mot du vocabulaire
n_D	taille du corpus en documents
n_V	taille du vocabulaire en mots
l	taille d'un document en mots
t	date
f	fréquence
k	indice de la thématique LDA
ϕ_k	distribution de probabilités sur le vocabulaire
θ^d	distribution de probabilités sur le document
e	distance euclidienne
v^w	vecteur de représentation du mot w
ρ	corrélation de Spearman
cov	covariance
σ	écart type
K	seuil
ω	vitesse
a	accélération
l	retard

p	taille du retard
$k(t)$	signal de fréquence des thématiques
$c(t)$	signal de fréquence de co-occurrence des thématiques
y_i	valeur réelle du signal
$\hat{y}_i$	valeur prédite du signal

. . .

Chapitre 1

Introduction

Dans ce chapitre d'introduction, nous présentons tout d'abord le contexte dans lequel s'inscrivent les travaux réalisés dans le cadre de cette thèse de doctorat. Nous présentons quelques notions générales utiles à la compréhension des travaux. Nous présentons en détail les différents jeux de données utilisés pour réaliser les expérimentations. Enfin, nous résumons les différentes contributions qui ont été faites durant cette thèse et présentons les différentes parties de ce manuscrit.

1.1 Contexte

Le contexte dans lequel s'effectue ce travail est particulier. Nous présentons ici le contexte général autour du sujet de thèse, le cadre scientifique dans lequel il s'inscrit et les applications industrielles qui peuvent en découler.

1.1.1 Contexte général

La thèse présentée ici s'intitule "**Détection de nouveauté au plus tôt dans des flux de données textuelles**". Il est courant, dans la société, de lire des

textes : des livres, des articles de presses ou de blogs, des courriels, des posts sur des réseaux sociaux, des messages, etc.. La masse de données textuelles traitées chaque jour est donc très importante aussi bien pour un humain que pour des systèmes informatiques. Lorsque ces données sont traitées, il est courant de se demander si ce qu'elles contiennent est attendu, c'est-à-dire qui correspond à ce que l'on a pu voir auparavant. Que ce soit de nouvelles informations, des évènements importants, de nouveaux mots, de nouvelles expressions, tout ceci est important de manière à comprendre ce qu'il se passe dans le flux de données.

Pour des entreprises, il est primordial d'analyser l'ensemble des données textuelles produites soit dans l'entreprise (des courriels, des comptes-rendus, des formulaires d'interventions) soit à l'extérieur (des courriels envoyés au service client, des posts sur les réseaux sociaux, des articles de presse). La détection de la nouveauté, que nous abordons dans ce manuscrit, a pour objectif d'analyser ces données afin de comprendre, à partir de peu d'informations, des phénomènes émergents, c'est-à-dire qui prennent de l'ampleur. Au final, cela a pour but d'anticiper des réponses de la part de l'entreprise pour mieux traiter ces nouvelles informations.

1.1.2 Contexte scientifique

L'analyse de grands volumes de données textuelles, en particulier de données en flux, relève des domaines du Traitement Automatique des Langues et de la Fouille de Textes. La problématique générale que pose l'entreprise se place dans un contexte de TDT (*Topic Detection and Tracking* [1]) et s'articule en deux problèmes complémentaires, à savoir a) le suivi de thématiques dans le temps et b) la détection de signaux faibles.

Le suivi de thématiques dans le temps est un problème étudié dans le domaine du traitement automatique des langues. Plusieurs travaux, depuis l'apparition des premiers modèles de *clustering* appliqués aux données textuelles, se sont concentrés sur leur adaptation aux évolutions temporelles. C'est le cas pour les algorithmes

classiques de clustering de type k-means [3] et pour les algorithmes probabilistes de type *Latent Dirichlet Allocation* (LDA) [4]. Des travaux comme ceux de [5], [6], [7] ont permis d'étendre le concept de thématique probabiliste sur une dynamique temporelle. Ceux-ci forment un nombre fixe de thématiques dont le contenu évolue dans le temps. L'inconvénient est qu'elles ne permettent pas de faire évoluer le nombre de thématiques. Certaines méthodes [8] permettent de faire évoluer ce nombre dans le temps, mais il faut généralement attendre d'avoir une quantité importante de données ce qui, nous le verrons, peut constituer un frein pour l'entreprise.

Le problème b) a été attaqué de manière très générale avec la détection de signaux faibles, [9] et quelques travaux se concentrent sur les données textuelles [1, 10]. Ce domaine est souvent confondu avec les domaines de la détection d'anomalies, d'évènements ou d'évolution du langage. Pour les deux premiers, il est nécessaire d'observer soit des valeurs extrêmes uniques (anomalies) soit de très grosses quantités sur une temporalité très courte (évènements). Pour le troisième, il est nécessaire d'attendre d'avoir de grosses quantités de données sur des temporalités très longues pour observer des changements importants. Enfin, dans tous ces domaines, l'évaluation est complexe et, nous le verrons, est souvent faite a posteriori de manière qualitative. Dans ce travail, les verrous scientifiques majeurs consistent à détecter de la nouveauté au plus tôt, c'est-à-dire sans attendre que le volume de documents soit nécessairement suffisant pour capturer les nouvelles thématiques et de développer des méthodes d'évaluations quantitatives pour se comparer aux méthodes existantes.

1.1.3 Applications pour EDF

Électricité de France (EDF) est une entreprise française et est le premier producteur et fournisseur d'électricité en Europe. Fort de ses plus de 25 millions de clients particuliers en France, il est nécessaire d'analyser les avis et les retours de ces clients. Pour cela, l'entité Commerce du groupe surveille l'évolution des thématiques

Catégorie	Exemple de courriels
Technique	"J'avais été prévenu qu'un employé Enedis devait passé pour relevé mon compteur le, date mais cleui-ci ne s'est jamais manifesté. Par conséquent, j'aimerai savoir sur quelle base vous allez établir ma facture 2018."
Contrat	"Bonjour nous souhaiterions changer notre adresse de livraison pour notre contrat qui n'est plus au adresse suite à une division de parcelle. Cordialement"
Montant	"Bonjour, mon échéancier a augmenté en septembre pourriez vous me dire pourquoi ? Cordialement"
Solidarité	"Bonjour, je reviens vers vous suite à la demande d'aide pour l'énergie. J'aurais voulu savoir à quel montant j'aurais le droit si le dossier est traité. Merci. Cordialement"

TABLE 1.1: Exemples de classifications de courriels EDF.

discutées dans différents types de corpus textuels : des tweets, des réclamations, des courriels, des articles de blogs, etc. Un plan de classement prédéfini par des experts métiers permet de recourir à des algorithmes de classification supervisée performants et ainsi trier les différents documents textuels dans des catégories liés aux métiers ou aux types de retours. Nous illustrons, dans le tableau 1.1, quelques exemples de catégorie ainsi que des courriels clients qui leur sont associés.

Cette classification permet d'améliorer les réponses métiers au sein d'EDF et ainsi avoir une vision plus globale des retours clients afin d'améliorer leur satisfaction. Bien que ces algorithmes de classification soient performants, un certain nombre de documents se retrouvent mal ou même non classés. Cela peut être dû au fait que les catégories changent : elles évoluent au cours du temps (par exemple avec l'apparition de nouveaux termes), ou au fait que de nouvelles catégories thématiques apparaissent. Ce phénomène est rendu d'autant plus complexe que ces thématiques peuvent avoir des dynamiques spécifiques et irrégulières (cycliques par exemple) qui sont difficilement perceptibles si on ne prend pas en compte un historique ou des connaissances métiers préalables.

Dans ce contexte, l'analyse des dynamiques de ces thématiques et la détection de la nouveauté à partir de flux de documents est un sujet primordial pour EDF. L'enjeu

pour la Direction Commerce d'une détection de thématiques émergente au plus tôt est d'augmenter la performance opérationnelle en comprenant plus rapidement de nouvelles attentes que les clients peuvent avoir. La Direction Commerce peut ensuite implémenter de nouveaux modèles et de nouvelles réponses commerciales afin de répondre au mieux aux demandes des clients.

De manière plus générale, l'entreprise souhaite être en mesure de suivre les thématiques des textes dans le temps, que celles-ci soient récurrentes ou qu'elles apparaissent et disparaissent au fur et à mesure du temps, et de mettre à jour les plans de classement qu'elle utilise quotidiennement pour ses analyses. Pour être en mesure d'identifier ces nouvelles thématiques au plus tôt, il est nécessaire d'effectuer des travaux de recherche et de mettre en place des modèles permettant cette détection. Bien que ces modèles soient spécifiquement pensés pour les besoins de la Direction Commerce, ils pourront trouver leur application dans l'ensemble des métiers du groupe EDF et dans l'ensemble des entreprises analysant les retours textuels de leurs clients.

1.2 Notions générales

Les travaux présentés dans ce manuscrit s'inscrivent dans le domaine du Traitement Automatique des Langues (TAL). Nous nous concentrons ici sur l'évolution temporelle de concepts classiques de ce domaine : les mots, les documents et les thématiques. Nous présentons donc ici quelques notions générales nécessaires pour aborder ce manuscrit. Nous abordons deux domaines : le TAL et le traitement des séries temporelles.

1.2.1 Traitement automatique des langues

Le Traitement Automatique des Langues est un domaine qui rassemble l'ensemble des techniques qui permettent de rendre les contenus textuels exploitables par la

machine. C'est un domaine multidisciplinaire qui traite aussi bien de la recherche d'information, de l'analyse syntaxique, de la génération automatique, de la traduction ou de la catégorisation de documents.

Un document $d_i \in \mathcal{D}$ est composé de mots w tirés d'un vocabulaire $\mathcal{V}$, où $\mathcal{D}$ est un ensemble de taille n_D et $\mathcal{V}$ un ensemble de taille n_V. Le vocabulaire $\mathcal{V}$ est construit lors d'une étape de tokenisation qui consiste à découper les séquences de mots en jetons (ou *tokens*). Le vocabulaire $\mathcal{V}$ est composé de n_V jetons et chaque document de taille L est représenté comme une suite de jetons $(w_1, ...w_L)$ avec $w_i \in [1, n_V]$. Un certain nombre de prétraitements classiques comme la lemmatisation ou la racinisation permettent de sélectionner des sous-ensembles de ces jetons et donc de raccourcir la taille du vocabulaire V.

La racinisation [11] est un procédé qui consiste à réduire un mot à sa racine. Il permet d'enlever les préfixes, suffixes, pluriels ainsi que les formes dérivées du mot en question. Le but est d'obtenir une forme tronquée d'un mot, commune à toutes les variantes morphologiques. Par exemple, en français, les mots "cheval, chevaux, chevalier, chevalerie" partageront la racine "cheval" qui pourra être enregistrée dans le vocabulaire.

La lemmatisation est un procédé qui consiste à déterminer le lemme d'un mot, c'est-à-dire à revenir à sa forme canonique. Ce processus nécessite un dictionnaire propre à la langue utilisée, car chaque mot sera traité par rapport à son contexte. Par exemple, en français, cela est particulièrement utile pour rassembler les différentes conjugaisons. Les mots "est, sois, fut, étais, fussions" partageront le même lemme "être".

Une fois les prétraitements effectués et la liste de vocabulaire obtenue, l'ensemble des documents peuvent être représentés sous la forme d'une séquence de jetons et ainsi conserver l'ordre des mots dans le document. Une autre manière consiste à simplifier cette représentation sous forme de sac de mots. C'est-à-dire que l'ordre des mots n'est plus conservé et un document d_i peut ainsi être représenté sous

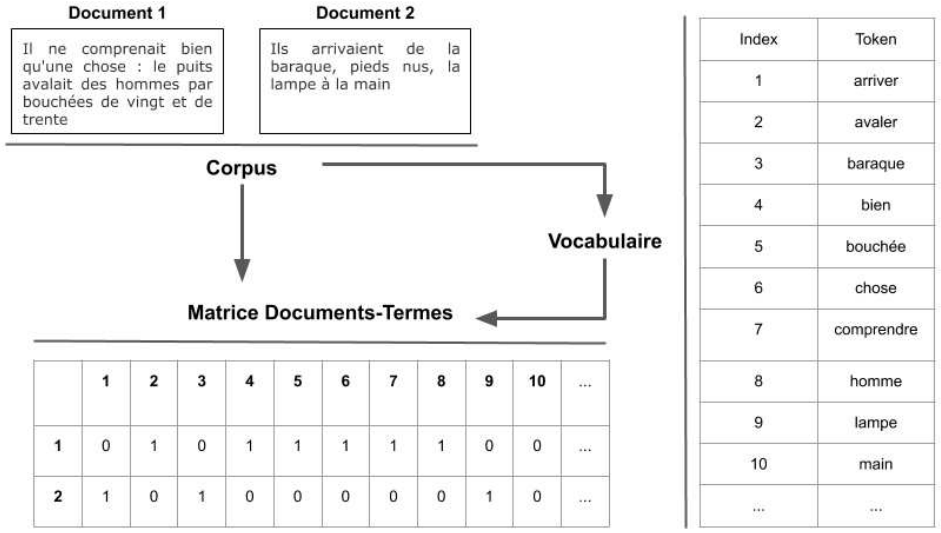


FIGURE 1.1: Construction d'une matrice documents-termes. À partir des documents, on construit la liste du vocabulaire contenant les jetons puis la matrice en question.

la forme d'un vecteur v_d de taille n_V où chaque entrée à l'index j correspond au nombre d'occurrences du mot w_j dans le document d_i. La représentation la plus simple consiste à représenter l'ensemble du corpus $\mathcal{D}$ sous la forme d'une matrice X de taille $n_D * n_V$ ou l'entrée x_{ij} est le nombre d'apparitions du jeton w_j dans le document d_i. D'autres méthodes que nous étudierons plus tard se basent sur d'autres informations que le nombre d'apparitions du jeton. Les vecteurs de représentations peuvent être normalisés, en divisant les lignes de X par le nombre de mots dans le document d_i. Cela permet d'obtenir des fréquences relatives qui seront désignées par l'acronyme TF (*Term Frequency*) dans ce manuscrit. Cette normalisation est à la base d'une première méthode de pondération, fréquemment utilisée dans le domaine du TAL. Nous allons maintenant présenter 3 manières de représenter les mots, les documents et les thématiques. Ces algorithmes seront utilisés fréquemment dans le reste du manuscrit.

TFxIDF [12] (***Term Frequency - Inverse Document Frequency***) : c'est une mesure permettant de représenter l'importance d'un mot par rapport à un document

dans un corpus. C'est un produit de deux statistiques : la fréquence du terme (TF) et l'inverse de la fréquence dans le document (IDF).

La fréquence du terme est normalisée, on la définit pour un mot w_j dans un document d_i comme :

$$TF(w_j, d_i) = \frac{f_{w_j,d_i}}{\sum_{w' \in d} f_{w',d}} \tag{1.1}$$

L'inverse de la fréquence dans le document est une mesure de l'information apportée par un mot dans un corpus, elle est définie par :

$$IDF(w_j, \mathcal{D}) = log \frac{n_D}{|d \in \mathcal{D} : w_j \in d|} \tag{1.2}$$

où n_D est la taille du corpus et $|d \in \mathcal{D} : w_j \in d|$ représente le nombre de documents ou le mot w_j apparaît.

$$TFxIDF(w, d, \mathcal{D}) = TF(w, d) * IDF(w, \mathcal{D}) \tag{1.3}$$

Un poids TFxIDF important est atteint avec une fréquence de terme élevé (dans un document donné) et une fréquence de terme faible dans le reste du corpus. La métrique du TFxIDF permet de filtrer plus simplement les mots outils ("avec, mais, car, être...") ainsi que les mots très fréquemment employés qui apparaissent dans beaucoup de documents et qui apportent donc peu d'informations.

Représentation vectorielle des mots (*Embeddings*) : c'est un ensemble de techniques qui permettent de représenter les mots sous forme de vecteurs dans un espace multidimensionnel. L'idée générale consiste à rapprocher dans l'espace les mots sémantiquement proches. Ce rapprochement est effectué en prenant en compte le contexte autour des mots : soit la co-occurrence dans l'ensemble du document pour les architectures en sac de mots soit le contexte proche pour les approches

séquentielles. Plusieurs approches permettent de former ces vecteurs de mots. Dans ce manuscrit, nous nous concentrerons sur 2 types de méthodes que nous détaillerons dans un prochain chapitre : les approches basées sur la factorisation de matrice et les approches basées sur les réseaux de neurones. À partir des co-occurrences de mots dans des fenêtres de contexte (c'est-à-dire autour d'un mot en question), l'algorithme apprend à positionner les mots de façon à ce que ceux dont le sens est similaire soient proches dans l'espace de représentation.

PPMI (*Positive Pointwise Mutual Information*) : c'est une mesure de pondération des mots qui se base sur leur co-occurrence dans les documents d'un corpus. Elle se base sur la probabilité d'apparition d'un mot x ou y et permet de construire une matrice de taille n_V sur n_V avec n_V étant la taille du vocabulaire entier. Elle se définit comme :

$$\text{PPMI}(x, y) = \max\left\{\log \frac{p(x,y)}{p(x)p(y)}, 0\right\}.$$

avec $p(x)$ étant la probabilité d'apparition du mot x et $p(x, y)$ la probabilité que les mots x et y apparaissent dans le même document. Elle est dérivée de la mesure de PMI (*Pointwise Mutual Information*) sur laquelle on ajoute une contrainte de positivité. En effet si $PMI(x, y) < 0$, nous avons $\frac{p(x,y)}{p(x)p(y)} < 1$ et donc $p(x, y) < p(x)p(y)$, ce qui voudrait dire que x et y ont tendance à plus apparaître individuellement et donc que leur co-occurrence apporte peu d'informations.

La mesure de PPMI permet de représenter les mots selon des vecteurs de dimension, n_V mais permet aussi d'identifier facilement les mots les plus fréquents et les n-grammes intéressants. Quand cette valeur est proche de 0, les deux mots ne forment pas un concept unique : ils co-occurrent par chance. Quand un des mots à une probabilité d'occurrence faible, mais une probabilité jointe d'occurrence avec l'autre mot forte, cela signifie que les deux mots représentent un concept unique.

Plusieurs extensions de la PPMI ont été développées. Certaines seront utilisées dans la suite de ce manuscrit :

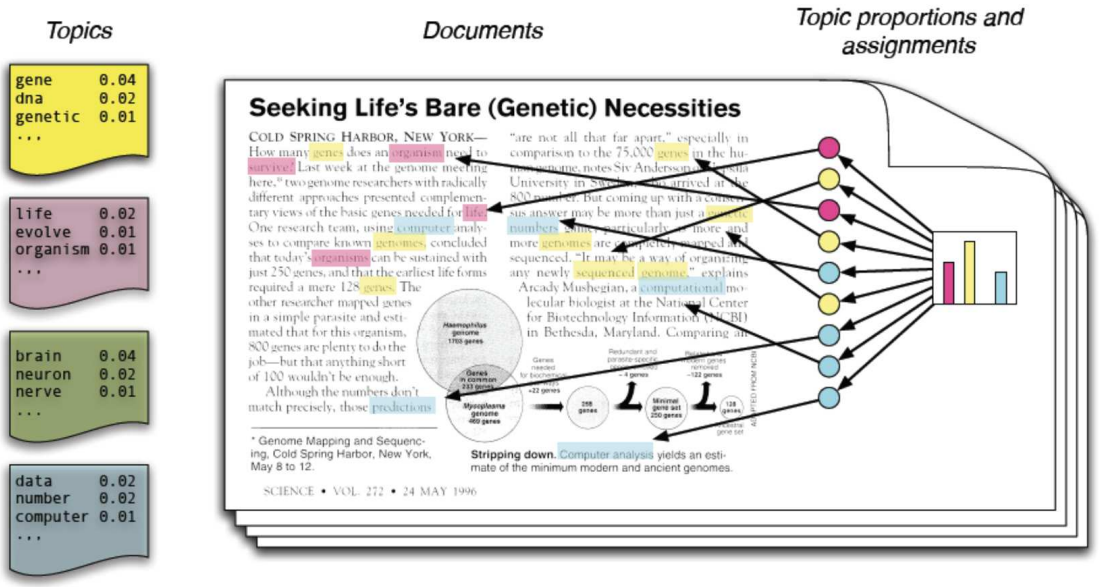


FIGURE 1.2: Illustration d'un modèle thématique type LDA. Les 4 thématiques sont décrites par rapport aux mots les plus probables à gauche.

— Shifted-PPMI [13]. $SPPMI(x, y) = max\{PMI(x, y) - log(s), 0\}$

— PPMI^k [14]. $PPMI^k(x, y) = max\{log(\frac{p(x,y)^k}{p(x)p(y)}), 0\}$

Représentation à base de thématiques : c'est un ensemble de technique qui permet de construire des thématiques de manière non supervisée. L'idée générale consiste, à partir de l'étude des co-occurrences entre les mots dans les documents, à générer des distributions de probabilités sur le vocabulaire qui font office de thématiques. L'algorithme le plus classique est *Latent Dirichlet Allocation* [4] mais d'autres existent dans la littérature telle que pLSA [15] ainsi que les approches basées sur de la factorisation de matrice non négative (NMF) [16]. Bien que les thématiques n'aient pas de titre attribué comme peuvent être des catégories manuellement construites (comme les catégories prédéfinies de "Sport", "Économie" ou "Politique"...), elles sont décrites par les mots les plus probables qui leur sont associés. Par exemple, sur la figure 1.2, nous retrouvons quatre thématiques dont les mots les plus probables nous donnent des indications sur leur contenu. Nous décrirons en détail dans un prochain chapitre comment fonctionnent ces algorithmes et comment nous nous en servons pour détecter de la nouveauté.

1.2.2 Séries temporelles

Dans ce manuscrit, nous nous intéressons à des signaux évoluant dans le temps. En effet, lorsque nous pensons à de la nouveauté, nous nous basons forcément sur une notion de passé et peut-être, de futur. Notre sujet de recherche implique donc un certain nombre d'aspects temporel dont il faut expliquer les bases.

Une série temporelle est une suite d'observation, donc de points de données, qui se suit dans le temps. C'est un type de données spéciales car les données y sont **dépendantes** les unes des autres, c'est-à-dire que leur valeur à un instant t dépend des observations précédentes à $t-1, t-2, ..., t-n$. C'est un type de données centrales dans le domaine scientifique et qui est aussi bien utilisé en économie, météorologie, traitement du signal ou statistique. Ces données temporelles sont donc ordonnées et nécessitent donc des traitements spécifiques avant d'être exploitables. Généralement, ce type de données est utilisé dans des tâches de *clustering*, de classification, de détection d'anomalies, mais surtout de prévision [17].

Notre objectif, dans ce manuscrit, est d'analyser des dynamiques et de trouver de la nouveauté. Nous voulons donc observer les évolutions de différentes séries temporelles tirées des données textuelles. Les données textuelles sont des données qui sont régies par des lois de probabilités et dont l'évolution n'est pas soumise qu'à un bruit aléatoire. Les observations présentes dans les séries temporelles que nous allons analyser sont liées (la notion de **dépendance**) et la loi qui régit cette liaison est stable, c'est-à-dire que nous avons une dépendance **stationnaire**. Une série temporelle $z_1, z_2, ..., z_t$ est dite stationnaire si pour toute fonction f, $f(z_1, z_2, ..., z_t)$ et $f(z_{1+k}, z_{2+k}, ..., z_{t+k})$ sont régi par la même loi.

1.3 Jeux de données

Pour ce travail de recherche, nous avons travaillé avec plusieurs jeux de données textuelles. Notre but est de développer des algorithmes génériques fonctionnant sur les données d'EDF. Ces données ne pouvant être partagées et dans un souci de reproductibilité nous travaillons aussi avec des données publiques. Dans ce manuscrit, nous présenterons en majorité des applications sur des données publiques, mais nous les illustrerons sur les données EDF.

Dans le domaine de la détection de la nouveauté, il n'existe pas de jeu de données annotées nous permettant de résoudre nos tâches de manière supervisée. Nous avons vu que notre but principal est de détecter des thématiques ayant des dynamiques complexes, mais les thématiques ne sont pas explicitement annotées comme "émergente" ou "cyclique". Nous devons donc nous concentrer sur des jeux de données existants qui présentent trois spécificités :

1. ce doit être des données **textuelles** : nous nous concentrons exclusivement sur des données textuelles. Bien que nous utilisions des techniques qui ne sont pas originellement appliquées sur ce type de données, notre cas d'application est bien lié au texte. Nous développons des outils capables de s'appliquer sur des corpus de différentes langues, mais nous n'avons pas pour objectif de traiter des corpus multilingues.
2. ce doit être des données **temporelles** : elles doivent évoluer dans le temps et présenter un nombre important d'évolutions (que nous appellerons ici "instant"). En effet, nous ne pouvons pas nous contenter d'un jeu de données comportant moins d'une dizaine d'instants : notre but est de détecter des phénomènes temporels complexes et il est critique de pouvoir évaluer la réactivité des systèmes que nous développons.
3. ce doit être des données **catégorisées** : chaque document présent dans le jeu de données doit être associé à une ou plusieurs catégories. Ces catégories

nous permettent d'avoir une vérité terrain quant à leur dynamique temporelle : nous pouvons observer des catégories "émergente" ou "cyclique" et ainsi mesurer la performance de nos modèles. Que la catégorisation soit faite automatiquement en amont ou manuellement par des experts métiers ne nous importe pas : dans ce travail, nous prenons le parti de faire confiance à la catégorisation des documents.

Dans ce manuscrit, nous avons utilisé trois jeux de données différents dans deux langues différentes. Deux jeux de données sont publics et un jeu de données appartient à EDF et ne peut être diffusé. Le tableau 1.2 résume le contenu des jeux de données utilisés.

— *New York Times Annotated Corpus* (NYTAC) [18] [1] : ce corpus est composé de 1,8 million d'articles de presse en anglais publiés par le *New York Times* entre 1987 et 2007. Parmi eux, 1,5 million (83%) d'articles ont été manuellement annotés par catégorie. Chaque article peut être associé à plusieurs catégories selon une hiérarchie prédéfinie. Dans ce manuscrit, nous nous basons sur les catégories les plus générales comme *terrorism*, *motion pictures*, *politics*, *restaurants*.

— Article scientifique : ce corpus est composé d'environ 8000 résumés d'articles scientifiques en anglais publiés dans des conférences internationales entre 1990 et 2005. Chacun de ces articles scientifiques a été publié dans une conférence spécialisée qui a été catégorisée en 5 domaines : *theory*, *database*, *datamining*, *visualization*, *medical*.

— Courriels clients EDF : ce jeu de données privé appartient à l'entreprise EDF. Il contient des courriels envoyés par des clients à l'entreprise entre octobre 2018 et octobre 2019. Afin d'être traités le plus rapidement possible,

1. https://catalog.ldc.upenn.edu/LDC2008T19

	# docs	Langue	# catégories	Temporalité	Périodicité
NYTAC	1.8M	Anglais	13	1995-2005	Jour
SCI	8337	Anglais	4	1990-2005	Annuel
EDF	100k	Français	13	Oct. 2018-Oct. 2019	Jour

Table 1.2: Résumé des jeux de données utilisés dans ce manuscrit

ces courriels sont classés automatiquement dans des catégories définies au préalable par des experts métiers d'EDF. Ces catégories sont au nombre de 13 et concernent des problématiques comme le "contrat", la "relation client", les "factures", les "interventions techniques" ou encore la "solidarité". Pour des raisons de réglementations RGPD (Règlement Général sur la Protection des Données), l'ensemble de ces courriels sont anonymisés avant tout traitement et toutes les informations permettant l'identification d'un client sont supprimées.

1.4 Contributions et organisations du manuscrit

La thèse présentée dans ce manuscrit a pour objectif d'étudier et de proposer des algorithmes de détection de la nouveauté dans des données textuelles afin de mettre à jour des mots, documents ou thématiques relatifs à une "émergence douce". Ce type d'émergence est définie en opposition au domaine de la détection d'évènements soudains. Ces algorithmes sont utilisés dans les systèmes d'informations de l'entreprise EDF afin d'améliorer la compréhension des attentes clients.

Nous décrivons nos travaux dans la suite de ce manuscrit qui se compose de la façon suivante :

Chapitre 2 - Définition de la nouveauté

Le chapitre 2 aborde la définition de la nouveauté afin de développer une définition claire et générale dans le cadre des données textuelles. Nous introduisons le concept de la nouveauté de manière générale dans la littérature en détaillant l'état de l'art dans la section 2.2. Cette section permet d'apporter une distinction entre les concepts de nouveautés, d'anomalies, d'évènements et d'étudier les différentes familles d'approches développées dans la littérature. La section 2.3 apporte des précisions sur la définition de la nouveauté au niveau des données textuelles et fait la distinction sur les différents niveaux de détection. Enfin, nous montrons en section 2.4 les difficultés autour de l'évaluation de telles méthodes et nous nous concentrons sur un travail de comparaison des méthodes existantes. Ce travail a donné lieu à la publication suivante.

Clément Christophe, Julien Velcin, Jairo Cugliari, Philippe Suignard, Manel Boumghar. **How to detect novelty in textual data streams ? A comparative study of existing methods**. In *AALTD Workshop @ECML-PKDD 2019*

Chapitre 3 - Détection des éléments nouveaux

Dans le chapitre 3, nous présentons des techniques de détection d'éléments nouveaux : des mots, des documents et des thématiques. Pour cela, nous varions les techniques de représentation en explorant d'abord les modèles thématiques probabilistes dans la section 3.3. Nous utilisons ensuite une observation faite par rapport aux mouvements dans des espaces de plongements afin de détecter des mots associés à des thématiques émergentes. Ces modèles sont évalués en simulant l'arrivée de thématiques émergentes dans différents corpus. Ces travaux ont donné lieu aux publications suivantes :

Clément Christophe, Julien Velcin, Manel Boumghar. **Utilisation de techniques de modélisation thématiques pour la détection de nouveauté dans des flux de données textuelles**. In *EGC 2018, vol. RNTI-E-34, pp.239-250.*

Clément Christophe, Julien Velcin, Jairo Cugliari, Manel Boumghar, Philippe Suignard. **Monitoring geometrical properties of word embeddings for detecting the emergence of new topics**. En cours d'évaluation.

Chapitre 4 - Surveillance de plans de classement prédéfini

Le chapitre 4 aborde la question des nouveautés de volumes. Nous utilisons les catégories de plans de classement prédéfinis et nous surveillons les dynamiques afin de lancer des alertes lorsque celles-ci semblent devenir anormales. Comme notre hypothèse de base consiste à dire qu'une nouveauté apparaît lorsque la dynamique ne correspond plus à ce qui est attendu, nous basons notre approche sur un système de prévision de série temporelle. Nous développons un système basé sur des variables exogènes extraites du contenu textuel et nous analysons ses erreurs de prédictions avec une méthode d'analyse séquentielle. Nous testons cette approche sur les données du *New York Times* et sur les données courriels d'EDF. Ce travail

a donné lieu à la contribution suivante :

Clément Christophe, Julien Velcin, Jairo Cugliari, Philippe Suignard, Manel Boumghar. **Change detection in textual classification with unexpected dynamics**. En cours de révision pour *Expert System With Applications 2021*.

Chapitre 2

Définition de la nouveauté

Dans ce chapitre, nous étudions la définition du concept central de ces travaux de thèse : la nouveauté. Nous voyons que ce concept n'est pas très bien défini et dépend du type de domaine qui est étudié ainsi que de la tâche que nous voulons résoudre. Nous commençons par présenter les différentes définitions de la nouveauté qui existent dans la littérature en général puis nous nous concentrons sur la définition précise de la nouveauté pour les données textuelles. Enfin, nous voyons comment nous pouvons évaluer concrètement des modèles de détection en fonction des tâches à résoudre.

2.1 Introduction

Le terme "nouveauté" n'est pas un concept très bien défini dans la littérature et sa définition exacte dépend fortement de la tâche que nous cherchons à résoudre. De manière générale, la nouveauté est définie comme "ce qui ne ressemble à rien de ce qui a déjà été observé". La tâche de détection de la nouveauté peut, quant à elle, être définie comme le fait de reconnaître des entités qui diffèrent, dans une certaine mesure, de ce qui a pu être observé dans le passé. Ce champ de recherche a mené

au développement de multiples approches qui sont, souvent, appliquées sur des jeux de données de grands volumes avec une très grande majorité d'exemples classifiés comme "normaux".

La détection de la nouveauté est une tâche étudiée dans différents domaines utilisant de larges bases de données. Cela peut aller de la médecine, la détection de changement dans des processus industriels, la détection d'intrusion dans des systèmes de sécurité, la vidéosurveillance, la robotique et la fouille de textes. La complexité et la diversité de ces systèmes sont telles qu'il est impossible, en tant qu'humain, d'observer cette nouveauté à l'oeil nu. Il existe plusieurs types de nouveauté qui ne sont pas connus *a priori*, ce qui rend impossible l'utilisation d'algorithmes classiques de classification multi classes. Le domaine de la détection de la nouveauté résout ce problème en construisant un modèle qui apprend de lui-même les caractéristiques de la "normalité" dans un jeu de données. Les observations qui ne font pas partie de cette classe "normale" sont ensuite traitées de diverses manières avant d'être déclarées anormales.

Le domaine de la détection de la nouveauté est proche des domaines de la détection d'anomalie et de la détection d'*outlier*. Un *outlier* est défini comme un point isolé par rapport aux autres points d'un jeu de données. Un *outlier* est identifié avant qu'un concept de "normalité" soit associé aux données. Une anomalie est un point isolé qui est identifié dans des données déjà considérées comme "normales".

Bien qu'il existe des définitions pour les concepts d'anomalies, d'*outlier* et de nouveauté, nous voyons qu'elles sont très proches et qu'il est difficile de faire la distinction. Dans ce chapitre, nous présentons différents travaux portant sur la nouveauté en essayant d'en extraire une définition générale et applicable dans notre cadre industriel. Nous commençons par décrire des travaux dans la littérature en général puis nous nous concentrons sur la nouveauté spécifiquement dans les données textuelles. Enfin, nous étudierons les différentes manières d'évaluer un système de détection de nouveauté.

2.2 La nouveauté en général

Bien que le concept de nouveauté ait été étudié depuis longtemps dans de nombreux domaines, nous ne trouvons pas de définition précise qui permettrait de l'appliquer simplement à notre problématique. Comme nous l'avons vu dans l'introduction, nous nous intéressons à la détection de nouveauté "émergente". Ce domaine est proche de la détection de "signaux faibles". Le concept de nouveauté est aussi très proche des concepts d'anomalies et d'*outlier*. Dans cette partie, nous présentons quelles sont les principales différences entre ces trois concepts et comment la nouveauté au sens de "signal faible" se démarque.

La nouveauté comme signal. Le terme de nouveauté est spécifiquement lié à un signal : une nouveauté est une évolution anormale ou inattendue d'un signal, c'est-à-dire que l'état observé est différent de ce que nous pouvions attendre. La nouveauté se matérialise par un changement anormal ou inattendu dans la nature du signal. Ce changement inattendu, matérialisé au début de l'observation par des signaux faibles, est continu et peut amener à l'observation de signaux forts, plus faciles à détecter. C'est ce processus de passage de signaux faibles à forts qui caractérise la nouveauté. Dans des jeux de données volumineux, des signaux faibles seront toujours observables, mais une grande majorité d'entre eux resteront à l'état faibles, ce que nous considérerons comme du "bruit". Cette transition de l'état de signal faible à fort est illustrée dans la figure 2.1. Tandis que la nouveauté correspond à un signal temporel, donc à un enchaînement d'observations, les anomalies et *outliers* correspondent à une observation unique dans le temps.

Dans la littérature, plusieurs recherches nous permettent de mieux caractériser cette notion de nouveauté dans le sens d'évolution d'un signal faible. C'est le cas des travaux de Hiltunen [19] [20] qui nous apporte plusieurs définitions de signaux faibles. En effet, son manuscrit de thèse [20] fait référence à une étude menée par Kuusi [21] dans laquelle les participants (des scientifiques finlandais) devaient exposer leurs avis

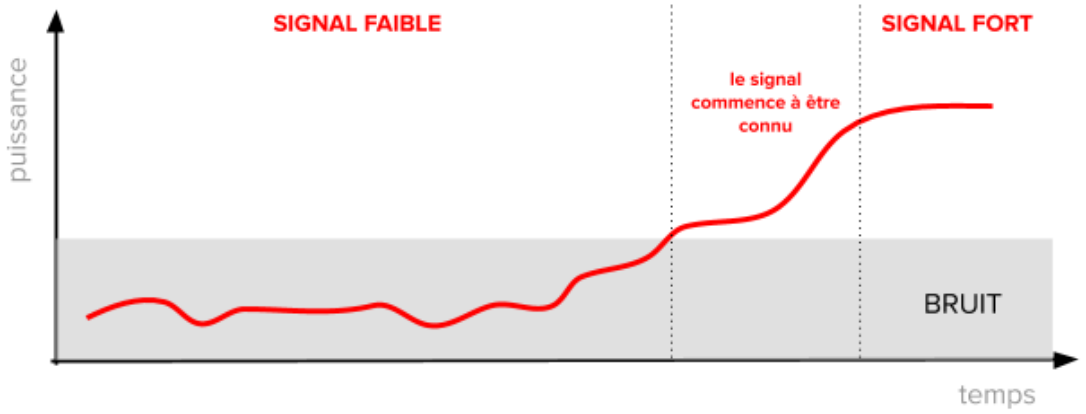


FIGURE 2.1: Signal modélisant l'apparition d'une nouveauté.

sur les caractéristiques d'un "signal faible". Deux définitions sont majoritairement ressorties :

- *Un signal faible "nouveau" est un signe annonciateur de changement, qui devient fort en se mélangeant à d'autres signaux. L'intérêt d'un tel signal est déterminé par l'objectif des observateurs. Un signal faible "nouveau" requiert : soutien, masse critique, croissance de son espace d'influence et acteurs dédiés.*
- *Un signal faible "nouveau" doit venir ou être reconnu par un groupe d'experts. Ce signal anticipe des phénomènes ayant des impacts sur le futur et peut inclure des caractéristiques qui doivent être détectées au plus tôt. Un signal faible nouveau se renforce par lui-même au cours du temps et est une alerte précoce d'une tendance émergente.*

D'autres dimensions de signaux faibles ont été mises en évidence par d'autres chercheurs. Par exemple, Rossel [22] met en lumière de nouvelles problématiques. Premièrement, la difficulté de se savoir en face d'un signal faible. Dans un second temps, il ajoute que l'identification d'un signal faible comme une expression précoce d'un changement dépend de notre propre interprétation en temps qu'humain. Sa conclusion consiste à dire que la détection des signaux faibles repose sur la possibilité de prendre du recul sur nos propres connaissances et nos limitations. Il recommande d'expliciter au maximum nos attentes afin de mettre la tâche de détection

de nouveauté en perspective.

Une deuxième question présente dans les travaux de thèse de Hiltunen [20] est la suivante : “Un signal faible est-il un signe de l'émergence d'un problème ou alors est-il le problème lui-même ?”. Elle fait référence aux travaux de Schultz [1] pour lequel les termes de “signaux faibles”, “tendance émergente” et “annonce de changement” veulent plus ou moins dire la même chose : la source du changement qui existe seulement dans un petit nombre d'entités (par exemple de documents), qui, multiplié par un certain nombre d'entités constituerait une vraie tendance. Mannermaa [23] définit un signal faible comme un phénomène qui a peu de probabilité d'apparaître, mais qui a un gros potentiel d'influence.

Au vu des différents travaux que nous venons de décrire et des besoins d'EDF, nous retiendrons la deuxième définition de l'étude de Kuusi [21] comme quoi un signal faible se renforce au cours du temps et est une alerte précoce d'une tendance émergente.

Différentes approches de détection selon les domaines. Tandis que la définition de nouveauté ne fait pas consensus dans la littérature en général, certains travaux comme ceux de [9], [24], [25] ont étudié et regroupé les différentes méthodes existantes pour détecter la nouveauté. Les travaux de [24] différencient seulement 2 manières de détecter la nouveauté : les approches statistiques et les approches basées sur les réseaux de neurones. De son côté, [9] classifie les techniques de détection de la nouveauté en 5 catégories distinctes :

— Les méthodes **probabilistes** : nous cherchons à estimer la densité d'une classe normale et suppose que des zones de basse densité ont peu de chances de contenir des données normales.
— Les méthodes basées sur **la distance** : nous partons du principe qu'une nouveauté va apparaître loin de ses plus proches voisins.

1. http://infinitefutures.com/essays/prez/holescan/sld005.htm

— Les méthodes basées sur **la reconstruction** : en entraînant un modèle de régression, nous supposons que l'observation d'une erreur importante entre la prédiction et la valeur réelle donne du poids à un score de nouveauté.
— Les méthodes basées sur **le domaine** : le but est de définir une frontière autour des données d'entraînements considérées comme normales. Cette frontière peut être construite de manière automatique ou à l'aide de connaissances métiers.
— Les méthodes basées sur **des techniques de théorie de l'information** : nous calculons l'apport informationnel des données d'entraînements grâce à l'entropie ou d'autres techniques basées sur la théorie de l'information de Shannon. Elles se basent sur le principe qu'une nouveauté va modifier significativement le contenu informationnel d'un jeu de données.

2.2.1 Les méthodes probabilistes.

Les approches basées sur les méthodes probabilistes estiment la fonction de densité qui a pu générer les données. Cette distribution peut ensuite être étudiée pour définir une frontière au-delà de laquelle les points sont considérés comme nouveaux. Dans le cas classique, des tests statistiques sont utilisés pour déterminer si les données dans un échantillon de test ont pu être générées par la distribution dite "normale". Plusieurs exemples de travaux se basant sur ces tests statistiques sont listés dans [9]. Ces tests ne font pas la distinction entre véritable nouveauté ou simple anomalie dans les données.

Plusieurs travaux ont contribué à la définition mathématique de la nouveauté dans des scénarios non supervisés. La tâche principale consiste à construire une règle de décision qui permet de faire la différence entre une classe normale (souvent dites "nominale" dans la littérature) et la classe de nouveauté. La classe normale est donc la classe correspondante à la loi de distribution sous-jacente des données. Dans la littérature, lorsque seules les étiquettes de la classe normale sont disponibles, c'est

un problème de *one-class classification* que nous appelons aussi “Détection de nouveauté inductive”. Les approches classiques présument que les nouveautés sont des anomalies par rapport à la distribution de la classe normale et construisent une méthode de détection en estimant le niveau de la densité de cette dernière [26–29]. Cette famille de méthode suppose souvent que la nouveauté sera uniformément distribuée sur le support de la densité de la classe normale. Cette dernière supposition est souvent fausse en pratique et ces méthodes ne fonctionnent pas lorsqu'il y a toujours un chevauchement important entre les densités des deux classes.

Souvent, le classifieur aura à disposition des exemples aléatoires $x_1, x_2, ...x_m$ associés à la classe normale. Cet étiquetage sera, dans la plupart des cas, obtenu avec l'aide d'experts du domaine. Cependant, il n'existe pas d'exemples explicitement associés à la classe nouvelle, ce qui ne permet pas d'utiliser des algorithmes de classification classique.

2.2.2 Les méthodes basées sur la distance

Les méthodes de détection de la nouveauté basées sur la distance se basent sur des mesures de distances spécifiques permettant de calculer la similarité entre deux points de données. Deux familles de techniques distinctes sont utilisées : les méthodes basées sur les plus proches voisins et les méthodes de *clustering*.

Les méthodes basées sur les plus proches voisins (kNN pour *k-Nearest neighbour* partent du principe que des exemples dits “normaux” dans les données ont des voisins proches dans l'espace de représentation. Les exemples “anormaux” (anomalies ou véritables nouveautés) sont isolés dans cet espace. Dans ce cas, le choix de la distance à calculer entre les différents exemples est primordial. Si la distance euclidienne classique est le choix le plus populaire, il existe d'autres définitions de distances [30] :

— Distance Euclidienne : la distance euclidienne entre deux vecteurs ϕ_1 et ϕ_2 est définie par :

$$EucliDist(\phi_1, \phi_2) = \|\phi_1 - \phi_2\|_2 \tag{2.1}$$

Cette distance est symétrique, supérieure ou égale à 0, mais n'a pas de borne supérieure.

— Dissimilarité Cosinus : la dissimilarité cosinus est définie par rapport à l'angle entre deux vecteurs ϕ_1 et ϕ_2 :

$$CosDiv(\phi_1, \phi_2) = 1 - \frac{\phi_1.\phi_2}{\|\phi_1\|\|\phi_2\|} \tag{2.2}$$

La dissimilarité Cosinus est symétrique, borné entre 0 et 2.

— Distance de Jaccard : elle se base sur l'indice du même nom et se définit comme :

$$JacDist(\phi_1, \phi_2) = 1 - \frac{\sum_i min(\phi_{1_i}, \phi_{2_i})}{\sum_i max(\phi_{1_i}, \phi_{2_i})} \tag{2.3}$$

Elle est toujours comprise entre 0 et 1.

— Divergence de Kullback-Leibler : utilisé pour comparer des distributions de probabilités ϕ_1 et ϕ_2, elle se définit comme :

$$KLDiv(\phi_1, \phi_2) = \sum_i \phi_{1_i} log \frac{\phi_{1_i}}{\phi_{2_i}} \tag{2.4}$$

Elle n'est pas symétrique. Elle est strictement supérieure à 0, mais n'a pas de borne supérieure.

— Divergence de Jensen-Shannon : elle est une version symétrique de la Divergence de Kullback-Leibler :

$$JSDiv(\phi_1, \phi_2) = \frac{1}{2} KLDiv(\phi_1, \phi) + \frac{1}{2} KLDiv(\phi_2, \phi) \tag{2.5}$$

avec $\phi = \frac{1}{2}(\phi_1 + \phi_2)$

Cette version est symétrique et bornée entre 0 et $ln2$.

— Distance de Mahalanobis :

$$MahaDist(\phi_1, \phi_2) = \sqrt{(\phi_1 - \phi_2)^T S^{-1} (\phi_1 - \phi_2)} \tag{2.6}$$

où S^{-1} correspond à l'inverse de la matrice de covariance.

Ces mesures de distances sont rarement efficaces pour traiter des données en grande dimension. C'est pourquoi il est maintenant courant d'utiliser des méthodes dans lesquelles les anomalies et les nouveautés sont détectées via la recherche de sous-ensembles vides. C'est le cas notamment dans [31–34]. Les auteurs de [31] proposent d'utiliser une somme pondérée de la distance par rapport aux k plus proches voisins pour tous les points de données et de considérer comme *outlier* les points présentant la plus grande distance. Ils optimisent leur technique en ne considérant que les hypercubes de l'espace contenant peu de points de données : un hypercube contenant beaucoup de points aura de fortes chances de ne pas contenir d'*outlier*. Ce type de technique est étendu dans [32–34].

Une technique classique dans le domaine de la détection d'anomalie ou d'*outlier* consiste à calculer le *Local Outlier Factor* (LOF) de chaque point. Le LOF d'un point est basé sur le ratio entre la densité locale autour d'un point et la densité autour de ses voisins. Il prend des valeurs importantes, car il quantifie à quel point un point est isolé par rapport à ses voisins. C'est une technique pour identifier des anomalies uniques (un seul point), mais peu efficaces pour détecter des groupes isolés.

Enfin, certaines méthodes se basent sur des approches de *clustering* pour détecter des anomalies et des nouveautés. Dans ce type de configuration, le but est de faire la différence entre la classe dite "normale" et le reste des exemples qui seront considérés comme des *outliers*. Souvent, la classe normale est identifiée via un certain nombre de points de références. La distance entre un point de donnée et le point de référence le plus proche est utilisée pour quantifier l'anormalité d'une observation dans un

espace. C'est le cas dans [35–37] qui utilisent l'algorithme de *k-means* pour identifier ces points de références. D'autres algorithmes de *clustering* ont été utilisés pour la détection de nouveautés : *fuzzy c-means* [38], *possibilistic c-means*[39], *wavelets* [40] et *Hidden Random Markov Fields*(HMRFs) [41].

2.2.3 Les méthodes basées sur la reconstruction

Les méthodes basées sur la reconstruction se basent sur l'entraînement d'un modèle de régression. Lorsque de nouveaux points de données "test" arrivent dans le jeu de données, l'erreur de reconstruction par rapport au modèle précédemment entraîné permet de calculer un score de nouveauté. Ce type de méthode est traditionnellement utilisée pour des applications où la sécurité est critique et où les données arrivent en flux. Généralement, ces approches sont basées soit sur des réseaux de neurones soit sur des méthodes de réduction de dimensions.

Les méthodes basées sur les réseaux de neurones comme [42] cherchent à estimer la capacité de la rétropropagation d'un *Multi-Layer Perceptron*(MLP) pour détecter la nouveauté. Généralement, ces méthodes nécessitent de fixer un seuil sur les valeurs de sortie, seuil au-delà duquel le point est considéré comme nouveau. Une autre approche consiste à explorer les réseaux de neurones probabilistes dans lesquels il y a autant de noeuds que de points de données. Les liens entre les noeuds représentant une forme de distance entre les exemples, ces réseaux parviennent à classifier chaque point et donc quantifier l'incertitude autour de l'appartenance d'un point à une classe. Une grande incertitude est alors un signe de nouveauté. Dans le passé, les méthodes de détection de nouveauté basées sur des réseaux de neurones utilisaient souvent le concept de cartes auto-organisatrices (*Self-organizing maps* (SOMs)) [43]. Ces SOMs sont généralement utilisées pour identifier des *clusters* dans un cadre non supervisé. Si les SOMs sont entraînées sur des données dites "normales", nous pouvons calculer une distance entre un point et le centre des *clusters* afin de détecter la nouveauté. L'avantage de ces SOMs réside dans le

fait qu'elles conservent la structure topologique des données. Elles ont été utilisées pour plusieurs applications de détection de nouveauté [44, 45] et développées pour introduire un aspect dynamique [46, 47]. Plus récemment, l'utilisation massive des méthodes basées sur des réseaux de neurones a permis de nouveaux développements autour de la détection de la nouveauté. C'est le cas dans les travaux de Ghosal [48] et Kliger [49] où des architectures récentes comme les *Convolutional Neural Network* (CNN) et les *Generative Adversarial Networks* (GAN) sont investigués pour détecter de la nouveauté au niveau des documents.

Enfin certaines méthodes partent du principe que les données peuvent être projetées dans un espace de plus petite dimension afin de mieux distinguer les données "normales" et la nouveauté. La plupart de ces méthodes sont basées sur l'algorithme de *Principal Component Analysis*(PCA) pour effectuer la réduction de dimension [50–52]. Des extensions de cette PCA ont aussi été utilisées pour la détection de la nouveauté : *kernel*-PCA [53–55] ou *t-distributed stochastic neighbor embedding* (t-SNE) [56].

2.2.4 Les méthodes basées sur le domaine

Les méthodes basées sur le domaine permettent de détecter les nouveautés ou les anomalies en calculant une frontière autour des données d'entraînements. En se concentrant spécifiquement sur la construction de la frontière, elles sont insensibles à la densité des données dites "normales". L'appartenance aux classes normale ou anormale est calculée par rapport à cette frontière. Traditionnellement, cette dernière est calculée via des algorithmes de type *Support Vector Machine* (SVM). Ces SVM sont couramment utilisés pour résoudre des problèmes de classification. Ils définissent des frontières linéaires en maximisant la marge entre deux classes. Deux extensions majeures à l'algorithme SVM l'ont adapté au problème de la détection de la nouveauté : l'approche de *Support Vector Data Description* (SVDD) [57] et les *one-class SVM*.

Les SVDD définissent la frontière autour des données comme étant la limite de l'hypersphère ayant le plus petit volume, mais englobant le plus grand nombre de données "normales". Ce type de méthode fonctionne moins bien lorsque l'espace est en très grande dimension ou lorsque les données ne sont pas distribuées de manière sphérique. Certains travaux ont donc étendu ce concept de SVDD pour l'utiliser dans différentes applications [58–60].

L'idée des *one-class SVM* proposé par Schölkopf [61] permet de définir une frontière non linéaire dans un espace transformé selon une fonction noyau. Cette approche nécessite de définir a priori un certain nombre d'exemples "normaux" qui peuvent être de l'autre côté de la frontière. Le paramétrage de cette valeur influence grandement les performances d'un tel système. Les travaux de Roth [62, 63] tentent de fixer ce paramètre en utilisant une procédure de validation croisée. Aussi, les résultats dépendent fortement du choix de la fonction noyau choisi pour transformer l'espace. Ces *one-class SVM* sont généralement utilisées pour résoudre une tâche de détection de nouveauté dans des séries temporelles [64–66].

2.2.5 Les méthodes basées sur la théorie de l'information

Enfin, les méthodes basées sur la théorie de l'information partent du principe qu'une nouveauté ou une anomalie va significativement modifier le contenu informationnel d'un jeu de données. Généralement, ces méthodes utilisent des mesures comme l'entropie. Après avoir calculé l'entropie sur la totalité du jeu de données, elles cherchent à identifier les points qui auraient le plus d'influence sur cette valeur dans le cas où nous les retirerons. Ces méthodes ne considèrent pas la distribution des données. Elles dépendent fortement du choix de la métrique utilisée pour calculer le contenu informationnel et sont extrêmement coûteuses à calculer. Cependant, certains travaux ont bien appliqué ce type d'approche à des problèmes concrets [67–70].

Dans cette thèse, nous nous utiliserons des approches des 3 premières catégories, à savoir les méthodes **probabilistes**, basées sur la **distance** et sur la **reconstruction**.

2.3 La nouveauté dans des données textuelles

Lorsque nous travaillons avec des données textuelles, la nouveauté peut se définir de plusieurs manières. La structure et les modes de diffusion particuliers de ce type de données font que la nouveauté se caractérise à des niveaux de granularité différents. Une nouveauté peut se présenter, dans des données textuelles, au niveau des thématiques, des documents, des phrases, des mots ou des chaînes de caractère. Pour chaque algorithme utilisé, il faut préciser la granularité à laquelle nous analysons les données et adapter la méthode si besoin. Dans ce manuscrit, nous nous intéressons principalement à la nouveauté au niveau des mots, des documents et des thématiques.

2.3.1 La nouveauté au niveau des mots et des chaînes de caractères.

La nouveauté au niveau des **mots** peut se matérialiser de plusieurs manières. Premièrement, si un mot ne faisant pas partie de notre vocabulaire initial apparaît dans notre corpus, cela constitue un type de nouveauté. Il faut définir le sens du mot, le représenter dans un espace, étudier son contexte, etc. Dans la littérature, on dit qu'un mot est *Out-of-Vocabulary* car il est inconnu des différents modèles de langues et de représentations. Dans certains travaux [71], ce problème de représentations des mots nouveaux est réglé en étudiant les différentes chaînes de caractères connues qui le composent. Par exemple, si le nouveau mot inconnu est "`bonjour`", il pourra être représenté sous la forme de 2 chaînes de caractères connues comme "`bon`" et "`jour`".

Un système de détection de nouveauté pourrait lancer une alerte dès lors qu'un nouveau mot apparaît dans le vocabulaire et ainsi le représenter via les chaînes de caractères qui le composent. Cependant, un tel système serait extrêmement sensible aux fautes de frappe et d'orthographe qui composent les jeux de données utilisés dans l'industrie. Aussi, il ne permettrait pas de faire la différence entre une anomalie, du bruit (des évènements qui n'apparaissent qu'une seule fois) et de véritables nouveautés émergentes. Au-delà de l'apparition simple de mots jusque-là inconnus, la fréquence d'utilisation d'un mot peut varier de manière anormale et ainsi représenter une nouveauté intéressante à étudier. De plus, lorsque nous étudions des groupes de mots proches dans le texte (n-grammes de mots), nous pouvons aussi analyser leur fréquence de co-occurrence. Plusieurs travaux se basent sur l'observation des n-grammes pour détecter des évènements ou des anomalies [2, 72]. De tels systèmes se basent uniquement sur l'évolution de la fréquence des mots dans le temps en développant des systèmes basés sur des métriques classiques comme TFxIDF [1] ou en en créant de nouvelles. C'est le cas dans la méthode TopicSketch [2] dans lequel un algorithme est construit pour surveiller des mesures physiques comme la vitesse et l'accélération de la fréquence des mots et des n-grammes. Dans HUPC [72] et ET-EPM [73], les auteurs utilisent une métrique dite d'utilités pour ensuite extraire des motifs représentatifs de la nouveauté. Bien que ces méthodes soient efficaces pour la détection de nouveauté dans certaines applications, elles ne basent leur approche que sur des mesures dérivées d'un simple comptage des mots : le sens et le contexte du vocabulaire ne sont pas pris en compte.

Enfin, l'émergence des techniques de plongement de mots comme word2vec [74], glove [75], fasttext [71] ou encore plus récemment BERT [76] et ELMO [77] ont permis de se baser sur l'analyse des groupes de mots sur des fenêtres de contexte assez larges afin de représenter au mieux leur sens dans un espace vectoriel. Certains travaux [78–80] s'intéressent à l'évolution temporelle de ces espaces de représentations pour détecter des changements dans le sens des mots ou dans leurs utilisations.

Nous avons donc étudié plusieurs manières de modéliser le sens des mots dans des données textuelles et nous verrons dans les chapitres suivants comment ces techniques de modélisation peuvent être utilisées pour la détection de la nouveauté. Cependant, la nouveauté n'est pas seulement présente au niveau des mots : elle peut apparaître à d'autres niveaux.

2.3.2 La nouveauté au niveau des thématiques.

Un document textuel est composé de mots, mais c'est tout d'abord leur sens et leur voisinage qui déterminent les sujets abordés. Même si le sens des mots n'est pas modélisé, nous pouvons extraire les thématiques abordées dans un document en étudiant les co-occurrences entre les mots : c'est le principe des algorithmes de modélisation thématiques comme LSA [81], PLSA [15] ou LDA [4]. Avant d'étudier en détail le fonctionnement de ces algorithmes, nous allons préciser la notion de thématique et expliquer comment la nouveauté peut se manifester à ce niveau-là.

Dans un ensemble de documents, une thématique correspond à une distribution de probabilité sur l'ensemble du vocabulaire, c'est-à-dire que les mots les plus probables d'une thématique peuvent nous permettre de caractériser cette dernière. Par exemple, si les mots les plus probables sont : *tacle*, *attaquant*, *ballon*, *but*, *match*, nous pouvons raisonnablement penser que nous sommes face à une thématique qui traite du sujet "football". La plupart du temps, les algorithmes de modélisations thématiques ne donnent pas de titre automatiquement aux thématiques : ils se contentent de retourner les distributions de probabilités sur le vocabulaire. Selon l'algorithme utilisé, un document pourra être composé d'une seule ou de plusieurs thématiques.

En étudiant l'évolution de ces thématiques dans le temps, nous pouvons mettre en lumière plusieurs phénomènes intéressants du point de vue de la nouveauté :

— Modification d'une thématique : les mots qui composent une thématique peuvent évoluer. Par exemple, si le mot qui, originellement, décrivait le mieux la thématique devient de moins en moins probable avec le temps, nous pouvons dire que la thématique change de sens.
— Création d'une nouvelle thématique : des mots qui n'apparaissaient pas du tout ensemble auparavant commencent petit à petit à se mélanger et à former une thématique à part entière. C'est le cas par exemple avec l'émergence de nouveaux concepts tels que l'informatique ou encore dans le cas d'évènements particuliers tels qu'une guerre.
— Disparition d'une thématique : au contraire, une thématique peut disparaître si les mots qui la composent ne co-occurrent plus ensemble.
— Fusion de deux thématiques : deux thématiques peuvent se mélanger au point de former une seule même thématique. Par exemple, quand un article de presse titre : "*À huit jours des Jeux olympiques au Brésil, un homme lié à Daech a été arrêté à Rio*", les thématiques "*Terrorisme*"et "*Jeux Olympiques*" sont mélangées. Elles pourront former une seule thématique si cette tendance continue.
— Éclatement d'une thématique : lorsqu'une thématique prend beaucoup d'ampleur, elle peut être séparée en plusieurs sous thématiques si ces dernières utilisent des vocabulaires spécifiques. Ça peut être le cas, par exemple, si une thématique "*Sport*" se sépare en "*Basketball*" et "*Football*".

Plusieurs travaux de la littérature ont travaillé sur ces différents phénomènes. Par exemple, les travaux de Blei [6] modélisent le glissement sémantique des thématiques, c'est-à-dire la modification des thématiques. D'autres [82] modélisent la création de nouvelles thématiques via le processus de restaurants chinois. Dans ce manuscrit, nous nous concentrons surtout sur les phénomènes de modifications et de créations de thématiques. Notre objectif est de minimiser la quantité de données nécessaires pour effectuer cette détection.

2.4 Évaluation de la nouveauté

Nous avons vu que la nouveauté est un concept abstrait et que sa détection peut être approchée de diverses manières et sur différentes entités. L'ensemble des méthodes de détection de la nouveauté utilisent donc des méthodes d'évaluations différentes, en fonction des concepts qu'ils veulent détecter. Nous avons travaillé à l'unification de ces méthodes d'évaluations en définissant des scénarios d'apparitions de la nouveauté, en testant sur des données textuelles entièrement simulées et en définissant des méthodes précises d'évaluation. Ce travail nous permet de déterminer quelles sont les méthodes les plus efficaces pour résoudre notre tâche de détection de la nouveauté.

2.4.1 Les mesures d'évaluation

Les algorithmes travaillant sur la tâche de la détection de la nouveauté ne font souvent pas la différence entre nouveauté, évènement et anomalie, comme nous l'avons décrit précédemment. De plus, la détection ne s'effectue pas au même niveau d'abstraction : identification des mots, documents ou thématiques nouvelles. Enfin, il n'existe pas de jeu de données explicitement annoté par rapport à une définition générale de la nouveauté. Il est donc impossible d'utiliser des méthodes d'évaluations classiques d'approches supervisées. Afin de contourner ce problème, la plupart des travaux de la littérature se rapprochent d'une tâche existante, en résolvant des problèmes différents, et utilisent donc des méthodes d'évaluations différentes. Cette dernière observation rend les comparaisons difficiles. De plus, comme nous travaillons sur des données textuelles ayant souvent un rapport étroit avec les évènements réels connus du grand public (Twitter, article de presse, article scientifique, etc.), un certain nombre de méthodes travaillant sur la détection de la nouveauté, d'évènements ou sur l'évolution du langage de manière plus générale évaluent leur approche de manière qualitative plutôt que quantitative : c'est-à-dire

en illustrant avec un certain nombre d'exemples sur lesquels la méthode fonctionne (ou pas). Nous voulons, de notre côté, obtenir des résultats quantitatifs qui permettent de comparer les méthodes entre elles de la manière la plus objective possible.

D'autres méthodes utilisent, elles, des méthodes d'évaluation quantitative qui mettent en lumière divers phénomènes. Par exemple, [1] évaluent leur approche à l'aide de courbes DET (*Detection Error Tradeoff)*) permettant d'observer le compromis entre les taux de faux négatifs et de faux positifs dans les données. Ils constituent leur vérité terrain en sélectionnant, a priori, un certain nombre d'évènements à détecter. Dans [83], les auteurs choisissent de s'évaluer via une mesure de précision à N (P@N) afin de détecter des mots liés à des évènements choisis manuellement. Enfin, [84] utilisent des mesures classiques de précision, rappel et F-Mesure permettant de mesurer la performance de leur méthode.

Il est aussi courant d'évaluer au préalable la méthode sur des données synthétiques [2], pour lesquelles nous maîtrisons l'ensemble des paramètres de génération. En plus des méthodes d'évaluation classique, l'application TopicSketch évalue aussi les évènements détectés en calculant des métriques comme la cohérence par rapport à des évènements pré identifiés dans des données connues. Enfin, des travaux comme [72, 73] évaluent leur approche en comptant le nombre d'alertes lancées et le retard par rapport à des évènements là aussi pré identifiés.

Dans nos travaux, nous utiliserons des données simulées ainsi que des données réelles. Nous simulerons l'arrivée de la nouveauté ce qui nous permettra d'utiliser des mesures de précision, rappel et F-Mesure ainsi que des mesures de retards.

2.4.2 Notre définition de la nouveauté

Nous avons vu que la nouveauté peut se présenter sur des variations de signaux connus et sur l'apparition d'éléments nouveaux. Nous faisons, dans ce manuscrit,

la distinction entre ce que nous appelons “les nouveautés de volume” et les “nouveautés de structure”. Les nouveautés de volume représentent des variations inattendues dans un signal connu tandis que les nouveautés de structure correspondent à des modifications sous-jacentes de la structure des données : mélange de nouveaux termes qui forment une nouvelle thématique, etc..

Dans ce chapitre, nous étudions plus en détail les nouveautés de volumes. Nous considérons 3 types de nouveautés qui peuvent être détectés par les méthodes sélectionnées. Nous avons vu que la nouveauté est liée à un signal évoluant dans le temps et nous nous intéressons aux 3 types de mouvements présentés dans la Figure 2.2. Nous avons un signal correspondant à un signal faible devenant, progressivement, un signal fort : c’est ce que nous appellerons un signal **émergent** et c’est le cas qui nous intéressera le plus par la suite dans le cadre industriel. Nous étudions aussi un signal de type évènement : un signal faible qui devient très rapidement très fort puis redescend soit à son niveau précédent, soit à un niveau intermédiaire. Enfin, nous étudions un signal cyclique, c’est-à-dire un signal ayant une certaine périodicité. Le signal cyclique n’est pas forcément considéré comme de la nouveauté. Cela dépend de ce que l’utilisateur veut détecter. Cependant, nous choisissons de le garder dans notre étude afin de tester si les méthodes étudiées le considèrent comme nouveau ou non.

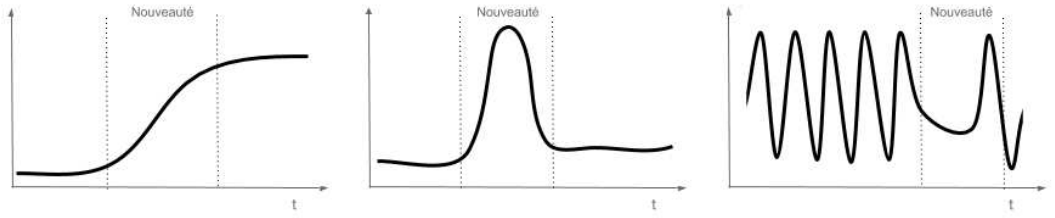


FIGURE 2.2: Les différents types de nouveautés : l’émergence, les évènements abrupts, les dynamiques cycliques

Nous proposons une méthodologie précise qui nous permet de mesurer l’influence des différents paramètres sur nos résultats. En plus des trois familles de signaux (émergent, évènement et cyclique), nous faisons varier la vitesse d’apparition de la

nouveauté afin d'obtenir, au final, 9 scénarios à tester. Ces scénarios sont présentés sur la Figure 2.3. Pour chaque expérience, nous mélangeons un scénario de nouveauté avec d'autres catégories appartenant à un signal constant.

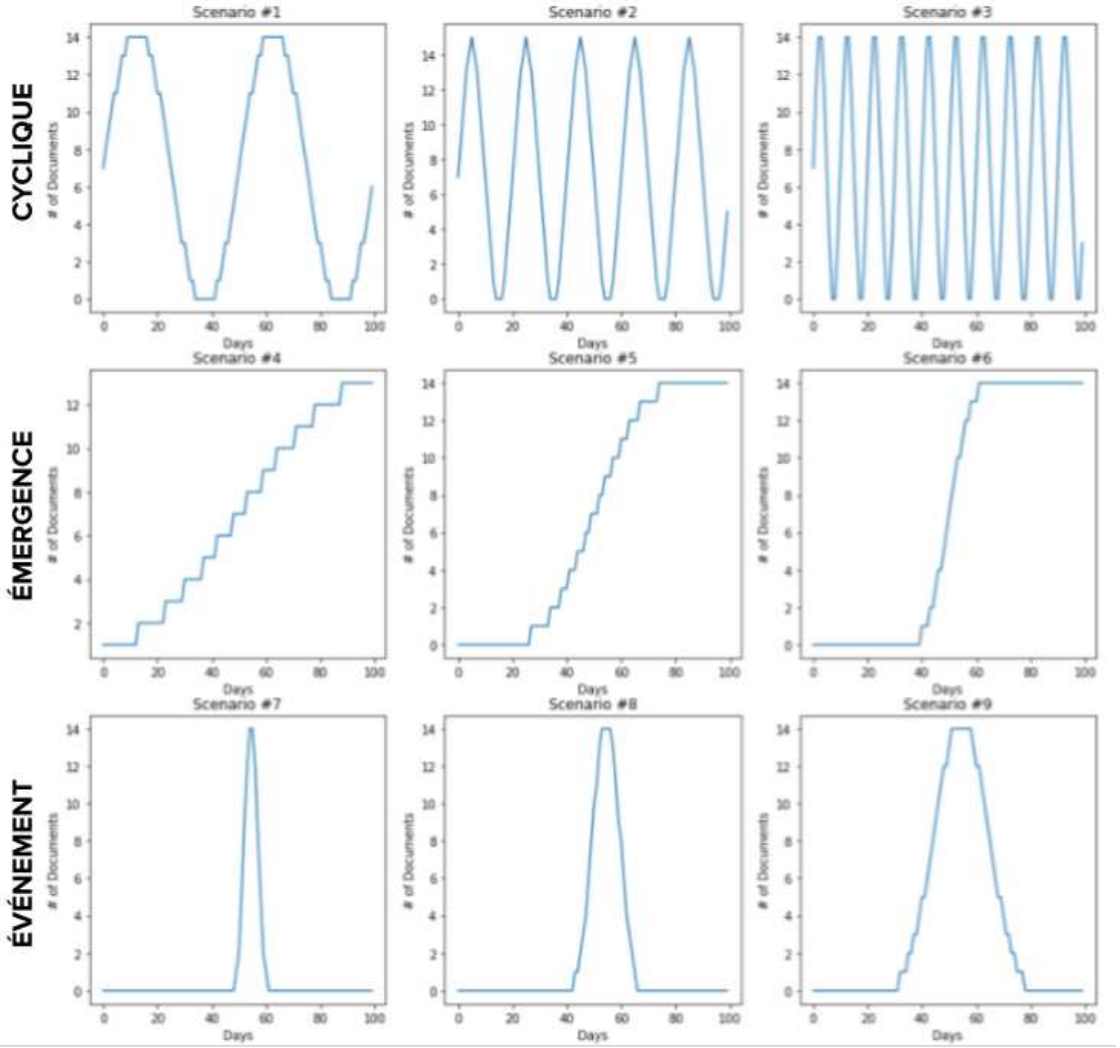


FIGURE 2.3: Scénarios simulés dans cette expérience.

Nous fixons deux tâches précises, associées à des mesures d'évaluation spécifiques, sur lesquelles nous voulons comparer les méthodes de la littérature :

— Tâche 1 : le but est de détecter les mots associés à la nouveauté introduite. Dans le cas où nous possédons une vérité terrain pour ces mots, il est possible d'utiliser des mesures classiques de précision, rappel et f-mesure.

— Tâche 2 : le but est de détecter les documents responsables de l'apparition de la nouveauté. Là aussi, si nous possédons une vérité terrain, nous pouvons utiliser les mesures classiques de précision, rappel et f-mesure.

2.4.3 Comparaison des méthodes de la littérature

Afin de résoudre les tâches énoncées précédemment, nous choisissons de tester des algorithmes provenant de différents domaines d'application : certains sont utilisés pour la détection d'évènements, notamment sur Twitter, d'autres pour la détection de nouvelles histoires dans des articles de presse et d'autres pour l'observation de l'évolution du langage dans le temps. Nous présentons quels sont ces algorithmes et comment nous les avons adaptés à notre tâche.

— ***Detections, bounds, and timelines : UMass and TDT-3*** [1] **(TF-IDF)** : ce papier est un des précurseurs dans le domaine de la détection de nouvelles histoires et donc de nouveauté. Les auteurs représentent les documents dans un espace construit avec la métrique TFxIDF [11] et calculent une dissimilarité cosinus (voir Eq. 2.2) pour effectuer une recherche par rapport aux plus proches voisins et ainsi identifier les documents nouveaux. Ils partent du principe qu'un document nouveau apparaîtra loin de ses plus proches voisins.

— ***Structured event retrieval over microblog archives*** [83] **(BS)** : ce travail est originellement conçu pour la détection d'évènements dans des données Twitter. Les auteurs développent une métrique de "*Burstiness Score*" basée sur l'évolution de la fréquence de chaque mot dans le temps. Pour détecter de nouveaux mots, nous considérons les mots avec le "*Burstiness Score*" le plus élevé.

— ***Towards effective event detection, tracking and summarization on microblog data*** [85] **(DF)** : cette méthode se base sur une métrique

Modèle	Tâche 1	Tâche 2
TF-IDF		X
BS	X	
DF	X	
OLDA	X	
TopicSketch	X	

TABLE 2.1: Tâche originellement résolue dans les modèles utilisés. **Tâche 1** : détecter des mots. **Tâche 2** : détecter des documents

de "*Document Frequency*" et sur une distance de Jaccard 2.3. Ils identifient des groupes de nouveaux mots chaque jour.

— ***On-line trend analysis with topic models : twitter trends detection topic model online data*** [84] **(OLDA)** : ce papier présente une méthode qui détecte des évènements sur des périodes de temps. Ils développent une variante du modèle LDA [4] qui met à jour la proportion de mot/thématique chaque jour. Les nouvelles thématiques sont identifiées via une distance de Jensen-Shannon 2.5 en les comparant aux thématiques des jours précédents. Nous utilisons cette technique pour détecter des mots nouveaux en utilisant les mots les plus probables des thématiques identifiées comme nouvelles.

— ***TopicSketch : Real-time bursty topic detection from Twitter*** [2] **(TopicSketch)** : ce papier présente une méthode adaptée à la détection d'évènement dans des données Twitter. Les auteurs proposent de surveiller la vitesse et l'accélération des fréquences des mots afin de détecter de la nouveauté. Chaque jour, le modèle est capable de lancer une ou plusieurs alertes sur un certain nombre de mots du vocabulaire.

Le Tableau 2.1 nous donne des informations sur les tâches qui sont originellement résolues par les différents algorithmes que nous avons choisis. Afin d'évaluer et de comparer ces méthodes dans les mêmes conditions, nous avons besoin de les adapter aux tâches pour lesquelles elles n'étaient pas prévues à l'origine. Même si cette adaptation n'est pas optimale et ne représente pas forcément la meilleure manière

de résoudre une tâche, elle nous donne des indications sur le potentiel, les forces et les faiblesses de chacune de ces approches.

Nous avons adapté la méthode **TF-IDF** sur la tâche de la détection de mots nouveaux en utilisant l'*Inverse Document Frequency* de chaque mot. Cette mesure est traditionnellement associée à la rareté et nous voulons vérifier si le fait d'avoir des valeurs importantes chaque jour est un bon indicateur de nouveauté. Pour la méthode **BS**, nous agrégeons les *Burstiness Score* des mots présents dans un document pour calculer un score associé au document. Nous testons plusieurs manières d'agrégation : par rapport à la moyenne, la médiane ou par rapport à un certain percentile. Pour utiliser la méthode **DF** pour la détection de nouveauté au niveau des documents, nous nous inspirons de [1] et nous utilisons une recherche basée sur les plus proches voisins dans un espace de représentation construit pour les documents sur la base de leur *Document Frequency*. En utilisant **OLDA** pour détecter de nouvelles thématiques, nous sélectionnons les documents les plus probables pour les thématiques identifiées comme nouvelles. La méthode **TopicSketch** est utilisée pour lancer des alertes quand la fréquence d'un terme est considérée comme anormale. Pour la détection des documents, nous considérons ceux qui contiennent les mots détectés au moment où ils sont détectés.

Pour simuler nos données textuelles et ainsi obtenir des jeux de données avec certaines thématiques constantes et des thématiques ayant un scénario temporel correspondant à ceux présenter dans la figure 2.3, nous utilisons un modèle de mélange [2]. L'avantage d'utiliser de type de modèle pour simuler complètement notre jeu de données réside dans le fait que nous contrôlons l'ensemble des paramètres. Si les algorithmes ne fonctionnent pas dans ces environnements favorables, il y a de fortes chances pour qu'ils ne fonctionnent pas sur des données réelles où il y a typiquement plus de bruits. Pour chaque document d, nous lui assignons une thématique z et nous tirons aléatoirement 100 mots de sa distribution de probabilité parmi

2. Le code pour la simulation des données est disponible à https://github.com/clechristophe/NoveltySimulator

les 10.000 mots du vocabulaire. Il est important de noter que nous utilisons des approches basées seulement sur des sacs-de-mots et donc l'ordre dans lequel nous ordonnançons nos mots n'est pas important. Nous ne générons donc pas des documents, au sens strict du terme, mais plutôt des sacs-de-mots tirés d'une distribution. Le modèle de mélange qui nous permet de générer ces données dépend d'un hyperparamètre α qui contrôle le recouvrement des mots dans le tirage du vocabulaire. En faisant varier ce paramètre, nous sommes capables de générer des thématiques plus ou moins proches en termes de divergence de Kullback-Leibler 2.4. Nous avons simulé des probabilités de distributions sur le vocabulaire pour chaque thématique et associé une thématique par document. Nous générons donc les documents en les organisant temporellement selon les signaux des scénarios de la figure 2.3.

Ces données nous permettent d'évaluer les algorithmes présentés précédemment ainsi que l'impact des différents paramètres de générations sur les résultats. Le fait d'avoir des données simulées pour lesquelles nous connaissons la thématique considérée comme de la nouveauté nous permet d'avoir une vérité terrain pour les deux tâches que nous avons définies. Nous évaluons ces tâches avec des mesures la précision, le rappel et la F-Mesure.

Avant de présenter les résultats généraux, nous voulons déterminer l'influence du paramètre α de génération des données. Ce paramètre permettant de contrôler la divergence de Kullback-Leibler 2.4 (KL-Div) entre les thématiques générées, nous générons 6 types de jeux de données avec une KL-Div prenant les valeurs suivantes : 0.01,0.05,0.1,0.5,0.9 et 0.99. Notre hypothèse est que, comme la plupart des algorithmes se basent sur une mesure de similarité entre les documents ou sur l'évolution de la fréquence des mots, la détection devrait être plus difficile lorsque les thématiques sont plus proches. Nous testons cette hypothèse en observant l'évolution des performances pour la tâche de détection des documents. Nous voyons sur la figure 2.4 que notre hypothèse s'avère être globalement vraie, surtout pour les deux méthodes basées sur la recherche par rapport aux plus proches voisins et TopicSketch. Enfin, il est intéressant de noter que, bien que le niveau de

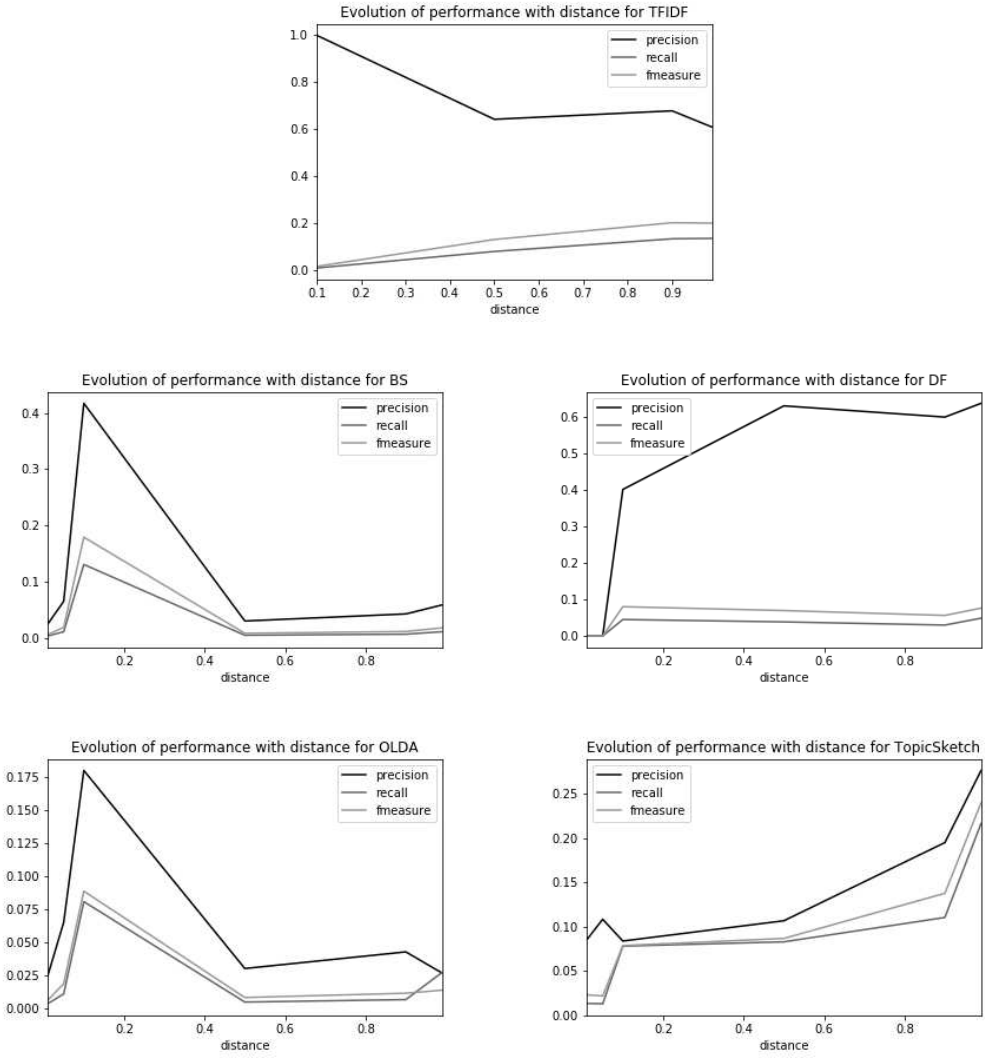


FIGURE 2.4: Évolution des performances par rapport à la divergence de Kullback-Leibler entre les thématiques pour chaque algorithme testé.

détection est très bas pour les méthodes DF et OLDA, elles montrent un résultat optimal pour KL-Div = 0.1.

Certaines méthodes que nous avons choisi d'évaluer dans ce travail sont spécialisées dans la détection d'évènements : OLDA et TopicSketch. Les évènements sont associés à des thématiques qui apparaissent très rapidement et en grande quantité avant de disparaître presque aussi rapidement. C'est un phénomène qui est souvent observable dans des données de type Twitter. C'est pour cela que nous choisissons d'évaluer l'influence de la vitesse d'arrivée de la nouveauté dans nos données

simulées sur les performances générales. Notre hypothèse est que ces algorithmes devraient être plus performants quand la nouveauté apparaît très rapidement. Nous observons dans la figure 2.5 que les deux méthodes sont plus sensibles à une pente forte : la détection est plus rapide lorsque la nouveauté apparaît rapidement. Cela confirme notre hypothèse de départ et nous pouvons conclure que ces méthodes sont bien adaptées à la détection d'évènements et moins à la détection de nouveauté qui apparaît doucement dans le temps.

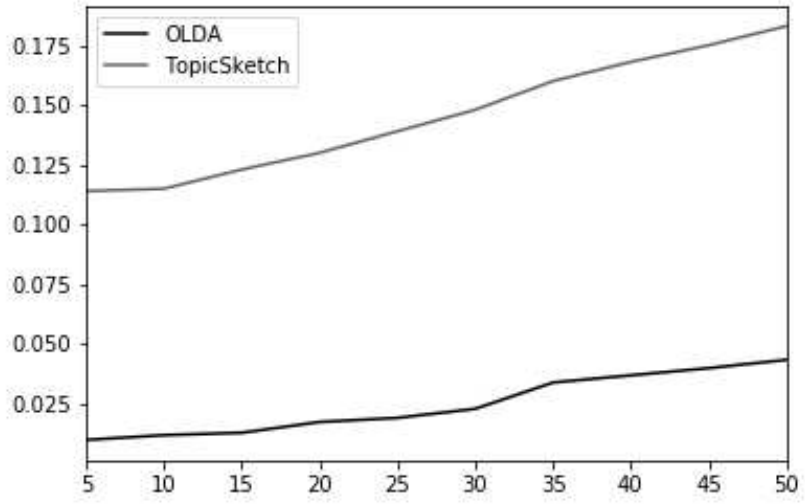


FIGURE 2.5: Évolution de la f-mesure par rapport au coefficient de pente d'apparition de la nouveauté.

Les tableaux 2.2 et 2.3 montrent les résultats généraux obtenus par chaque algorithme sur chaque scénario testé. Sur le tableau 2.2, nous voyons que, malgré les résultats assez faibles, c'est la méthode basée sur TopicSketch qui montre les meilleures performances. Nous voyons aussi que OLDA ne fonctionne pas pour détecter de nouveaux mots ou documents dans le cas de scénarios cycliques. Finalement, nous nous apercevons que le rappel est, généralement, plus faible, ce qui montre que peu de mots corrects sont effectivement détectés par les algorithmes. En termes de détection de documents, sur le tableau 2.3, les résultats sont plus divers, seules les méthodes basées sur TopicSketch et TF-IDF se différencient. Les

Résultats pour la tâche 1 : Détection des nouveaux mots									
Méthodes	Scénario 1			Scénario 2			Scénario 3		
	P	R	F	P	R	F	P	R	F
TF-IDF	0.073	0.005	0.010	0.085	0.004	0.009	0.089	0.006	0.012
BS	0.044	0.004	0.007	0.025	0.006	0.010	0.007	0.009	0.008
DF	0.186	0.005	0.010	0.179	0.004	0.008	0.205	0.006	0.013
OLDA	0	0	0	0	0	0	0	0	0
TopicSketch	**0.239**	**0.106**	**0.138**	**0.286**	**0.104**	**0.153**	**0.281**	**0.106**	**0.151**
Méthodes	Scénario 4			Scénario 5			Scénario 6		
	P	R	F	P	R	F	P	R	F
TF-IDF	0.247	0.006	0.012	0.400	0.003	0.007	0.496	0.009	0.017
BS	0.173	0.005	0.010	0.308	0.003	0.006	0.375	0.010	0.020
DF	0.330	0.005	0.011	0.504	0.003	0.006	0.594	0.008	0.017
OLDA	0.130	0.006	0.011	0.197	0.002	0.005	0.420	0.010	0.021
TopicSketch	**0.483**	**0.105**	**0.172**	**0.552**	**0.103**	**0.170**	**0.716**	**0.109**	**0.187**
Méthodes	Scénario 7			Scénario 8			Scénario 9		
	P	R	F	P	R	F	P	R	F
TF-IDF	0.456	0.008	0.015	0.466	0.005	0.011	0.499	0.010	0.021
BS	0.352	0.008	0.017	0.366	0.005	0.010	0.377	0.011	0.021
DF	0.543	0.007	0.015	0.604	0.006	0.012	0.621	0.010	0.020
OLDA	0.302	0.009	0.017	0.248	0.006	0.013	0.341	0.010	0.022
TopicSketch	**0.608**	**0.108**	**0.182**	**0.641**	**0.106**	**0.179**	**0.762**	**0.111**	**0.193**

TABLE 2.2: Résultats de Précision (P), Rappel (R) et F-Mesure (F) pour chaque algorithme évalué sur la tâche 1 pour 9 scénarios d'arrivée de la nouveauté.

algorithmes basés sur la recherche par rapport aux plus proches voisins (TF-IDF et DF) fonctionnent mal sur les scénarios cycliques ce qui confirme notre hypothèse de départ. Ce tableau montre que, même si les résultats sont globalement assez faibles, les algorithmes n'ont pas le même comportement sur tous les scénarios d'arrivée de la nouveauté.

2.5 Conclusion

Dans cette section, nous avons apporté un éclairage sur la définition de la nouveauté aussi bien au niveau général qu'au niveau textuel. Nous avons montré les différentes familles d'approches utilisées pour résoudre ce problème et mis en avant

Résultats pour la tâche 2 : détection de nouveaux documents.									
Méthodes	Scénario 1			Scénario 2			Scénario 3		
	P	R	F	P	R	F	P	R	F
TF-IDF	0	0	0	0.062	0.001	0.003	0.093	0.002	0.005
BS	0.028	0.005	0.008	0.083	0.013	0.023	**0.125**	0.018	0.032
DF	0	0	0	**0.193**	0.052	**0.071**	0	0	0
OLDA	0.029	0.005	0.008	0.084	0.014	0.024	0.109	0.018	0.029
TopicSketch	**0.051**	**0.036**	**0.040**	0.062	**0.056**	0.043	0.096	**0.078**	**0.083**
Méthodes	Scénario 4			Scénario 5			Scénario 6		
	P	R	F	P	R	F	P	R	F
TF-IDF	0.600	0.023	0.04	**0.825**	0.055	0.103	**0.856**	0.046	0.085
BS	0.048	0.005	0.010	0.102	0.011	0.019	0.258	0.027	0.049
DF	**0.638**	**0.026**	0.050	0.650	0.029	0.056	0.741	0.046	0.087
OLDA	0.048	0.005	0.010	0.077	0.011	0.019	0.108	0.027	0.043
TopicSketch	0.150	0.096	**0.114**	0.165	**0.123**	**0.139**	0.229	**0.155**	**0.183**
Méthodes	Scénario 7			Scénario 8			Scénario 9		
	P	R	F	P	R	F	P	R	F
TF-IDF	**0.818**	0.278	0.375	**0.843**	0.167	**0.259**	**0.852**	0.085	0.149
BS	0.604	**0.481**	**0.536**	0.326	0.133	0.189	0.170	0.034	0.057
DF	0.770	0.054	0.100	0.710	0.071	0.122	0.838	0.052	0.090
OLDA	0.104	0.181	0.132	0.069	0.080	0.070	0.061	0.045	0.038
TopicSketch	0.216	0.186	0.197	0.209	0.128	0.150	0.221	**0.142**	**0.166**

TABLE 2.3: Résultats de Précision (P), Rappel (R) et F-Mesure (F) pour chaque algorithme évalué sur la tâche 2 pour 9 scénarios d'arrivée de la nouveauté.

leurs avantages et leurs inconvénients. Nous avons fait la distinction entre différents termes comme "nouveauté", "évènement", "anomalies" et "outliers" afin de cadrer les tâches que nous voudrons résoudre par la suite. Nous avons montré que les travaux de la littérature n'utilisent pas forcément les mêmes méthodes d'évaluation et n'essaient pas de détecter la nouveauté au même niveau d'abstraction (mots/documents/thématiques).

En conséquence, cela ne permet pas d'avoir de comparaisons claires entre les méthodes existantes et rend difficile l'évaluation de nos méthodes. C'est pour cela que nous avons présenté un travail de comparaison de ces algorithmes. Nous avons fixé un cadre par rapport à plusieurs tâches que nous voulons évaluer avec des métriques d'évaluations associées. En testant ces algorithmes sur des scénarios d'arrivée de

la nouveauté sur des données simulées, nous avons pu contrôler l'ensemble des paramètres de génération des données et ainsi observer leur effet sur les résultats. Bien que les résultats soient globalement faibles, il faut rappeler que nous cherchions à détecter un petit ensemble de mots et de documents parmi de grands ensembles de données. Ce travail nous a surtout permis de mettre en lumière des approches intéressantes pour continuer nos travaux. Par exemple, pour l'ensemble des tâches, nous avons remarqué que l'algorithme TopicSketch [2] était celui présentant les meilleurs résultats. Enfin, la méthode TF-IDF [1] est une méthode utilisant des propriétés simples du texte. Nous l'utiliserons comme baseline dans les prochains chapitres, car elle présente l'avantage d'être simple à mettre en place et a des performances intéressantes. Dans la suite de ces travaux, en plus de résoudre des tâches de détection de nouveaux mots ou de nouveaux documents, nous nous intéressons aussi à l'aspect temporel de la détection : nous voulons réduire le retard entre l'apparition d'une nouveauté et sa détection.

Chapitre 3

Détection des éléments nouveaux

Dans le chapitre 2, nous avons rappelé que, dans notre cas d'application, la nouveauté se matérialise sous deux formes : les nouveautés de structure et les nouveautés de volume. Les nouveautés de structure apparaissent lorsque de nouveaux éléments, jusque là inconnus, arrivent dans nos données. Cela peut être des mots qui changent de sens ou des thématiques qui émergent. Dans ce chapitre, nous présentons nos travaux autour des nouveautés de structures et donc sur la détection de ces éléments nouveaux.

3.1 Introduction

Lors de l'apparition de nouvelles technologies, comme l'informatique ou internet, de nouveaux sujets de conversation apparaissent. Avant les années 1950, le mot "ordinateur" n'existait pas. Le mot "internet" a été créé dans les années 1970. Le mot "souris" ne pouvait désigner qu'un animal avant la création de l'objet informatique dans les années 1960. L'ensemble de ces mots qui, soit n'existaient pas, soit évoluaient dans des contextes différents ont connu des changements liés à des phénomènes nouveaux dans le monde réel.

Dans l'ensemble des données textuelles exploitables dans le monde (articles de presse, réseaux sociaux, articles scientifiques) nous retrouvons des dynamiques particulières : des mots apparaissent, ils changent de sens, ils apparaissent dans de nouveaux contextes au point de former des thématiques identifiables, ces thématiques prennent de l'ampleur ou disparaissent. Chaque nouveau document écrit peut être l'occasion pour le langage d'évoluer. Il peut être intéressant de modéliser et d'analyser cette évolution.

Pour une entreprise comme EDF, il est nécessaire d'analyser ces mouvements pour comprendre ce qui change dans l'esprit des clients et dans les retours qu'ils font sur l'entreprise soit sur les réseaux sociaux soit directement par courriel. Cette compréhension permet à l'entreprise d'anticiper la mise en place de réponses marketings ou industriels afin d'améliorer ses services et son rapport avec les clients.

Dans ce chapitre, nous nous intéressons en particulier à la détection de nouveaux éléments, c'est-à-dire à des mots qui changent de sens ou des thématiques qui émergent dans nos données. Nous nous concentrons d'abord sur les différents espaces de représentations qui existent en exposant leurs avantages et leurs inconvénients. Nous présentons nos travaux qui portent sur la détection de nouveautés à l'aide de modèles thématiques. Enfin, nous détaillons nos travaux sur l'utilisation de modèles de plongements appliqués à la détection de la nouveauté.

3.2 Modèles de représentations

Nous avons parlé, en introduction, de "mouvement" au niveau du langage. La notion de mouvement sous-entend la présence d'un espace de représentations et de points qui évoluent dans cet espace. Afin de représenter des données textuelles dans un espace, nous devons les transformer en données numériques. Nous avons succinctement présenté, dans la section 1.2, certaines méthodes de pondération des mots

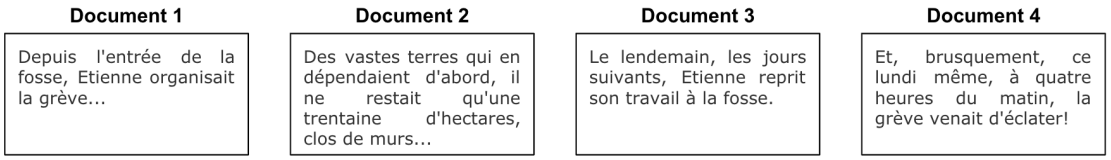


FIGURE 3.1: Exemple d'un corpus de documents

qui permettent de représenter des mots ou des documents dans des espaces. Nous approfondissons ici ces méthodes.

3.2.1 Pondérations dans l'espace des mots

Pour représenter des mots ou des documents sous forme de données numériques, il est classique de se baser sur des mesures de pondérations en très grande dimension. C'est le cas par exemple des mesures de type TFxIDF et PPMI que nous avons abordées en introduction et que nous allons approfondir ici. Ces mesures se basent sur des matrices dont la taille est égale à la taille du vocabulaire. La mesure la plus basique pour représenter des données textuelles dans l'espace des mots consiste à compter les mots présents dans les documents et de construire une matrice Documents-Termes. Nous avons présenté le processus de construction de cette matrice dans la section 1.2. Supposons un ensemble de documents, présenté sur la figure 3.1.

Après des étapes de nettoyage des données et indexation des termes, la méthode de représentation basique correspond à un simple comptage dans une matrice documents-termes (figure 3.2).

L'une des méthodes de représentations les plus classiques dans le domaine du TAL consiste à se baser sur la mesure de pondération TFxIDF [86]. Nous avons explicité dans la section 1.2, les formules mathématiques qui permettent de la calculer. Une matrice de représentation en TFxIDF permet de pondérer l'importance des mots dans chaque document par rapport à sa fréquence d'apparition dans les autres

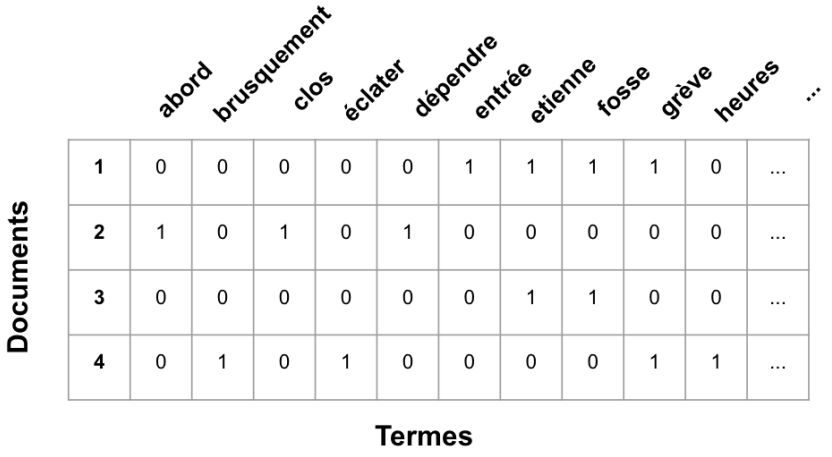


FIGURE 3.2: Exemple d'une matrice documents-termes

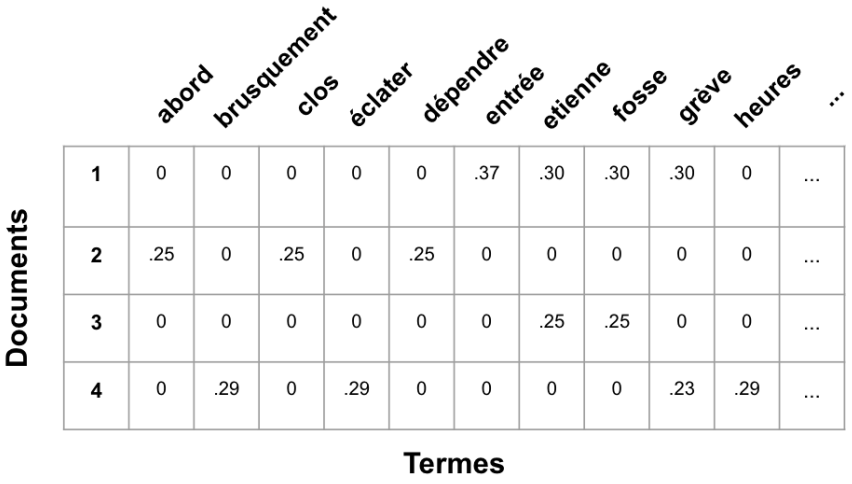


FIGURE 3.3: Exemple d'une matrice TFxIDF

documents. Notre matrice documents-termes est transformée en matrice TFxIDF (figure 3.3).

Une autre méthode nommée **Okapi BM25 (*Best Matching*)**[87] est couramment utilisée dans le domaine du TAL. Là où TFxIDF se base seulement sur la fréquence du terme et l'inverse de la fréquence du document, Okapi BM25 ajoute deux hyperparamètres pour la pondération. Sa formule est la suivante :

$$BM25(d,w) = IDF(w,d)\frac{f(w,d).(k_1+1)}{f(w)+k_1*(1-b+b*|D|/avgdl)}$$

où $f(w, d)$ est la fréquence du terme w dans le document d, $|D|$ est le nombre de termes dans le document d, $avgdl$ est la taille moyenne des documents du corpus et b et k_1 sont des hyperparamètres à définir. Le paramètre k_1 permet d'atténuer l'effet des fréquences de mots élevés sur le score TF. C'est ce qu'on appelle la caractéristique de saturation. Le paramètre b (entre 0 et 1) contrôle l'importance de la taille du document par rapport à la taille moyenne du corpus.

3.2.2 Pondérations dans un espace compressé

Les schémas de pondérations dans l'espace des mots demandent de manipuler des matrices très grandes dont la majorité des dimensions apportent peu d'informations. Il est courant de compresser ces matrices et donc de travailler avec des représentations de mots en plus petites dimensions. Cette volonté d'utilisation d'espaces compressés est à la base des modèles de plongement de mots ou *embeddings*.

Ces techniques permettent de représenter les mots dans des espaces vectoriels où les mots ayant un sens similaire sont proches. Les techniques présentées précédemment dans la section 3.2.1 représentent aussi les mots sous forme de vecteurs, mais elles partent du principe que les mots sont indépendants les uns des autres. En réalité, la présence d'un mot dans un document, ou dans une phrase, dépends du contexte autour de lui. Les modèles de plongements tels que [71, 74–77] introduisent cette notion de dépendance. Cette hypothèse découle de l'hypothèse distributionnelle de Harris [88] selon laquelle des mots sémantiquement proches ont tendance à partager des contextes similaires.

La famille de modèles de plongement la plus classique est Word2Vec [4]. C'est une famille, car elle comporte deux algorithmes différents : le *Continuous Bag of Words* (CBOW) et le *Skip-Gram with Negative Sampling* (SGNS). L'idée principale derrière cette famille est de développer un modèle d'apprentissage qui permet de prédire l'apparition d'un mot d'après les mots qui l'entourent (le contexte). Cet

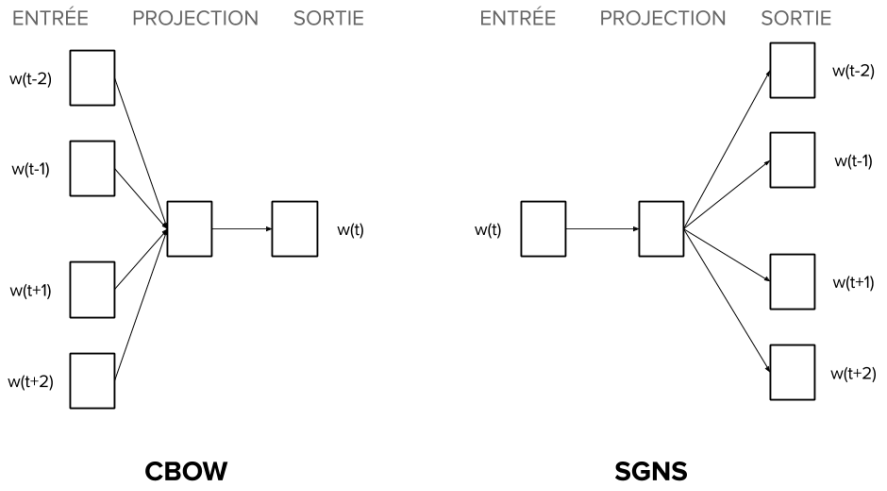


FIGURE 3.4: Différences entre les approches CBOW et SGNS.

apprentissage est effectué à l'aide d'un réseau de neurones et les poids appris par ce réseau constituent les vecteurs de représentations de mots.

Sur un ensemble de documents, le modèle s'arrête sur les mots de chaque document pour soit utiliser le mot courant w_i pour prédire ses voisins (le contexte) : c'est l'approche SGNS ; soit utiliser le contexte pour prédire le mot courant w_i : c'est l'approche CBOW. Ces différences sont illustrées sur la figure 3.4. Les vecteurs de mots correspondent à la partie "Projection" sur cette figure : ce sont les poids appris par le réseau.

Dans ce type d'espace de représentation, les mots ayant un sens similaire sont proches. Cette famille de représentation vectorielle de mots permet de généraliser des transformations géométriques sur les vecteurs de mots pour, par exemple, passer d'un masculin à un féminin ou d'un verbe à son participe. Par exemple, il est possible d'effectuer des calculs vectoriels pour trouver le féminin d'un mot, ou la capitale d'un pays. Nous pouvons avoir des opérations du type $v_{roi} - v_{homme} + v_{femme} = v_{reine}$ ou $v_{pays} + X = v_{capitale}$ où X est un vecteur qui permet de faire cette transformation pour n'importe quel pays. Ce type de transformation est illustré dans la figure 3.5.

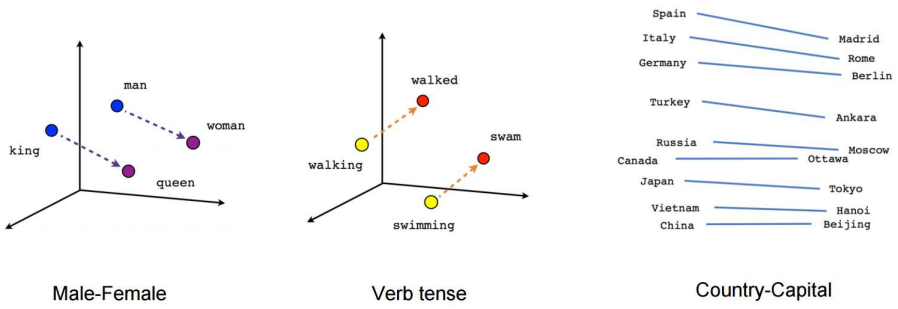


FIGURE 3.5: Transformations géométriques possibles dans des espaces vectoriels de type Word2Vec (Source : https://www.tensorflow.org/tutorials/text/word2vec)

D'autres modèles de plongements existent dans la littérature, tels que Glove [75] et FastText [71]. Glove [75] est un algorithme qui construit des vecteurs de mots en grandes dimensions qui prennent en compte les statistiques globales des mots dans le corpus en ajoutant une matrice de co occurrence des mots à l'apprentissage. FastText [71] est une extension de Word2Vec qui considère les mots comme un assemblage de n-grammes de caractères. L'idée générale pour l'apprentissage de représentations est la même que pour Word2Vec. Les vecteurs de mots sont la moyenne des vecteurs des n-grammes qui les composent. Ce type de modèle a l'avantage d'être moins sensible aux mots inconnus, car ceux-ci peuvent toujours être représentés par leurs n-grammes. Dans la suite de ce manuscrit, nous nous basons sur des représentations effectuées avec un modèle SGNS. Les approches Word2Vec sont facilement adaptables à un environnement dynamique où nous devons faire évoluer l'espace en ajoutant de nouveaux documents.

Le développement de ces modèles de plongements de mots basés sur l'hypothèse de Harris est souvent effectué sur des jeux de données en langue Indo-Européennes où cette hypothèse distributionnelle est effectivement vérifiée. Enfin, ce type de modèle nécessite de grandes quantités données pour représenter de manière stable les mots dans un espace. En effet, chaque nouvelle observation d'un mot dans un contexte précis a un effet important sur ses coordonnées dans l'espace. Il faut donc de nombreuses observations différentes pour compenser cette instabilité et représenter

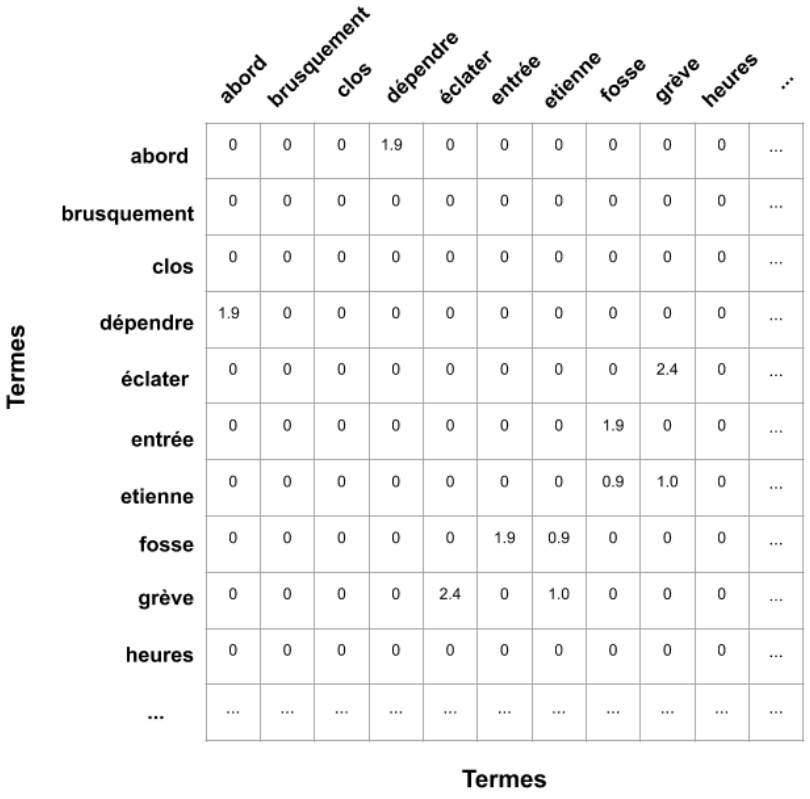


FIGURE 3.6: Exemple d'une matrice PPMI

au mieux le sens des mots.

La mesure de PPMI (*Positive Pointwise Mutual Information*) que nous avons étudié en introduction 1.2 et qui pondère les mots en fonction de leur fréquence de co-occurrence permet d'obtenir des matrices carrées ou chaque vecteur de mot à une taille égale à la taille du vocabulaire. Un exemple de matrice obtenue est présenté sur la figure 3.6.

Lorsque nous compressons ce type de matrice via un algorithme de *Singular Value Decomposition* (SVD), nous obtenons des vecteurs de mots plus stables, c'est-à-dire où chaque nouvelle observation d'un mot dans un contexte a un impact moindre sur ses coordonnées dans l'espace [89].

La SVD est une méthode qui permet de décomposer une matrice de base A de taille $n_V * n_V$ en trois autres matrices, dont une contient les vecteurs propres. Nous

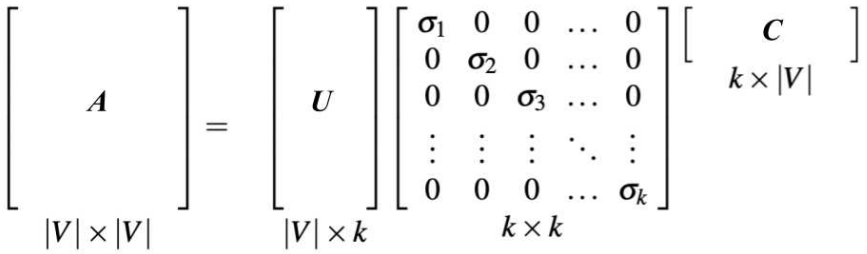


FIGURE 3.7: Fonctionnement d'une SVD tronquée sur une matrice PPMI

pouvons sélectionner k vecteurs propres qui formeront les k-dimensions de notre matrice finale, c'est ce que l'on appelle un procédé de SVD tronquée. Lorsque nous appliquons une SVD tronquée sur une matrice termes-termes, les vecteurs de mots de taille K sont contenus dans la matrice U. La décomposition de cette matrice de base est illustrée sur la figure 3.7. Il est courant de combiner les matrices U et V et de les normaliser pour obtenir de meilleures représentations des mots.

L'algorithme de SVD est couramment utilisé dans le domaine du TAL. En effet, il est à la base de l'algorithme LSA (*Latent Semantic Analysis*) [81] lorsqu'il est appliqué à une matrice Documents-Termes et permet d'introduire la notion de thématiques dans l'analyse.

3.2.3 Pondérations dans l'espace des thématiques

Dans la section 1.2, nous avons parlé de modèles thématiques probabilistes en évoquant rapidement le modèle *Latent Dirichlet Allocation* (LDA). Nous le présentons ici en détail avec certaines de ses extensions temporelles.

LDA est un modèle probabiliste utilisé pour décrire un corpus de n_D documents associés à un vocabulaire de taille n_V. Dans ce modèle, des variables latentes sont utilisées pour représenter des thématiques présentes dans chaque document. LDA

utilise le processus génératif suivant qui permet de simuler la création d'un document :

Algorithm 1: Latent Dirichlet Allocation

Génère, pour chaque thématique k, $1 \leq k \leq K$, une distribution sur les termes : $\phi_k \sim Dir(\beta)$, où ϕ_k et β sont des vecteurs de dimensions n_V;

for *chaque document* d **do**

  Tirer aléatoirement une distribution sur les thématiques $\theta^d \sim Dir(\alpha)$ où θ^d et α sont des vecteurs de dimensions K;

  for *chaque terme* n, $1 \leq n \leq N$ *dans* d **do**

    Choisi une thématique : $z_n^d \sim mult(1, \theta^d)$;

    Choisi un terme w_n^d de la thématique z_n^d avec la probabilité $P(w_n^d = v | z_n^d = k) = \phi_{k,v}$;

  end

end

Une thématique est donc décrite par sa distribution sur le vocabulaire. Afin de visualiser le sens d'une thématique, nous observons généralement les mots ses plus probables. Par exemple, la figure 3.8 montre les 30 mots les plus probables pour une thématique construite avec un modèle LDA sur le jeu de donnée des courriels EDF. En observant ces mots, nous pouvons induire que cette thématique concerne le paiement des factures reçues par courrier. Au niveau de la complexité, un modèle LDA s'exécute en $\mathcal{O}(n_D N K)$ avec n_D le nombre de documents dans le corpus, N le nombre de mots dans les documents, et K le nombre de thématiques.

Plusieurs extensions du modèle LDA ont été étudiées dans la littérature, notamment ses extensions temporelles.

Dans DTM [6], les auteurs développent un modèle qui capture l'évolution des thématiques dans un corpus d'articles scientifique provenant du journal Science. Ils découpent le corpus en fenêtres temporelles et, afin de construire un modèle LDA à

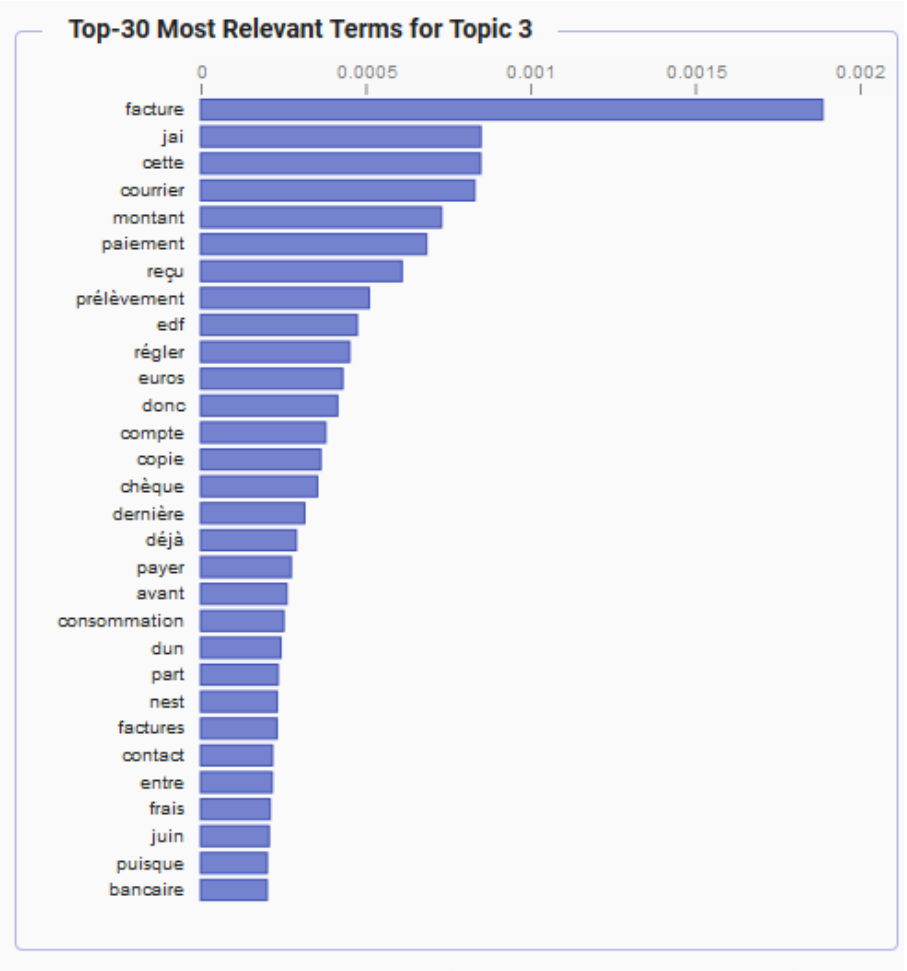


FIGURE 3.8: Visualisation des 30 mots les plus probables d'une thématique LDA construite sur le jeu de données EDF.

un temps t, ils utilisent les paramètres α et β du modèle LDA au temps $t-1$. DTM capture les dépendances entre les distributions de thématiques dans les documents et de termes dans les thématiques. Ces dépendances sont capturées à travers des distributions gaussiennes construites au temps d'avant. Cela force les nouvelles valeurs des paramètres à être distribuées autour des valeurs précédemment observées. Dans on-line LDA (OLDA) [90], les auteurs présentent une version en ligne de LDA qui capture automatiquement les thématiques et leur évolution dans le temps. Elle permet de construire un modèle à jour (mélange de thématiques par documents et mélange de termes par thématique) à chaque fois qu'un document apparaît. Ceci est fait grâce à une matrice évolutive dont les colonnes sont les distributions de

termes-thématiques générées par les documents reçus dans la fenêtre de temps. En combinant cette matrice à un vecteur représentant un poids pour chaque fenêtre de temps, OLDA détermine le paramètre β du modèle. En calculant les β de cette manière, les distributions de thématiques sont liées dans des modèles consécutifs. Un autre modèle dérivé de OLDA est présenté dans [84]. Les auteurs ajoutent la possibilité de mettre à jour le vocabulaire à chaque instant. Dans TM-LDA [91], les auteurs veulent apprendre les paramètres de transition entre les thématiques afin de minimiser l'erreur de prédiction des thématiques dans les documents suivants. Ce modèle est construit pour apprendre les paramètres des transitions de thématiques à partir d'une séquence de document organisée temporellement et de prédire la distribution future des thématiques dans les nouveaux documents. Dans ST-LDA-D — C [92] les auteurs modélisent la dépendance entre les thématiques avec la copule de Francks. Les copules sont des outils mathématiques utilisés pour modéliser des dépendances entre variables aléatoires. Pour lier des distributions de thématiques en t et $t-1$, on considère les vecteurs associés à ces distributions et on les lie coordonnée par coordonnée.

Tous ces modèles introduisent les concepts de thématiques dans l'analyse à partir de l'observation des mots et permettent de suivre leurs évolutions dans le temps. Cependant, ils ne prennent pas en compte le contexte autour des mots dans les documents, ce qui aiderait à modéliser leur sens sémantique.

3.3 Détection de nouveautés à l'aide de modèles thématiques.

Comme nous l'avons démontré dans les sections précédentes, les modèles thématiques probabilistes permettent de découvrir des thématiques automatiquement dans les données et de les faire évoluer dans le temps. Notre but, dans ce chapitre, est de

détecter des structures qui émergent progressivement dans les données. Ces structures peuvent être des mots, des groupes de mots ou bien des thématiques. Afin d'étudier l'apport des modèles thématiques pour ce type de tâche, nous expérimentons à l'aide de modèles de type LDA. Il existe des modèles tels que DTM [6] et OLDA [90] où l'aspect temporel est déjà pris en compte, mais ils sont construits pour observer des changements sur de très longues durées (besoin de beaucoup d'observations) ou très courtes (sensible au bruit).

3.3.1 Méthode générique

Pour étudier l'utilité d'un modèle LDA pour détecter la nouveauté, nous nous plaçons dans une configuration simple, c'est-à-dire avec seulement deux instants : un historique, contenant ce que l'on connaît et un contexte, qui contient certains documents considérés comme nouveaux. Cette formalisation est illustrée sur la figure 3.9.

Nous observons un certain nombre de documents $\mathcal{D} = \{(d_i, t_i), i \in \mathbb{R}\}$ avec i l'indice du document, d_i le texte du document et t_i sa date d'apparition. La date t_c correspond à la date de séparation entre notre historique et notre contexte moment où nous observons les documents arriver. w_h et w_c correspondent à la taille des fenêtres temporelles définissant l'historique et le contexte que l'on prend en compte. L'historique correspond à un sous-ensemble de documents $D_{hist} \subset \mathcal{D}$ où $D_{hist} = \{(d_i, t_i) \in \mathcal{D} / t_c - w_h \leq t_i < t_c\}$. Le contexte correspond à un ensemble de documents $D_{cont} \subset \mathcal{D}$ où $D_{cont} = \{(d_i, t_i) | t_c < t_i \leq t_c + w_c\}$.

3.3.2 Observation des distances

Notre but est de détecter des thématiques émergentes dans un corpus. Nous explorons l'utilité des modèles thématiques en nous plaçant dans un environnement à

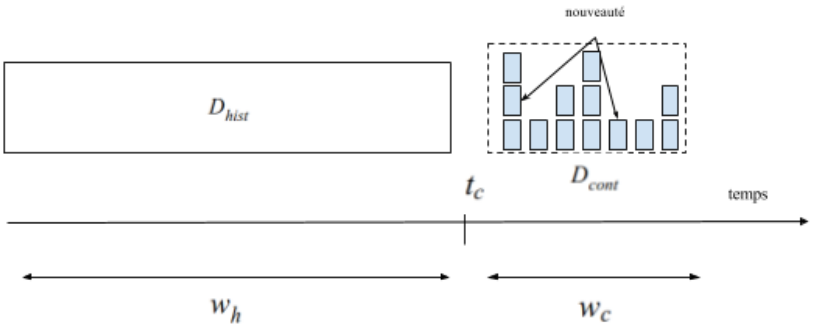


FIGURE 3.9: Modèle générique avec l'ensemble d'historique à gauche et l'ensemble de contexte à droite qui contient certains documents que nous devons détecter comme nouveau.

deux instants et cette émergence est donc matérialisée seulement par la proportion de documents nouveaux présents dans le contexte.

Pour résoudre cette tâche, notre hypothèse principale consiste à supposer que les documents nouveaux devraient anormalement être loin, en termes de distance, de leurs plus proches voisins. Cela correspond à la famille d'algorithme de détection de la nouveauté présentée dans la section 2.2.2. Nous associons, à chaque document du contexte, un score de nouveauté qui dépend des documents de l'historique. Plus le score de nouveauté est grand, plus le document peut être considéré comme nouveau. Nous pouvons utiliser les différentes définitions de distance présentées dans la section 2.2.2 pour la calculer entre nos différentes entités. Une entité peut représenter un document (les termes qui le composent) ou une thématique (les termes les plus probables). Nous expérimentons avec la dissimilarité cosinus afin d'obtenir une matrice de distance comme illustrée sur la figure 3.10.

Une fois cette matrice obtenue, nous voulons agréger ces distances pour calculer notre score de nouveauté. Nous calculons la moyenne de la distance par rapport aux plus proches voisins d'une entité du contexte. Nous détectons les documents **anormalement** distants les uns des autres. Nous devons donc fixer un seuil qui

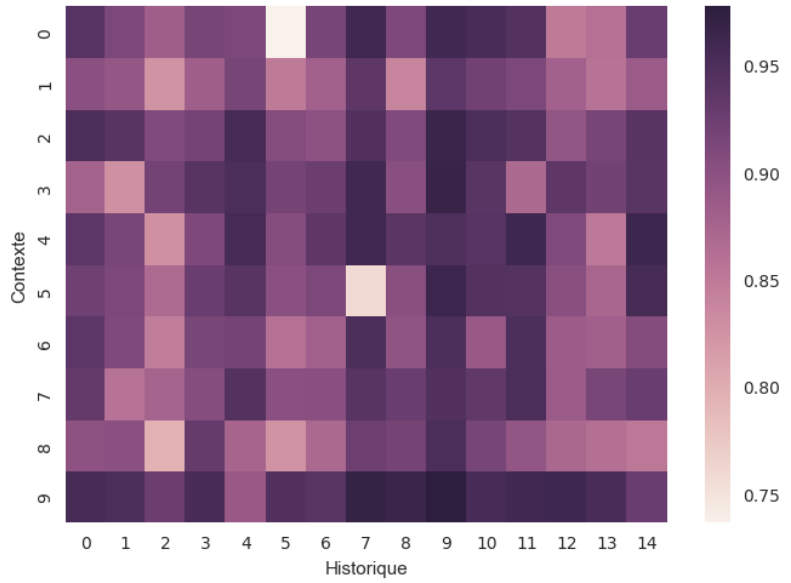


FIGURE 3.10: Représentation de la matrice de distance.

permet de définir cette notion d'anormalité. Nous fixons ce seuil par rapport aux valeurs observées autour de la moyenne + l'écart type. Cela nous permet de limiter le nombre de documents détectés et d'analyser seulement les valeurs extrêmes.

Moyenne des k plus proches voisins : $score(d) = \frac{1}{|V_k|} \sum_{\substack{i=0 \\ d'_i \in J}}^{k} diss(d, d'_i)$, et J est l'ensemble de taille k qui minimise la fonction $diss(d, d'_i)$

3.3.3 Modélisation

Afin d'explorer l'utilité des modèles thématiques pour la détection de la nouveauté, nous définissons trois approches. La première approche nous sert de baseline et représente les documents seulement par rapport aux mots et les approches thématiques ne sont pas analysés. Ensuite nous construisons deux modèles utilisant les thématiques pour représenter l'historique ou l'ensemble des documents.

Modèle de comparaison documents-documents La Figure 3.11 représente une comparaison des termes présents dans les documents de l'historique et du

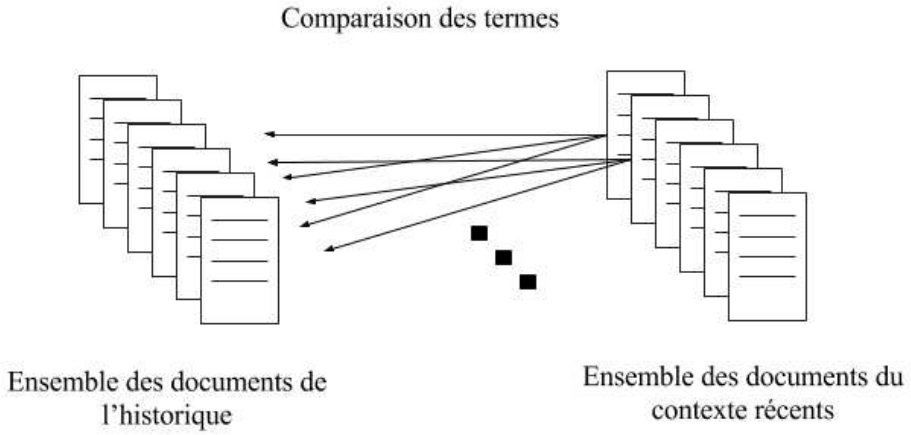


FIGURE 3.11: Comparaison des documents deux à deux.

contexte récent. Le modèle consiste à calculer une distance entre tous les documents deux à deux : à chaque document arrivant dans la fenêtre de contexte, nous le comparons avec tous les documents de l'historique. Les documents sont représentés par leur vecteur TFxIDF correspondant aux termes. Pour calculer la distance entre documents, il est classique d'utiliser la dissimilarité cosinus. Nous agrégeons les distances pour calculer le score de nouveauté correspondant à la moyenne des distances par rapport aux plus proches voisins. Nous classons l'ensemble des documents du contexte par rapport à leur score de nouveauté. Au niveau de la complexité, nous comparons tous les documents un à un, notre méthode s'exécute donc en $\mathcal{O}(n_D^2)$.

Modèle de comparaison thématiques-documents La Figure 3.12 introduit la notion de thématique dans le modèle. Ce dernier permet de comparer les termes des documents du contexte avec les termes les plus probables des thématiques de l'historique. Une thématique est décrite par rapport à ses termes les plus probables. Nous générons des documents "résumés" des thématiques grâce aux 100 mots les plus probables de chacune d'elles. Nous pouvons donc comparer les documents avec

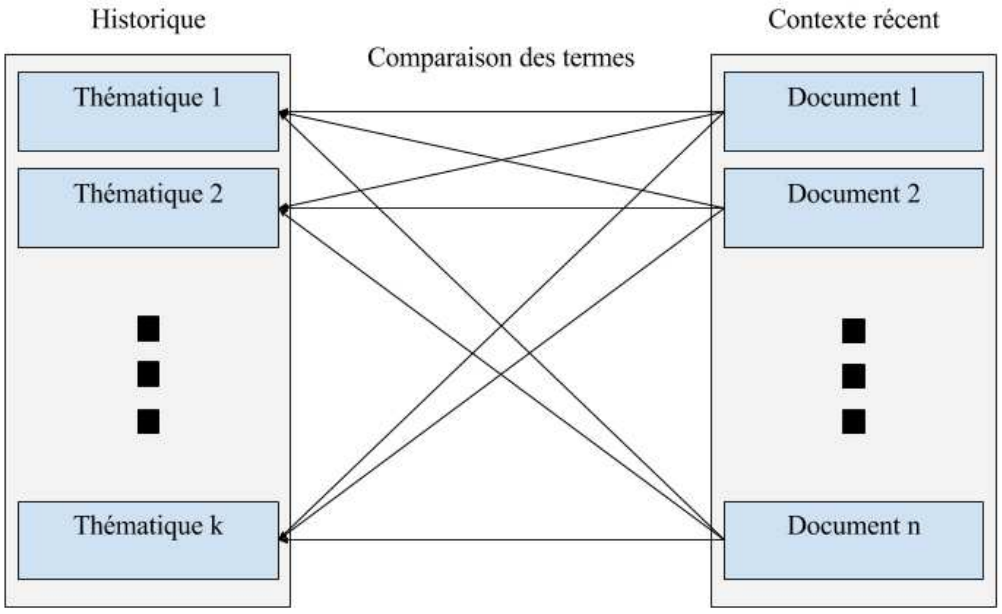


FIGURE 3.12: Comparaison des documents avec les thématiques de l'historique.

les thématiques dans un même espace. Pour ce modèle nous utilisons la dissimilarité cosinus 2.2 en prenant, pour score, la probabilité des termes dans la thématique avec laquelle on compare. Ce modèle détermine rapidement si un document fait partie des données "normales" ou, si il se trouve loin des thématiques de l'historique, de la nouveauté. Au niveau de la complexité, nous avons vu que la création des thématiques se fait en, $\mathcal{O}(n_D NK)$ mais celles-ci peuvent être calculées une fois avant l'arrivée de nouveaux documents. Pour ce modèle de comparaison, nous avons donc une complexité en $\mathcal{O}(n_D K)$ avec n_D le nombre de documents arrivant dans le contexte et K le nombre de thématiques.

Modèle de comparaison thématiques-thématiques La Figure 3.13 représente une variante du modèle précédent dans le sens où, au lieu de comparer directement les documents du contexte avec les thématiques de l'historique, nous allons

construire des thématiques sur ces documents puis déterminer si une ou plusieurs thématiques peuvent être considérées comme nouvelles (étape (a)). Pour comparer les thématiques du contexte avec celles de l'historique, nous calculons la distance entre les termes les plus probables de chaque thématique. En agrégeant les résultats de la matrice de distance (moyenne aux k plus proches voisins), nous pouvons classer les thématiques par ordre de nouveauté. Afin de sélectionner les thématiques qui pourraient nous intéresser, nous avons fixé un seuil : $threshold = mean(score(Z_d)) + std(score(Z_d))$. Ce seuil permet de traduire la notion de thématique *anormalement* distante des thématiques précédentes.

Une fois les thématiques nouvelles identifiées, nous utilisons les documents les plus probables de celles-ci et comparons leurs termes avec les thématiques de l'historique (étape (b)). Cela permet de déterminer quels sont les documents responsables de la nouveauté de la thématique. Au niveau de la complexité, nous devons attendre qu'un certain nombre de documents arrivent dans le contexte pour former des thématiques LDA en $\mathcal{O}(n_D NK)$. Une fois ces thématiques construites, la comparaison s'effectue entre les thématiques donc nous avons $\mathcal{O}(K^2)$ avec $K << n_D$.

3.3.4 Expérimentations et résultats

Méthodologie Afin de mesurer l'apport des méthodes de modélisation thématiques pour la détection de la nouveauté, nous testons nos approches sur un jeu de données où les documents sont classés par catégorie. Cela nous permet de simuler l'arrivée d'une catégorie qui fait office de nouveauté. Notre processus est le suivant : nous retirons, de notre jeu de données, tous les documents relatifs à une catégorie. Ensuite, nous séparons le jeu de données en deux avec d'un côté, les documents de l'historique (notre base de connaissances) et, de l'autre, les documents du contexte. Enfin, nous ajoutons, dans le contexte, quelques documents de la thématique que nous avons retirée. Ces documents font office de nouveauté et donc de référence pour notre expérimentation.

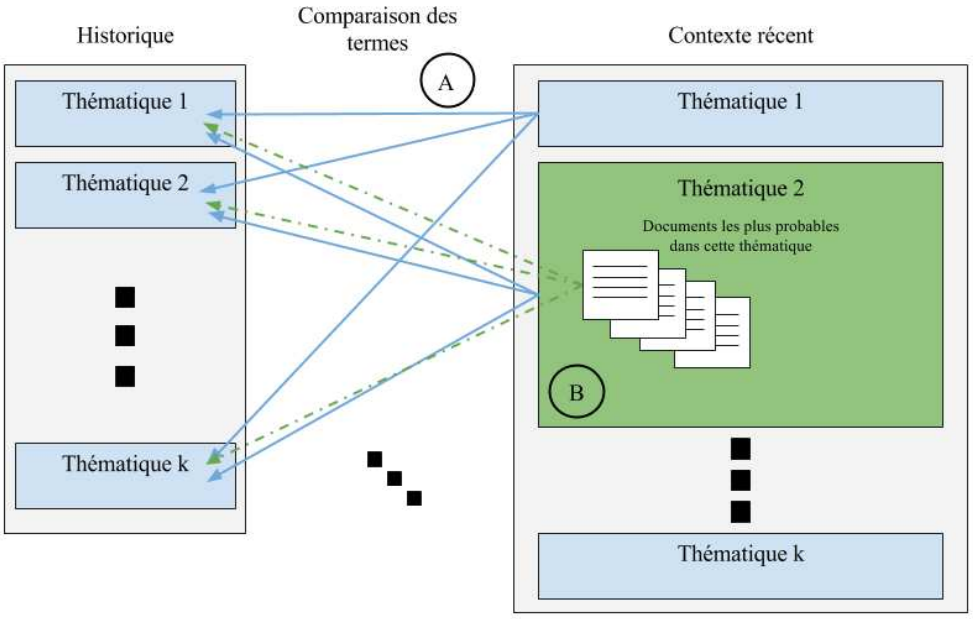


FIGURE 3.13: Comparaison des documents des thématiques nouvelles avec les thématiques de l'historique

Catégories	Nombre d'articles
database	2475
datamining	1133
medical	221
theory	2252
visu	2261

TABLE 3.1: Répartition des documents par catégories

Nous testons nos approches sur un jeu de données de résumés d'articles scientifiques publiés entre 1990 et 2005. Ces articles sont classés selon cinq catégories : *theory, database, datamining, visu* et *medical*. La répartition des articles dans les catégories est présentée sur le tableau 3.1.

Notre historique et notre contexte sont constitués des articles publiés, respectivement, entre 1990 et 1999 et entre 2000 et 2005. Nous testons 5 scénarios différents où un scénario correspond à une catégorie retirée. Aussi, la taille de notre nouveauté (donc le nombre de documents de la catégorie ajouté dans le contexte) varie. Nous

testons en ajoutant 1,5,20 et 100 documents afin de vérifier l'impact de la quantité sur la performance de la détection. Les ensembles d'historiques et de contexte sont composés d'environ 4000 documents chacun. Nous répétons l'opération 100 fois pour avoir des documents différents à détecter et donc des résultats plus stables.

Résultats Les résultats que nous présenterons dans cette partie sont composés de deux parties. Dans un premier temps, nous allons observer des mesures d'AUC (*Area Under Curve*) moyennes qui permettent de quantifier la qualité de détection de nos systèmes. Pour rappel, l'AUC mesure l'aire sous une courbe ROC, c'est une mesure très utilisée dans le domaine de la recherche d'information et nous pouvons l'utiliser, car nous avons ramené notre problème à un problème de classification binaire. Dans un second temps, nous présenterons la précision à 100 de nos systèmes. Cette mesure nous permet de nous concentrer uniquement sur le début de classement de nouveauté : c'est-à-dire les documents qui sont fortement susceptibles d'être présentés à des experts métiers. Il est important de noter que notre méthodologie permet de détecter de nouvelles catégories, mais ne mesure pas la nouveauté apparaissant au sein des catégories : par exemple, entre 1990 et 2005, nous pouvons imaginer que les articles de recherche sur le thème des bases de données ont beaucoup évolué et n'utilisent pas forcément les mêmes termes. Notre système n'intègre pas ces changements et cela peut expliquer les scores de détection assez faibles semblables à ce nous avions dans le chapitre 2.

Au niveau des mesures d'AUC, il est intéressant de voir qu'à partir de seulement 5 nouveaux documents introduits dans l'ensemble de contexte, nous observons des différences dans les tendances de détection entre les modèles. En effet, dans le tableau 3.2, quand nous ajoutons 5 nouveaux documents, nous voyons qu'il est plus difficile de détecter des catégories comme *datamining* et *medical*. Au contraire, les

catégories *theory* et *visu* sont plus faciles à détecter ($AUC > 0.7$). Lorsque nous utilisons des modèles thématiques, que ce soit pour résumer l'historique (tableau 3.3) ou pour le contexte (tableau 3.4), nous remarquons que les tendances s'inversent et que la catégorie *medical* est plus facile à détecter. Bien sûr, les catégories *database* et *datamining* restent assez difficiles à détecter. Cela peut s'expliquer par le fait que ces deux catégories partagent beaucoup de termes en communs et que les thématiques LDA ne font pas la différence entre ces deux ensembles. Au contraire, il semble que les catégories *theory* et *medical* utilisent des termes plus spécifiques qui permettent d'avoir des thématiques plus facilement identifiables. Enfin il est intéressant de noter que, dans le troisième modèle (tableau 3.4), la première étape identifie bien les nouvelles thématiques, car nous retrouvons bien les nouveaux documents insérés dans leur liste des 100 documents les plus probables.

# de documents	1	5	20	100
database	0.81±0.05	0.68±0.03	0.67±0.03	0.66±0.04
datamining	0.47±0.06	0.59±0.03	0.54±0.03	0.57±0.04
medical	0.71±0.04	0.61±0.03	0.62±0.04	0.66±0.02
theory	0.73±0.06	0.82±0.04	0.80±0.03	0.80±0.02
visu	0.67±0.03	0.75±0.04	0.72±0.04	0.71±0.03

TABLE 3.2: AUC moyennes du modèle de comparaison documents-documents

# de documents	1	5	20	100
database	0.73±0.04	0.68±0.03	0.66±0.04	0.65±0.02
datamining	0.33±0.05	0.47±0.03	0.43±0.04	0.44±0.04
medical	0.74±0.01	0.73±0.02	0.74±0.02	0.74±0.01
theory	0.75±0.03	0.75±0.02	0.76±0.02	0.75±0.02
visu	0.66±0.03	0.61±0.03	0.60±0.04	0.60±0.03

TABLE 3.3: AUC moyennes du modèle de comparaison thématiques-documents

Les résultats présentés dans le tableau 3.5 montrent le nombre moyen de documents nouveaux que nous retrouvons dans les 100 premiers documents classés par score de nouveauté (2.5% du classement). C'est, en quelques sortes, un zoom sur le début de la courbe ROC. Ces 100 premiers documents sont destinés à être présentés à des

# de documents	1	5	20	100
database	0.43±0.06	0.64±0.04	0.73±0.03	0.60±0.02
datamining	0.25±0.07	0.68±0.04	0.53±0.03	0.55±0.03
medical	0.62±0.06	0.71±0.02	0.69±0.03	0.69±0.03
theory	0.33±0.07	0.85±0.04	0.81±0.02	0.83±0.02
visu	0.63±0.05	0.57±0.04	0.53±0.04	0.65±0.03

TABLE 3.4: AUC moyennes du modèle de comparaison thématiques-thématiques

experts métiers pour interprétation.

La première colonne montre le niveau de détection lorsque nous comparons deux à deux les documents du contexte et de l'historique par rapport aux termes qu'ils utilisent. Nous voyons que les scores sont très faibles pour les catégories *datamining* et *medical* : c'est-à-dire que les documents nouveaux ne sont pas les plus éloignés, en termes de distance, de leurs plus proches voisins. Pour ce modèle, qui compare les documents deux à deux, la quantité de nouveauté introduite n'a pas d'effet sur la détection. En observant les résultats de cette première colonne, nous voyons que près de 17% des nouveaux documents correspondant à la catégorie *theory* sont dans les documents les mieux classés. Ce score est de 10 % pour la catégorie *database* et moins de 5% pour les autres.

La deuxième colonne présente les mêmes résultats une fois que nous avons résumé nos documents de l'historique sous forme de thématiques LDA. Nous observons une nette amélioration des scores relatifs aux catégories *theory* et *medical*. Une légère baisse de la catégorie *visu* et une baisse significative pour les catégories *database* et *datamining*. Cette dernière observation nous confirme que ces deux catégories posent des problèmes à cause des termes qu'elles utilisent : en effet, elles partagent beaucoup de termes et les nouveaux documents *database* sont forcément proches des thématiques relatives à la catégorie *datamining* dans l'historique.

Lorsque nous résumons notre fenêtre de contexte sous forme de thématiques, nous observons la même tendance que précédemment, à part pour la catégorie *visu* qui semble rester constante. Cette méthode permet d'identifier certains types de documents nouveaux sans comparer tous les documents : l'étape d'identification des nouvelles thématiques permet de diminuer le nombre de documents à comparer (100 documents par thématiques nouvelles identifiées). Aussi, les modèles thématiques permettent de manipuler des matrices plus légères (en comparant les thématiques de l'historique et du contexte) par rapport à la comparaison de documents deux à deux.

Modèle	1	2	3
database	9.45	7.65	4.25
datamining	1.65	0.75	0.25
medical	2.45	8.63	9.75
theory	16.80	17.45	22.75
visu	3.70	3.20	3.50

TABLE 3.5: Comparaison des mesures de précision à 100 des différents modèles

3.3.5 Conclusion

Nous avons montré que les modèles thématiques peuvent être utilisés pour la détection de nouveauté dans certains cas. Les modèles que nous avons développés sont des modèles simples basés uniquement sur des mesures de distances et nous observons des différences significatives par rapport à un modèle basé uniquement sur la comparaison des termes. Cependant, ils restent particulièrement sensibles à la distance entre les thématiques. Ils parviennent à détecter de nouvelles thématiques lorsque celles-ci sont très différentes de ce qui a été vu avant, mais sont incapables de détecter des thématiques dont les termes ont déjà été utilisés dans le même type de contexte. Nos expérimentations se basent sur une baseline que nous développons et sur deux approches basées sur les modèles thématiques. La méthode d'expérimentation que nous avons développée, à savoir l'insertion artificielle de thématiques annotées dans

le jeu de données, nous a limités par rapport aux méthodes de la littérature que nous pouvons utiliser pour nous comparer. En effet, le but premier de ce travail était de vérifier l'apport d'un modèle thématique de type LDA pour la détection de nouveauté par rapport à une méthode simple basée uniquement sur les distances entre les termes. Dans la suite de ce manuscrit, nous développerons des méthodes de détections et d'expérimentations qui nous permettent de nous comparer à d'autres algorithmes existants.

3.4 Détection de nouveautés à l'aide de modèles de plongement.

Dans la section précédente, nous avons exploré l'utilisation de modèles thématiques probabilistes pour détecter des documents comportant de la nouveauté. Nous avons vu que ce type d'approche n'est pas assez performant pour de la nouveauté "proche" de ce qui existait auparavant. Notre objectif, d'un point de vue industriel, est de détecter des thématiques émergentes ; thématiques qui se matérialisent par l'apparition de nouveaux termes ou par le changement de contexte autour de certains termes. Dans l'introduction, nous avons pris pour exemple le mot "souris", dont le contexte change (il passe du monde animal à l'informatique) **parce que** l'informatique est une thématique émergente. Contrairement aux travaux précédents [78, 79] qui cherchent les changements dans le passé, notre objectif consiste à détecter les signes associés à ce changement **dès qu'il commence à apparaître**. Nous sommes donc dans une logique d'anticipation et de détection de signaux faibles. Pour détecter le changement de contexte, nous étudions l'apport des modèles de plongements au niveau des mots.

3.4.1 Mouvements dans les espaces de représentations.

Les espaces vectoriels de mots permettent d'analyser le sens des mots, par rapport à leurs voisins et sont donc utiles pour étudier leurs mouvements. Nous avons présenté certains modèles permettant la construction de tels espaces dans la section 3.2.2. Ici, nous nous intéressons aux mouvements des mots dans ces espaces et nous basons le développement d'une méthode originale de détection de la nouveauté suite à une observation que nous avons faite.

État de l'art Nous avons, dans les chapitres précédents, parlé des rapprochements entre les domaines de la détection de la nouveauté et la détection d'évènements. Les évènements sont des observations faites sur des petites périodes : ce sont des approches **court terme**. Le domaine de la détection de la nouveauté se rapproche aussi du domaine de l'évolution du langage qui, lui, se concentre sur le **long terme**. En effet, les changements affectant le langage, c'est-à-dire les changements de sens des mots, sont la conséquence de l'apparition de nouveaux concepts, de l'émergence de nouvelles technologies qui changent les façons de vivre et, donc, de parler.

Plusieurs approches comme [6, 93, 94] ont utilisé des modèles thématiques pour observer ces changements. L'émergence des modèles de plongements comme [74] ont redynamisé ce champ d'études. Des travaux comme ceux de [78–80] ont analysé la structure des espaces de représentations vectorielles pour illustrer les changements de sens des mots sur de longues périodes. Dans [79] et [80], les auteurs quantifient le changement de sens ainsi que l'aspect polysémique des mots en observant la corrélation entre le mouvement dans l'espace de représentation et l'évolution de leur fréquence. Des modèles de représentations contextuelles comme BERT [76] ou ELMO [77] ont considérablement apportés au domaine du TAL, mais ils sont très récents, ils ont besoin de grosses quantités de données et leur aspect temporel est, pour l'instant, très peu étudié.

Jeux de données	langue	# de doc	# de cat	temporalité	# de mots
NYTAC	Anglais	1.8M	13	1995-2005	20.000
SCI	Anglais	8337	4	1990-2005	5.000

TABLE 3.6: Résumé des jeux de données utilisés.

Notre travail se situe entre le domaine de la détection d'évènement et celui de l'observation de changement dans le langage. Nous voulons être capables de détecter, rapidement si possible, les prémices de ces changements au niveau des mots qui portent l'émergence d'une nouvelle thématique.

Simulation de l'émergence d'une thématique Afin de détecter ce type de nouveauté, nous devons nous placer dans un scénario où nous avons une référence observable, c'est-à-dire des exemples de thématiques émergentes. Cela nous permet de mesurer l'efficacité de différents algorithmes et de développer de nouvelles méthodes. Dans la littérature, nous manquons d'exemples de thématiques émergentes. C'est pour cela que nous décidons de simuler cette émergence dans des jeux de données annotés par catégories. Nous utilisons les jeux de données du *New York Times* et des résumés d'articles scientifiques dont nous rappelons les caractéristiques dans le tableau 3.6.

Comme dans la section précédente, nous utilisons les annotations de catégories. Nous sommes dans un scénario où nous avons plusieurs instants. Au lieu de retirer une catégorie puis de l'ajouter, nous réorganisons sa dynamique. Ceci est effectué en sélectionnant une catégorie, puis en modifiant les dates associées aux articles afin que la dynamique générale corresponde à un signal d'émergence présenté sur la figure 3.14. Ce signal commence à un niveau faible, mais non nul afin que les mots associés à la catégorie émergente existent et soient représentés dans l'espace. Aussi, ce signal est bruité afin de mieux correspondre à la réalité.

Construction d'un référentiel au niveau des mots. Nous avons maintenant une référence au niveau des thématiques, c'est-à-dire que nous savons quelle

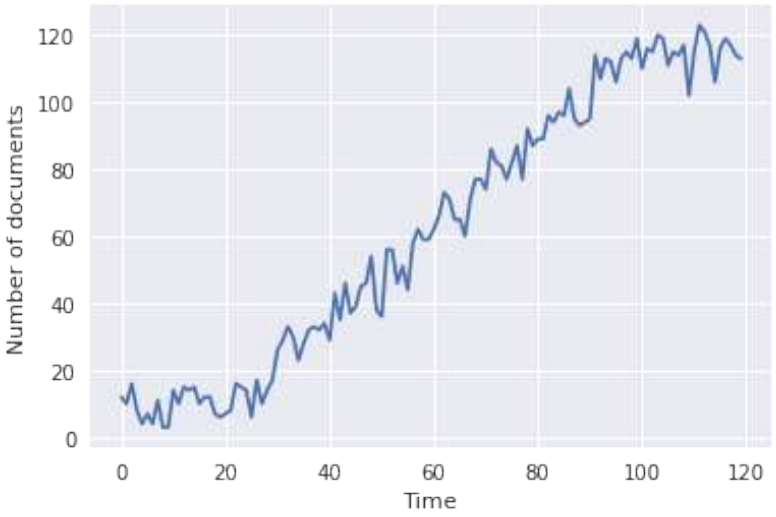


FIGURE 3.14: Signal utilisé pour simuler l'émergence d'une catégorie.

catégorie nous voulons détecter dans le jeu de données. Notre but étant de détecter des mots associés à cette catégorie émergente, nous devons définir un certain nombre de mots qui sont fortement liés à chacune des thématiques. Pour cela nous construisons un classifieur sur le début de notre jeu de données. Indépendamment de notre tâche de détection de nouveauté, nous résolvons une tâche de classification, pour laquelle nous voulons extraire les descripteurs (mots) les plus discriminantes pour chacune des thématiques.

Nous choisissons d'entraîner un classifieur de type *Naive Bayes*. Celui-ci présente l'avantage d'être simple, rapide, performant et surtout l'extraction des variables importantes est simple. Sur le jeu de données du *New York Times*, un classifieur de ce type obtient 78% de performance sur les 13 catégories utilisées dans cette expérience. Les mots les plus discriminants de certaines des catégories de NYT et de SCI sont présentés dans le tableau 3.7. Nous voyons bien que ces mots font partie du champ lexical lié à la catégorie. Nous sélectionnons les 100 mots les plus discriminants pour chaque catégorie pour construire notre référentiel : ce sont ces mots que nous voulons détecter lorsque nous introduisons artificiellement une catégorie émergente. Le nombre de 100 à été déterminé manuellement en observant la nature du vocabulaire : on considère que ces 100 mots font vraiment partie du champ

Database	Theory	Theater	Motion Pictures	Politics
query	problem	theater	film	party
data	algorithm	play	movie	government
database	bound	broadway	director	mayor
system	time	musical	hollywood	political
performance	polynomial	production	directed	election
object	approximation	show	actor	president
Restaurants	**Art**	**Murders**	**Basketball**	**Terrorism**
restaurant	art	police	game	attack
sauce	museum	officer	knicks	terrorist
dish	painting	murder	team	state
menu	artist	death	point	bombing
food	gallery	shot	player	official
dining	exhibition	kill	season	federal

TABLE 3.7: Variables les plus discriminantes pour certaines catégories de NYT et SCI.

lexical associé à la catégorie.

Observation des mouvements pour différentes dynamiques. Dans un espace de représentation vectoriel, les entités similaires (mots ou thématiques) sont proches si elles sont utilisées dans les mêmes contextes : leur sens dépend des entités autour d'elles. Au cours du temps, leur sens peut changer et donc leur représentation dans l'espace peut changer : il y a des mouvements. Dans ce chapitre, nous considérons des espaces de représentations de mots et nous observons un type de mouvement dans cet espace.

Nous avons $\mathcal{V} = \{w_1, \ldots, w_n\}$ un ensemble de mots. Pour chaque mot $w \in \mathcal{V}$, nous avons une représentation numérique contenue dans un vecteur $v^w \in \mathbb{R}^D$, construit via un modèle de plongement, disons $\mathcal{A}$ (e.g., `SVD`, `word2vec`, `Glove`, `fasttext`). Nous observons comment le vecteur v^w change lorsque nous mettons à jour l'algorithme $\mathcal{A}$ avec de nouveaux documents.

Nous avons $v_1^w, v_2^w, \ldots, v_T^w$ une séquence de vecteurs représentant le mot w à chaque instant $t = 1, \ldots, T$. Pour deux vecteurs consécutifs, nous quantifions la magnitude

du changement en mesurant la distance euclidienne :

$$e_w(v_t^w, v_{t-1}^w) = \|v_t^w - v_{t-1}^w\| \tag{3.1}$$

Les figures 3.15 et 3.16 montrent les mouvements euclidiens de deux mots ("*Terrorism*" et "*Film*") associés à certaines catégories ("*Terrorism*" et "*Motion Pictures*"). Ces deux catégories présentent des dynamiques différentes : "*Terrorism*" comporte un évènement majeur tandis que "*Motion Pictures*" est une catégorie émergente simulée.

Nous partons du principe que l'évolution de la fréquence des mots représentés correspond au signal de la catégorie, c'est-à-dire que si un mot est fortement lié à la catégorie, sa fréquence évolue de la même façon. Ces figures nous permettent d'établir notre hypothèse principale : **la corrélation entre le mouvement et la fréquence d'un mot ne semble pas être la même dans le cas d'un évènement ou d'une émergence**. En effet, sur la figure 3.15, nous observons une corrélation positive sur le modèle SGNS et une corrélation nulle via un procédé de type SVD. Sur la figure 3.16, nous observons une corrélation négative en général. Le lien entre le mouvement et la fréquence des mots a déjà été observé dans différentes études, [79, 80] mais nous mettons en lumière la différence entre des dynamiques spécifiques.

3.4.2 Modélisation

Définition des corrélations. Dans [79] et [80], les auteurs mettent en avant une corrélation entre mouvements et fréquence des mots. Nous calculons donc une corrélation de *Spearman* $\rho_{X,Y}$ entre un mouvement d_w et une fréquence f_w sur une partie du signal :

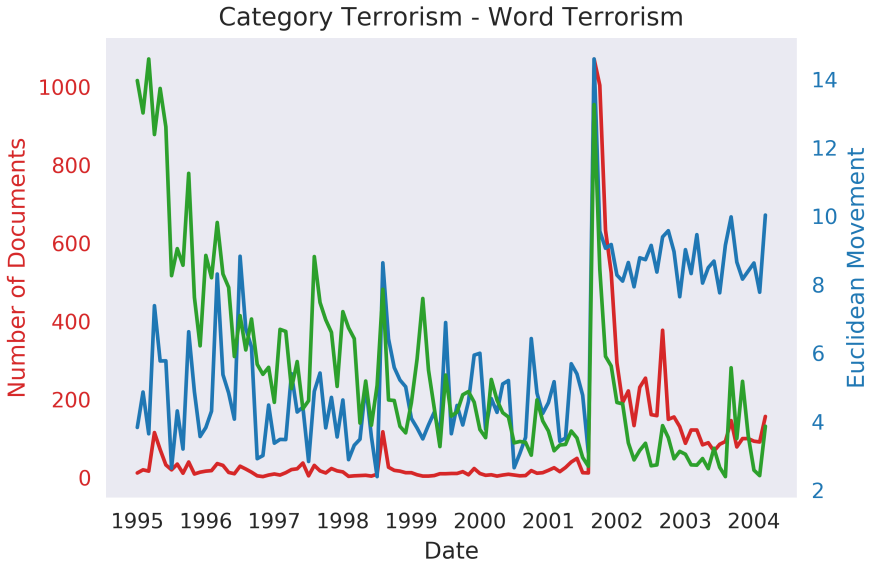


FIGURE 3.15: Évolution de la catégorie "*Terrorism*" (en rouge) et du mouvement du mot "*Terrorism*" pour les modèles SGNS (bleu) et SVD (vert).

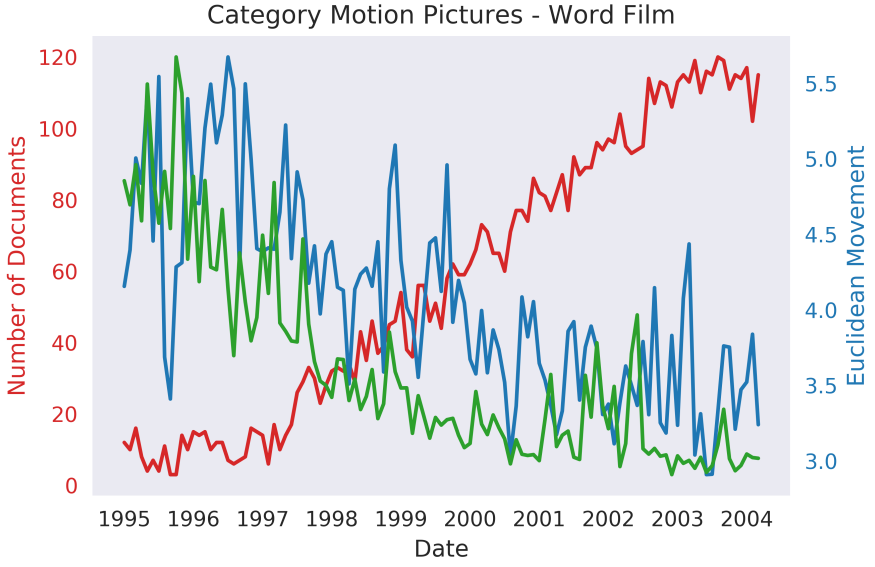


FIGURE 3.16: Évolution de la catégorie "*Motion Pictures*" (en rouge) et du mouvement du mot "*Film*" pour les modèles SGNS (bleu) et SVD (vert).

$$\rho_{rf_w, re_w} = \frac{\text{cov}(rf_w, re_w)}{\sigma_{rf_w}\sigma_{re_w}} \tag{3.2}$$

où f_w est le signal de fréquence du mot w et rf_w et re_w sont les variables de rangs des signaux f_w et e_w. *cov* correspond à la covariance et σ à l'écart type.

Construction des représentations. Comme nous l'avons dit dans la section précédente, nous introduisons, de manière contrôlée, une catégorie émergente dans notre corpus selon le signal présenté en figure 3.14. Les autres catégories sont laissées telles quelles : elles gardent leur dynamique naturelle. Quand la fréquence d'une catégorie augmente avec le temps, la fréquence des mots de son champ lexical augmente aussi. Nous analysons comment les vecteurs v^w changent lorsque nous ajoutons des documents pour mettre à jour l'espace de représentation.

Nous construisons deux espaces de représentations différents. Le premier est construit avec un algorithme de *Singular Value Decomposition* (SVD) sur une matrice de type *Shifted*-PPMI (SPPMI). Nous avons présenté ce type de matrice dans la section 3.2. L'algorithme de SVD permet de réduire la dimension des vecteurs de la matrice tout en gardant la structure générale. Le deuxième espace est construit avec un modèle de type Word2Vec : *Skip-Gram with Negative Sampling* (SGNS) [74] que nous avons présenté dans la section 3.2. Il a été démontré que ces deux modèles ont des résultats très proches en termes de qualité de représentation des mots tout en assurant une certaine stabilité [89]. Aussi, [13] a montré que les résultats des modèles SGNS peuvent être approchés via des méthodes de factorisation de matrices en utilisant la SPPMI. Pour rappel, la SPPMI est définie comme ceci :

$$\text{SPPMI}(x, y) = \max\left\{\log\frac{p(x, y)}{p(x)p(y)} - \log(s), 0\right\}. \tag{3.3}$$

où s est une constante.

Nous devons faire évoluer ces espaces dans le temps à chaque fois que nous ajoutons de nouveaux documents. Nous initialisons nos espaces de représentations à $t = 0$ avec les documents correspondant au premier instant de nos jeux de données : ceux de janvier 1995 pour les articles du *New York Times* par exemple. À chaque instant, les espaces sont mis à jour avec les documents de l'instant suivant tout en gardant le même vocabulaire qu'à l'initialisation (les nouveaux mots ne sont pas introduits dans l'espace). Pour un modèle SGNS, il suffit de mettre à jour les poids du modèle précédent avec les nouvelles observations. Pour le procédé SVD, nous devons ajouter une étape d'alignement. En effet, quand nous ajoutons de nouvelles observations, nous mettons à jour la matrice SPPMI puis nous devons recommencer la SVD. Deux SVD sur une même matrice ne produisent pas tout à fait les mêmes espaces, car certains axes de représentations sont inversés. Afin d'obtenir des mouvements interprétables, nous utilisons un processus de Procrustes orthogonal pour l'alignement entre deux SVD consécutives. Ce processus est notamment utilisé dans la littérature [79]. Ces représentations en chaînes nous permettent de calculer des mouvements et des corrélations sur l'ensemble de nos jeux de données.

Distribution des corrélations. Sur les figures 3.17 et 3.18, nous observons les distributions des corrélations entre le mouvement et la fréquence de tous les mots dans des espaces construits, respectivement, avec SGNS et SVD. La partie en verte correspond aux corrélations des mots effectivement émergents. Ces derniers correspondent à notre référence. Pour un modèle SGNS (figure 3.17, les corrélations sont majoritairement positives et certains mots (environ la moitié de notre référence) présentent une corrélation négative. Pour un procédé SVD, les corrélations sont majoritairement négatives et l'ensemble des mots de notre référence se trouve à l'extrême gauche de cette distribution, c'est-à-dire proche de -1. Nous arrivons, grâce à cette corrélation, à isoler les mots qui correspondent aux thématiques émergentes. Cette observation est faite a posteriori, c'est-à-dire à la fin de l'observation une fois que nous avons obtenu tous les documents. Notre de but est de

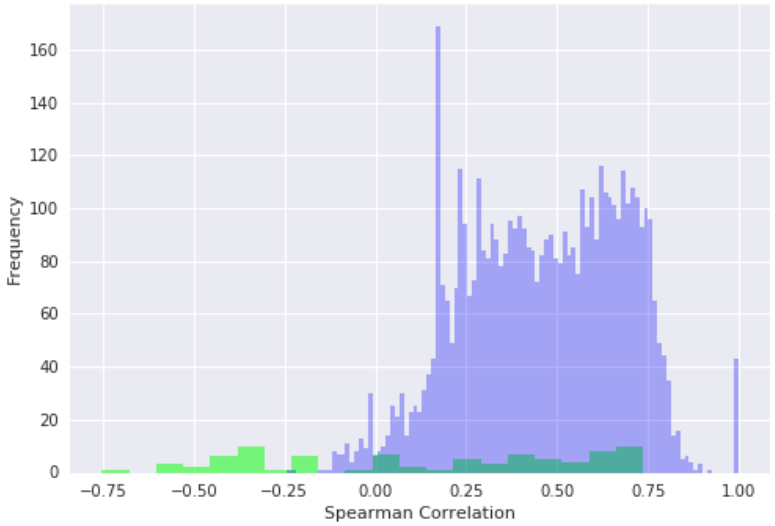


FIGURE 3.17: Distribution des corrélations entre mouvement dans un espace SGNS et fréquence des mots pour l'ensemble du vocabulaire sur le NYT. Les mots émergents sont en vert.

mettre en place un système pour surveiller cette corrélation au cours du temps et isoler des mots qui auraient des comportements émergents.

Détection de la nouveauté Auparavant, nous avons calculé puis observé la corrélation entre les signaux de mouvements et de fréquence sur le jeu de données complet. Notre but est de détecter les mots émergents via cette corrélation au fur et à mesure du temps. Pour cela, nous modifions donc le calcul de la corrélation pour la mettre à jour au fur et à mesure du temps (équation 3.4). Nous testons deux variantes : une corrélation qui varie sur une fenêtre de temps de taille n, c'est à dire entre $t-n$ et t et une corrélation entre $t=0$ et t.

$$\rho^t_{rf_w, rd_w} = \frac{\text{cov}(rf_w^{[t-n,t]}, rd_w^{[t-n,t]})}{\sigma_{rf_w^{[t-n,t]}} \sigma_{rd_w^{[t-n,t]}}} \tag{3.4}$$

Par rapport à l'équation 3.2, nous considérons des morceaux de signaux f_w et d_w. Un mot w est considéré comme émergent lorsque sa corrélation temporelle ρ^t_w passe

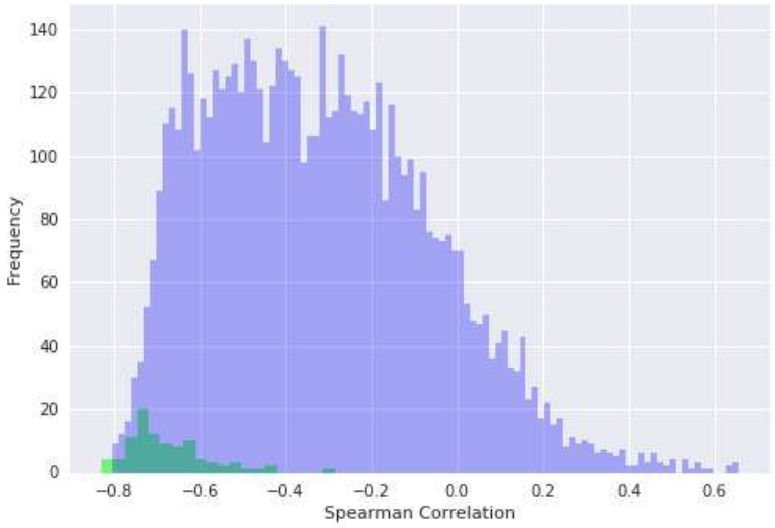


FIGURE 3.18: Distribution des corrélations entre mouvement dans un espace SVD et fréquence des mots pour l'ensemble du vocabulaire sur le NYT. Les mots émergents sont en vert.

en dessous d'un seuil K. C'est une méthode que nous avons nommé ***Correlation-based Embedding Novelty Detection (CEND)***.

Définition automatique du seuil de détection Une partie des performances de notre modèle CEND dépend du choix du seuil K. Nous devons choisir un seuil pour chacun des scénarios, c'est-à-dire pour chaque catégorie que nous testons et donc nous sommes dans l'impossibilité de le fixer manuellement. Ce seuil est une valeur qui représente le côté anormal de la corrélation. Nous observons donc la distribution de l'ensemble des corrélations sur le vocabulaire et nous estimons l'écart type σ par rapport à la moyenne $\bar{\rho}$. Cela nous permet de calculer des intervalles de variabilité à chaque instant. En rapportant cette distribution dans un espace gaussien, nous choisissons la valeur du 97.5ème quantile. Nous avons une valeur de seuil $K = -1.96 * \sigma + \bar{\rho}$. En testant sur le signal complet du *New York Times*, nous obtenons un seuil évoluant autour de -0.65 pour toutes les catégories dans un espace SVD. Cette valeur semble correspondre à la limite entre les mots en vert et en

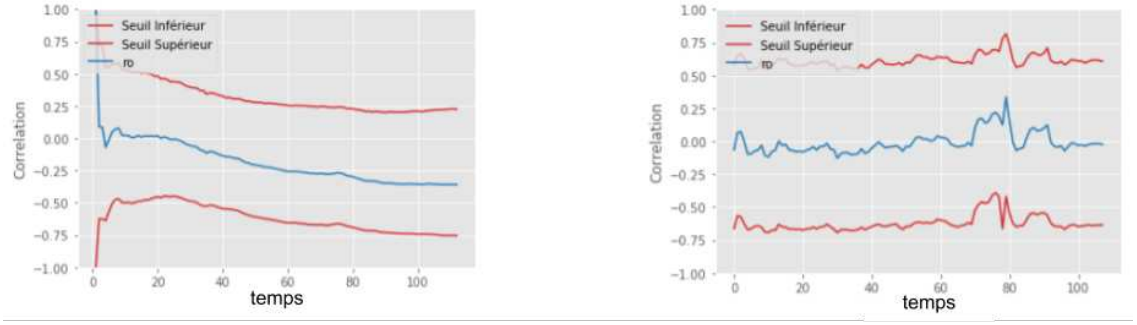


FIGURE 3.19: Évolution du seuil (rouge) dans le temps par rapport à la moyenne (bleu) sur le *New York Times*. À gauche, le seuil est calculé sur l'ensemble du signal. À droite, il est calculé sur une fenêtre de taille 10.

violet sur la figure 3.18. Sur la figure 3.19, nous observons l'évolution du seuil dans le temps. Lorsque celui-ci est calculé par rapport à l'ensemble du signal, c'est-à-dire pour $n = t$ dans l'équation 3.4, il a tendance à diminuer, car l'ensemble des mots ont leur sens qui se fixe dans l'espace : plus ils sont fréquents, plus leur contexte est bien défini, moins leur mouvement est important. Lorsque nous nous basons sur une petite fenêtre de temps pour calculer le seuil, il est globalement stable même s'il remontre entre les mois 70 et 80. Ces mois correspondent à l'évènement du 11 septembre 2001. Comme nous l'avons vu, les mots relatifs à des évènements très importants ont tendance a avoir des mouvements importants très corrélés a leur fréquence. Comme cet évènement touche un grand nombre de mots, la moyenne des corrélations et le seuil qui en découle connaissent un pic durant cette période.

3.4.3 Résultats

Nous avons construit un système permettant de détecter des mots associés à des thématiques émergentes. Nous avons aussi mis en place un système permettant d'avoir une référence. Nous devons nous comparer à d'autres méthodes de la littérature avant de pouvoir conclure sur les performances générales de notre modèle CEND.

Baselines Afin d'évaluer notre approche, nous nous comparons à quatre méthodes de la littérature : elles constituent nos baselines. Comme la tâche que nous résolvons est peu étudiée, nous sélectionnons des méthodes qui travaillent sur des problématiques proches et qui sont facilement adaptables pour fonctionner sur nos scénarios. Certaines de ces baselines ont été présentées dans les chapitres précédents, mais nous détaillons ici leur mécanisme d'alerte.

Notre première baseline est adaptée de [1] (TFIDF) dans laquelle une méthode permet de détecter des thématiques nouvelles. Cette méthode est basée sur la mesure de TFxIDF que nous avons présentée dans ce manuscrit. Elle permet de lancer des alertes sur des mots quand leur mesure de TFxIDF passe un seuil déterminé manuellement. Le second algorithme avec lequel nous allons nous comparer est [2] (TopicSketch). C'est une méthode qui se base sur la surveillance de mesures physiques (vitesse, accélération) des entités textuelles (mots et n-grammes). Cette méthode lance des alertes lorsque ces mesures passent un seuil défini manuellement. Dans [72] (HUPC), les auteurs développent une méthode pour extraire des structures représentatives de thématiques émergentes dans des données de réseaux sociaux. Après avoir isolé ces structures avec une mesure de "*Utility*", ils déterminent si celles-ci font partie d'une nouvelle thématique ou d'une thématique connue à chaque instant. [73] (ET-EPM) utilise la même métrique de "*Utility*" et la combine avec une mesure de nouveauté basée sur la prédiction de l'évolution d'un mot. Ils utilisent des méthodes d'analyses de graphes pour grouper les mots en thématiques.

Résultats généraux Dans le tableau 3.8, nous présentons les résultats généraux obtenus par notre méthode CEND avec des vecteurs de mots construits avec SNGS ou SVD. Nous comparons nos résultats avec ceux des baselines présentées précédemment. Pour chacune des méthodes, nous avons considéré les mots qui ont été détectés au moins une fois durant toute la période temporelle. Les valeurs de précision (P),

	NYTAC			SCI		
	P	R	F	P	R	F
TFIDF [1]	0.17	0.12	0.14	0.10	0.12	0.11
TopicSketch [2]	**0.48**	0.17	0.25	0.20	0.15	0.17
HUPC [72]	0.25	0.19	0.22	0.14	0.16	0.15
ET-EPM [73]	0.27	0.22	0.24	0.18	0.22	0.20
CEND-SGNS	0.37	0.33	0.34	0.22	0.32	0.26
CEND-SVD	0.32	**0.45**	**0.37**	**0.24**	**0.36**	**0.29**

TABLE 3.8: Performance moyenne des méthodes de détection de nouveauté

rappel (R) et F-Mesure (F) ne sont pas particulièrement hautes, mais nous devons les remettre dans leur contexte : nous cherchons à détecter les 100 mots de notre référence parmi un vocabulaire de taille 20 000 pour le *New York Times* et 5000 pour le jeu de données d'articles scientifiques. Pour les deux jeux de données, nous évaluons plusieurs fois notre approche : nous testons chaque catégorie comme émergente, en l'introduisant artificiellement. Comme le processus de simulation de cette émergence mélange les documents de la catégorie nouvelle, nous faisons les expériences 5 fois par catégorie afin de garantir des résultats stables.

Sur les deux jeux de données, notre approche est meilleure en termes de F-Mesure tandis que TopicSketch [2] a une meilleure précision sur le *New York Times*. Celle-ci peut être expliquée grâce au rapport entre précision et rappel : elle a une tendance a réduire les erreurs de détection en produisant moins d'alertes. Le rappel globalement inférieur montre qu'elle ne détecte pas beaucoup de mots émergents et donc le résultat global (F-Mesure) est inférieur à notre approche. Sur le jeu de données des articles scientifiques, les résultats sont inférieurs, car il y a moins d'instants à analyser : le corpus est séparé en seulement 15 instants (15 années) et donc notre méthode a seulement 14 valeurs de corrélation à analyser pour détecter l'émergence de ces nouveaux mots. Les résultats sont assez similaires entre CEND-SGNS et CEND-SVD bien que les représentations via SVD semblent être plus stables.

Afin de vérifier si les mots détectés le sont bien à cause de leur lien avec la catégorie artificiellement introduite, nous évaluons notre méthode avec un groupe de contrôle

	P	R	F
NYTAC	0.02	0.04	0.03
SCI	0.02	0.02	0.02

TABLE 3.9: Performance moyenne de CEND-SGNS sur un groupe de contrôle

où aucune catégorie n'est introduite. Au lieu d'introduire une thématique selon le signal représenté sur la figure 3.14, nous mélangeons les documents d'une catégorie et nous les introduisons selon un signal bruité constant. Nous pouvons donc vérifier, dans le tableau 3.9, que les performances de Précision/Rappel/F-Mesure sont significativement inférieures à celles présentées dans le tableau 3.8.

Notre méthode CEND lance des alertes chaque fois qu'une corrélation liée à un mot passe un seuil qui est défini automatiquement. Nous avons jusqu'ici regardé les résultats généraux pour les mots qui ont été détectés au moins une fois, mais il est aussi intéressant d'observer les mots qui ont été détectés plusieurs fois. En triant les mots par rapport au nombre d'alertes qu'ils ont lancées, nous pouvons évaluer notre modèle avec des métriques classiques dans le domaine de la recherche d'information : la courbe ROC (*Receiver Operating Characteristic*) et l'aire sous la courbe (*Area Under Curve* (AUC))

Sur la figure 3.20, nous observons une courbe ROC pour la catégorie *Motion Pictures* du *New York Times*. Cela signifie que nous avons introduit artificiellement cette catégorie en tant que thématique émergente et nous cherchons à en détecter les mots. Nous voyons, au début de cette courbe, que notre modèle CEND-SGNS a tendance a lancer plus d'alertes sur les mots de notre référence : ce sont ceux qui ont reçu le plus d'alertes et donc qui sont en haut de notre classement. Dans le tableau 3.10, nous observons la métrique d'AUC pour quelques catégories du *New York Times* et du jeu de données d'articles scientifiques. Nous observons aussi les quelques mots les plus détectés pour ces catégories. Nous remarquons que, bien que notre modèle ne détecte pas l'ensemble des 100 mots de notre référence, il a tendance à mettre en avant, c'est-à-dire à les détecter plus souvent, certains de ses mots. Globalement, l'AUC reste entre 0.70 et 0.85 ce qui paraît très satisfaisant.

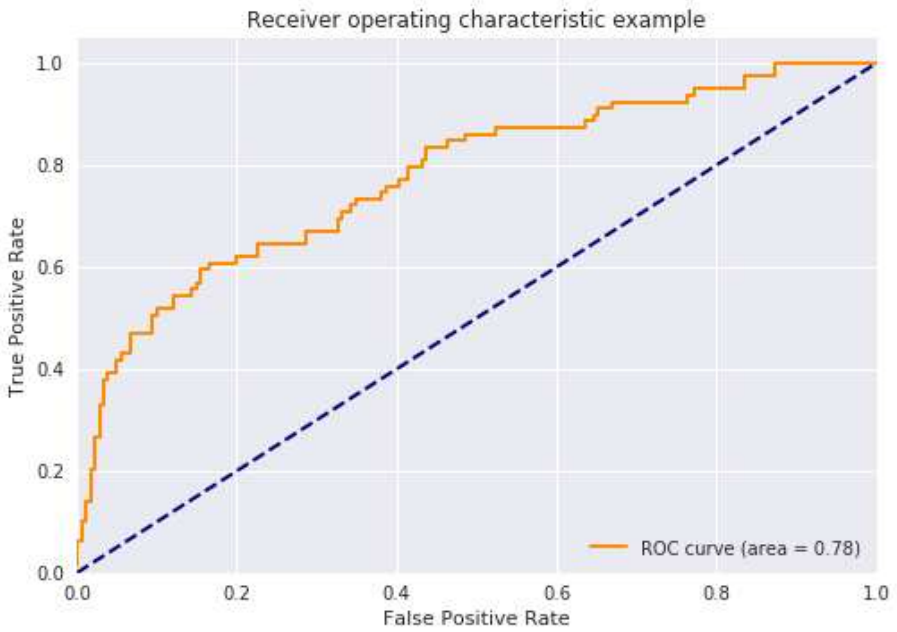


FIGURE 3.20: Courbe ROC pour la catégorie *Motion Pictures* du *New York Times*.

Catégorie	AUC	Mot_{SNGS}	Mot_{SVD}
Database	0.71	database algorithm access	query data database
Theory	0.79	general constant linear	problem algorithm polynomial
Theater	0.82	written character play	play broadway show

TABLE 3.10: Exemples d'AUC et des mots les plus souvent détectés pour certaines catégories de NYT et de SCI.

3.4.4 Cas d'applications EDF

Nous avons développé cette idée à partir d'une observation faite sur le jeu de données du New York Times, confirmée sur un jeu de données plus petit, mais toujours en anglais et toujours contenant des textes bien formatés : c'est-à-dire avec des concepts bien définis, une bonne orthographe et des vocabulaires précis. Avec l'idée

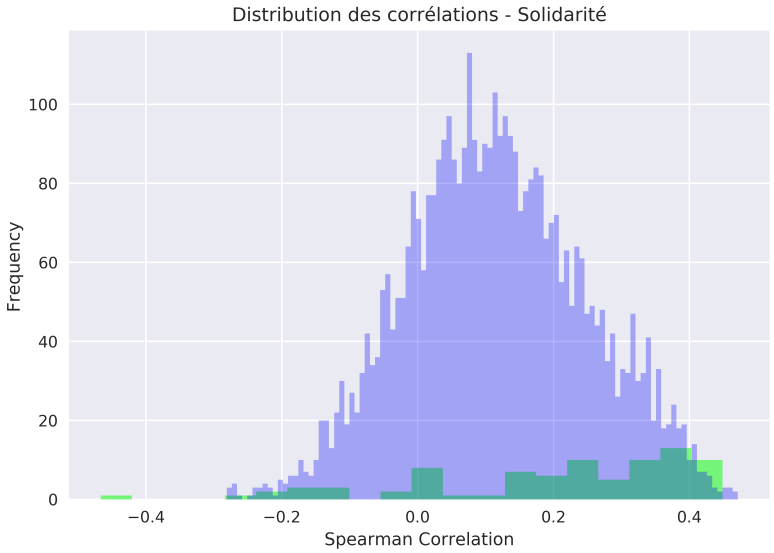


FIGURE 3.21: Distribution des corrélations entre mouvement dans un espace SVD et fréquence des mots pour la catégorie émergente Solidarité dans les courriels EDF.

d'appliquer ce modèle sur des cas d'applications industriels, nous l'avons testé sur le jeu de données des courriels EDF.

En appliquant le même processus d'expérimentation, c'est-à-dire en ajoutant artificiellement une catégorie de manière émergente, nous ne faisons pas les mêmes observations. En effet, la distribution des corrélations entre le mouvement des mots dans un espace SVD et leur fréquence (Figures 3.21 et 3.22) n'est pas la même que lors de nos expérimentations sur le *New York Times* (Figure 3.18). Dans le cas du *New York Times*, nous avons une distribution globalement négative, centrée autour de $-0,4$ et avec des mots identifiés comme émergents (en vert) proche de -1. Dans le cas des courriels EDF, la distribution des corrélations ressemble à une gaussienne centrée quasiment sur 0 et sur laquelle les mots émergents semblent être répartis plus à droite. Cette observation ne nous permet donc pas d'appliquer notre modèle CEND pour détecter les mots associés à des thématiques émergentes, car notre hypothèse de base n'est pas vérifiée dans ce cas.

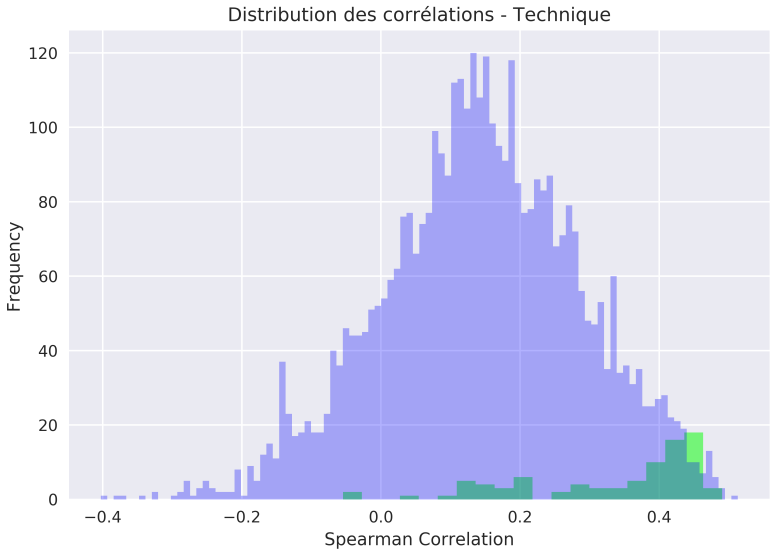


FIGURE 3.22: Distribution des corrélations entre mouvement dans un espace SVD et fréquence des mots pour la catégorie émergente Technique dans les courriels EDF.

Cette absence de corrélation entre mouvement et fréquence des mots peut s'expliquer par le fait que les courriels envoyés par les clients à EDF sont bien moins structurés que les jeux de données utilisés auparavant. En effet, les concepts abordés ont un vocabulaire bien spécifique, mais qui n'est pas forcément correctement maîtrisé par les clients qui ne sont pas experts du domaine. Les mots évoluent donc dans des contextes très différents, ce qui est illustré par un mouvement très important et très bruité. Sur la Figure 3.23, nous voyons, à gauche, que la corrélation négative semble être respectée pour l'exemple du mot "Sociale" lorsque la catégorie "Solidarité" est émergente. Globalement, le sens du mot se fixe dans l'espace, car son mouvement diminue, mais ce dernier est assez bruité et présente des variations importantes. La fréquence du mot augmente, mais est aussi globalement très bruitée. Dans ce cas précis, nous arrivons à une corrélation autour de $-0,4$ alors que nous avions des corrélations entre -0.9 et -1 pour tous les mots émergents sur les catégories du *New York Times*.

Dans un deuxième temps, les fréquences des mots sont elles-mêmes très bruitées

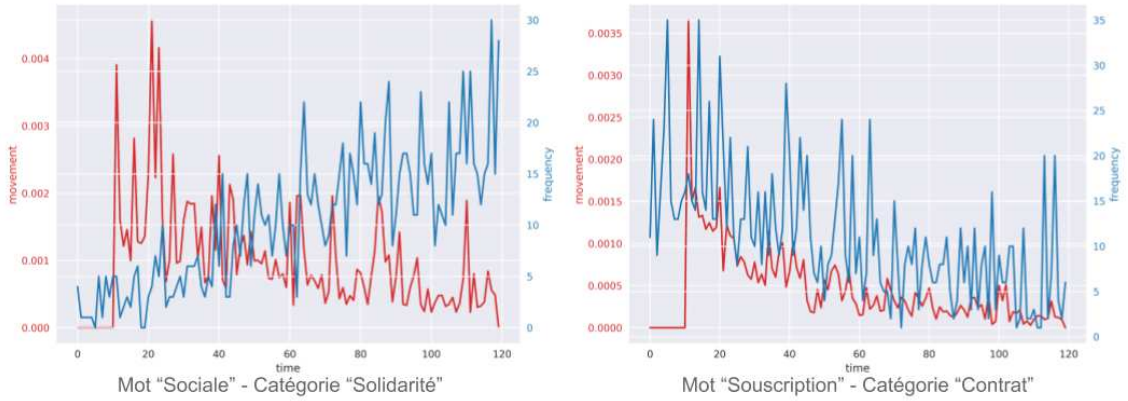


FIGURE 3.23: Évolutions des mouvements et fréquences des mots “Sociale” et “Souscription” dans les catégories émergentes “Solidarité” et “Contrat” respectivement

et, concernant les mots censés être émergents, elles ne correspondent pas au signal de la catégorie introduite. Par exemple, sur la Figure 3.23, le mot “souscription” est censé être le descripteur le plus important de la catégorie contrat. En effet, un client parle de souscription lorsqu'il souscrit à un nouveau contrat et donc nous nous attendons, lorsque nous introduisons artificiellement la catégorie “contrat”, à avoir un signal émergent pour ce terme. En fait, le mot “souscription” est utilisé dans une multitude de contextes : lorsqu'un client se plaint qu'un concurrent à voulu le faire souscrire à une offre (catégorie “Démarchage Abusif”), que la souscription à une option entraîne un surcoût (catégorie “Montant”), etc. Ce comportement rend obsolète notre méthode d'insertion artificielle, car l'émergence des termes ne correspond plus à l'émergence de la catégorie. Enfin, nous nous rendons compte que notre modèle est limité lorsque les concepts émergents que nous voulons détecter au plus tôt sont très instables et présentent beaucoup de vocabulaires en commun avec des thématiques connues.

3.4.5 Conclusion

Dans cette section, nous avons étudié l'apport des modèles de plongements pour la détection des mots associés à des thématiques émergentes. Nous nous sommes basés

sur l'hypothèse que les signaux associés à une émergence douce présentent un certain type de mouvement dans des espaces construits avec SVD et Word2Vec SGNS. Nous avons vérifié cette hypothèse et observé que les mots associés à une émergence douce présentent une corrélation négative entre leur mouvement et leur fréquence. Nous avons remarqué qu'une corrélation positive est plus présente lorsque nous utilisons un algorithme de type Word2Vec pour la modélisation. Cette corrélation positive est en fait une des caractéristiques principales de ce type de modélisation : plus un mot est utilisé dans un même contexte, plus la norme de son vecteur est grande [95]. La corrélation très négative des modèles SVD par rapport au SGNS peut être expliquée par le fait que la SVD est un algorithme beaucoup plus stable et moins biaisé envers les nouvelles observations comme SGNS. Cette hypothèse est introduite dans, [89] mais mériterait plus d'analyses théoriques. Le modèle CEND que nous avons développés a montré son efficacité sur des données correctement structurées, comme des articles de presses ou des articles scientifiques. Dans ce type de données, les textes sont écrits par des spécialistes métiers, les mots sont bien choisis et les champ lexicaux sont donc bien définis. En revanche, notre hypothèse de départ n'a pas été vérifiée sur les données courriels d'EDF. En effet, comme ils sont écrits directement par des clients et non par des spécialistes du domaine, les mots clés sont beaucoup moins facilement identifiables et apparaissent dans plusieurs catégories pas forcément émergentes. D'un point de vue industriel, les applications du modèle CEND sont possible en interne mais le choix du jeu de données est primordial.

3.5 Conclusion générale

Dans ce chapitre, nous avons étudié l'utilité des méthodes de modélisation thématique ainsi que l'apport des modèles de plongements de mots en grande dimension.

Nous avons vu que les modèles thématiques probabilistes permettent la détection de la nouveauté si celle-ci est assez différente de ce que nous avons déjà pu observer.

Si la nouvelle thématique partage beaucoup de termes et de contexte avec une thématique existante, elles auront tendance à se mélanger et la détection ne pourra pas avoir lieu.

Nous avons ensuite développé une méthode pour détecter les mots associés à des thématiques émergentes en utilisant les propriétés des espaces de représentations des mots en grande dimension. Nous avons basé notre méthode sur le fait que les mots associés à une thématique émergente présentent un type de mouvement spécifique dans ce type d'espace. Nous avons vérifié cette hypothèse en expérimentant sur plusieurs catégories et nous avons conclu qu'une corrélation très négative entre le mouvement et la fréquence d'un mot était signe de l'appartenance à une catégorie émergente ce qui n'est pas le cas avec les évènements : où les mots ont tendance à présenter une corrélation positive.

Plusieurs extensions sont facilement imaginables pour notre méthode. Nous avons montré, sur la figure 3.18, la différence entre les distributions des corrélations des mots nouveaux et des mots non nouveaux. Nous pourrions imaginer un test statistique qui permettrait de détecter au plus tôt l'apparition d'une distribution différente. Pour adapter cette méthode au contexte industriel, il serait intéressant de grouper automatiquement les mots d'un même champ lexical afin de mettre en lumière des nouvelles thématiques au lieu de nouveaux mots et ainsi de lier les deux approches développées dans ce chapitre.

Chapitre 4

Surveillance de plan de classement prédéfini

Dans le chapitre 2, nous avons rappelé que nous nous concentrons sur deux types de nouveautés à détecter dans les données textuelles : les nouveautés de **Volumes** et de **Structures**. Après avoir présenté nos travaux autour des nouveautés de structures dans le chapitre 3, nous présentons ici une méthode que nous avons développée pour détecter au plus tôt les nouveautés de volume.

4.1 Introduction

De nos jours, les entreprises reçoivent énormément de retours textuels de la part de leurs clients. Ces retours peuvent prendre la forme de Tweets, de courriels ou de retranscription de conversation téléphonique. Pour toutes ces entreprises, il est nécessaire d'analyser les thématiques discutées dans ces données afin de comprendre les besoins des clients. Nous avons vu que ces thématiques présentent des dynamiques différentes qui apportent des éléments d'informations essentiels pour l'amélioration de la relation client.

L'entreprise EDF reçoit près de 300.000 courriels clients par mois et a donc dû développer des algorithmes de classifications afin de répartir ces données dans diverses entités du groupe. Ces algorithmes ont été développés de manières supervisées avec l'aide d'une forte expertise métier pour déterminer les catégories qui composent ces données : c'est ce qu'on appelle le "plan de classement". Cependant, un certain nombre de courriels clients se retrouvent mal ou non classés. Cela peut être du a des changements dans la structure même des catégories : nous l'avons vu au chapitre précédent. Cela peut aussi être du a des dynamiques qui changent de manière imprévue dans ces catégories. Ces changements peuvent être un signe de nouveauté sous-jacente ou de la nécessité de mettre à jour les catégories du plan de classement. Il est essentiel pour EDF de détecter, le plus rapidement possible, ces changements inattendus.

La méthode que nous développons doit répondre à plusieurs objectifs :

1. Apprendre la dynamique attendue des catégories du plan de classement.
2. Lancer des alertes lorsque ces dynamiques deviennent anormales.
3. Être rapide pour détecter des changements de dynamiques, mais ne pas être trop sensibles aux valeurs extrêmes (anomalies).
4. Apporter une explication aux alertes.

Dans ce chapitre, nous présentons les différentes approches de la littérature qui permettent de résoudre ces tâches et nous développons une nouvelle méthode basée sur des approches de prédictions de séries temporelles et sur une méthode d'analyse séquentielle. Après avoir détaillé le coeur de notre modèle, nous expliquerons les différentes données textuelles à surveiller pour que celui-ci fonctionne. Nous montrerons, enfin, des résultats sur des données du *New York Times* et de courriels EDF.

4.2 Approches de la littérature

Tout d'abord, nous devons étudier les différentes approches de la littérature qui permettent de résoudre les tâches que nous avons définies.

Comme nous l'avons déjà décrit dans les chapitres précédents, il est courant en analyse de données textuelles de travailler à la détection d'évènements, d'anomalies ou de nouveautés. Des méthodes comme [1, 72, 73, 83, 85] sont en effet construites dans l'idée de lancer des alertes lorsqu'une observation anormale est détectée. Bien qu'elles soient développées dans l'idée de détecter des évènements et pas des nouveautés, nous les considérons comme les méthodes les plus proches de notre cas d'applications.

Nous utilisons la baseline de OLDA *Online-Latent Dirichlet Allocation* [84] dont nous avons décrit le fonctionnement dans les chapitres précédents. Dans OLDA des alertes sont lancées lorsque le sens d'une thématique est modifié par la publication de nouveaux documents. Enfin, nous avons déjà parlé de *TopicSketch* [2] que nous décrirons plus en détail dans la section 4.5.2. Même si ces deux méthodes surveillent des entités textuelles différentes (thématiques vs. mots), nous remarquons que la transformation de l'information textuelle vers des séries temporelles est commune aux deux. Nous avions vu dans le chapitre 2 que l'algorithme OLDA présentait une capacité à retenir l'aspect cyclique d'une thématique. Cependant, les performances sur des scénarios d'émergences et d'évènements étaient globalement très faibles. C'est pourquoi nous nous inspirerons, dans notre approche, de l'aspect thématique développé dans OLDA mais nous nous comparerons à deux baselines [1] et [2].

Nous avons montré que notre objectif principal consiste à surveiller l'évolution temporelle d'un plan de classement. Nous ne voulons pas détecter des valeurs anormales isolées, mais plutôt des agrégats d'anormalités qui nous donnent des informations sur de potentiels changements dans une catégorie. Cette tâche est peu courante dans le domaine de l'analyse textuelle. C'est pourquoi nous nous sommes inspirés de

méthodes développées dans d'autres champs d'application comme : les processus industriels [96], la santé [97], le contrôle qualité [98], la surveillance de lignes électrique [99] ou encore la cybersécurité [100]. L'ensemble de ces travaux font le choix de surveiller des séries temporelles en utilisant des méthodes d'analyse séquentielles telles que CUSUM (*Cumulative Sum Control Chart*) [101].

4.3 Modèle CDPred

L'idée générale derrière notre approche est que la nouveauté apparaît lorsque nous ne sommes plus capables de prévoir correctement le futur. Nous nous basons donc sur des méthodes de prédiction de série temporelle afin de prédire les dynamiques des catégories d'un plan de classement.

4.3.1 Prédiction univariée

Nous voulons prédire le comportement normal d'un signal $[y_1, y_2, ..., y_t, ..., y_N]$. Ce comportement normal à un instant t est exprimé comme :

$$y_{t+1} = h(y_{0,...,t}) + \epsilon_t$$

où ϵ_t correspond à du bruit blanc.

Nous prédisons uniquement la valeur suivante du signal y_{t+1} en utilisant les valeurs passées $y_{0,...,t}$. Pour cela, nous comparons deux algorithmes de prédiction :

— K plus proches voisins (KNN) : traditionnellement utilisé pour des tâches de classifications et de régression, on peut utiliser cet algorithme pour de la prédiction. En fixant une valeur l, entier positif représentant un retard maximum, on utilise les valeurs passées comprises dans ce retard $y_{t,...,t-l}$ pour prédire la valeur suivante y_{t+1}. Sur une fenêtre glissante, nous sommes

donc capables d'entraîner un modèle à partir des caractéristiques $[y_{t,...,t-l}]$ et des valeurs cibles à prédire y_{t+1}.

— *AutoRegressive Integrated Moving Average* (ARIMA) est un algorithme classique de prédiction de série temporelle. Il cherche à prédire les valeurs futures d'un signal seulement à partir d'un modèle de régression linéaire basé sur les valeurs passées.

4.3.2 Prédiction avec des variables exogènes

Nous utilisons un modèle de prédiction basé sur des caractéristiques extérieures observées dans les données textuelles : ces caractéristiques sont des variables dites exogènes.

On a $[y_1, y_2, ..., y_t, ..., y_N]$, une série temporelle où y_t est la valeur de la série au moment t et $[x_{1f}, x_{2f}, ..., x_{tf}, x_{Nf}]$ une série temporelle où x_{tf} est la valeur de la variable f au moment t. Notre modèle est de la forme :

$$y_{t+1} = h(x_{t,1}, x_{t,2}, ..., x_{t,F}) + \epsilon_t$$

où F est le nombre de variables utilisées et ϵ_t est un terme d'erreur. La fonction h permet de lier les entrées $x_t \in \mathbb{R}^F$ à $y \in \mathbb{R}^+$.

Pour estimer la fonction h, nous utilisons un modèle de type *Random Forest* [102]. Bien que cet algorithme soit souvent utilisé pour des tâches de classification, il fonctionne aussi pour des tâches de prédiction de série temporelle [103]. Par rapport à des approches plus récentes basées sur des réseaux de neurones, les *Random Forest* présentent l'avantage d'être "explicable" et de donner des informations quant à l'importance des variables. En effet, le comportement d'un modèle de type *Random Forest* est facilement explicable, car nous pouvons mesures l'importance des

variables, c'est-à-dire que nous pouvons observer, pour toute prédiction, quelles sont les variables qui ont joué un rôle important. Ce modèle permet d'estimer l'importance d'une variable en mesurant l'augmentation dans l'erreur de prédiction lorsque cette variable est modifiée.

Le pas de temps δt est constant : cela peut être des heures, des jours ou des mois. Le modèle *Random Forest* est entraîné sur un sous-ensemble des données avec $t \in \{1, ..., t_c\}$ où t_c est une constante marquant la fin de la fenêtre d'entraînement. Une fois que le modèle est entraîné, nous pouvons l'utiliser pour prédire $\hat{y}_{t+1}$ pour $t \in \{t_{c+1}, ..., t_N\}$. Nous obtenons une prédiction à un instant dans le futur puis, à l'aide d'un mécanisme de roulement, nous avançons pas à pas pour prédire la série temporelle entière. Le modèle est évalué par rapport à l'erreur entre la valeur prédite et la valeur réellement obtenue : $e_{t+1} = y_{t+1} - \hat{y}_{t+1}$. Nous faisons l'hypothèse qu'une grande erreur de prédiction est un signe de changement. C'est dans cette optique que nous surveillons l'évolution de e_t afin de détecter des comportements nouveaux.

4.3.3 Contrôle du changement

Notre but final est de lancer une alerte lorsque notre erreur de prédiction devient trop importante. Nous sommes intéressés par les changements qui ont lieu sur le long terme et non pas par les valeurs extrêmes (des pics ou des creux) qui pourraient s'apparenter à des anomalies. Afin de surveiller ces changements, nous utilisons une méthode populaire dans le domaine du contrôle de processus statistique : *Cumulative Sum Control Chart* que nous appellerons, dans ce manuscrit, CUSUM.

CUSUM est une méthode d'analyse séquentielle conçue pour détecter des changements de régime dans des processus. Pour cela, une variable s_t représentant une somme cumulée est calculée puis surveillée dans le temps [104]. Des écarts par rapport à une valeur cible sont successivement ajoutés pour obtenir des valeurs

cohérentes de la statistique CUSUM lorsqu'un processus s'écarte de la cible. Cela permet de définir une zone de confort dans laquelle la statistique doit évoluer.

La statistique g_t enregistre les valeurs de s_t en les sommant une à une. Elle déclenche une alerte lorsque sa valeur dépasse un seuil h prédéfini. Cette méthode ajoute aussi un paramètre d'oubli, c'est-à-dire que pour éviter que la valeur prenne trop d'ampleur et que les valeurs anciennes aient autant d'importance que les nouvelles, une valeur ν est retirée de g_t à chaque instant.

$$g_t = max(g_{t-1} + s_t - \nu, 0) \text{ ou } max(g_{t-1} - s_t - \nu, 0),$$

$$\text{une alarme est lancée si } g_t > h$$

Dans notre cas, nous surveillons l'erreur de prédiction qui doit rester dans une zone de variation contrôlée (notre zone de confort). Pour cela, nous avons $\hat{\theta}_t = \frac{1}{t-t_0} \sum e_k$ qui représente la statistique cumulée sur la période commençant juste après la dernière alarme lancée t_0. Au temps t, nous observons le changement de la statistique $s_t = e_t - \hat{\theta}_{t-1}$.

La méthode CUSUM dépend de deux hyper paramètres que nous devons soit fixer manuellement, soit via une heuristique. Ces paramètres h et ν contrôlent la sensibilité de notre système face au changement. Dans la plupart des applications connues de CUSUM, ils sont fixés manuellement après une expertise poussée des processus surveillés. Dans notre cas, nous devons surveiller plusieurs signaux en même temps (un signal par catégorie) et nous faisons le choix de définir une heuristique qui nous permet d'avoir une valeur de paramètre h qui dépend du signal. Le paramètre ν est, quant à lui, fixé manuellement à 2 pour tous les signaux.

Le paramètre h est mis à jour en fonction de la valeur maximale observée durant la phase d'apprentissage sur le signal surveillé : cela permet d'avoir une méthode qui n'est pas dépendante de l'amplitude de nos signaux. Nous avons donc :

$$h = \frac{1}{2} max(y_{1,\dots,t_c}),$$

Cette valeur de seuil représente la sensibilité de notre modèle à notre erreur de prédiction. Puisque nous voulons limiter le nombre de fausses alarmes ainsi que la détection de pics temporaires, nous avons défini cette heuristique a posteriori, c'est-à-dire durant les expérimentations sur nos jeux de données.

Notre approche est donc basée sur l'utilisation d'un modèle de prédiction basé sur des variables exogènes tirées du texte et sur l'analyse de son erreur avec une méthode CUSUM. Nous nommons cette approche **CDPred** pour *Change Detection Prediction*

4.4 Expérimentations

Nous avons choisi d'utiliser un modèle de type *Random Forest* pour prédire l'évolution de nos signaux. Ce type de modèle permet d'observer facilement l'importance des variables utilisées. Nous présentons ici quelles sont les variables que nous avons choisi d'observer et quelles sont celles qui ont apporté le plus d'informations à nos modèles.

4.4.1 Jeux de données

Pour ces travaux, nous avons testé nos méthodes sur deux jeux de données : celui du *New York Times* et celui des courriels EDF. Ces jeux de données présentent l'avantage d'avoir assez d'instants pour que des dynamiques "anormales" puissent apparaître.

Pour le *New York Times*, nous choisissons de surveiller 13 catégories différentes. Ces catégories ont été choisies dans le but de mélanger des dynamiques différentes. Certaines catégories, comme la politique ou le terrorisme, sont régulièrement sujettes

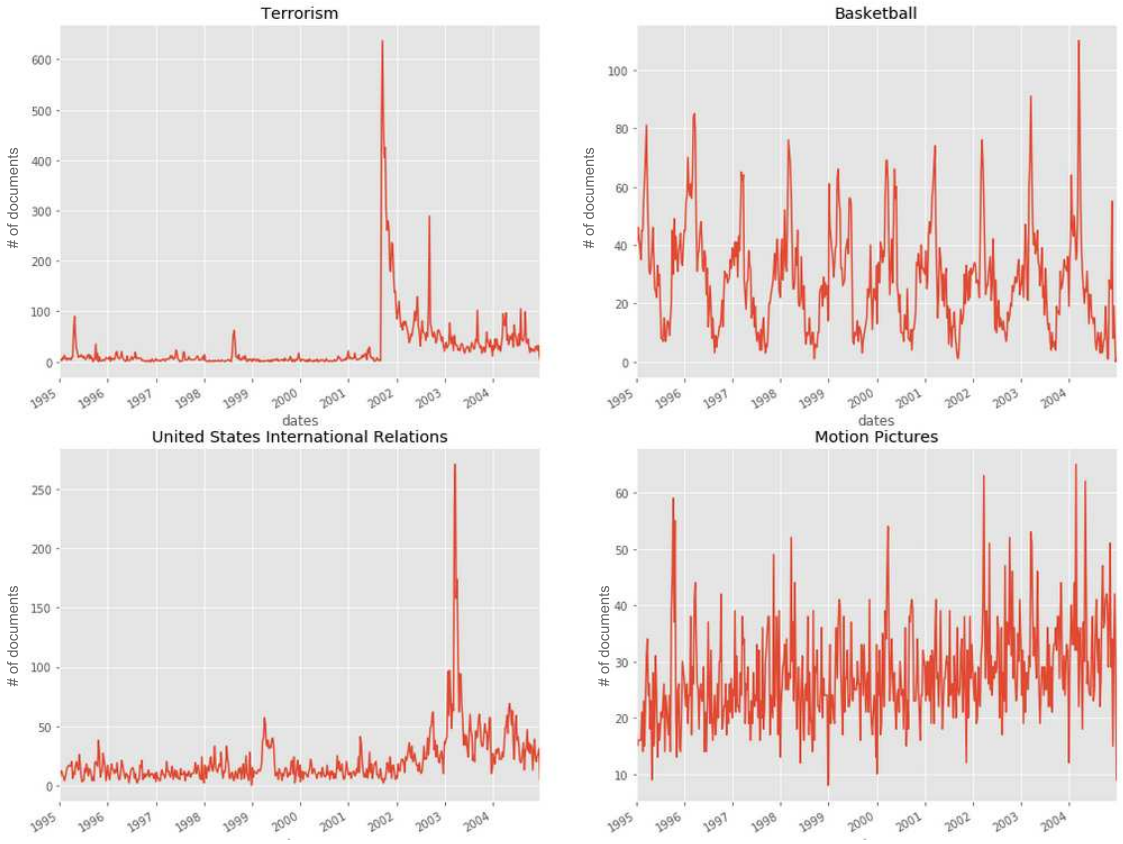


FIGURE 4.1: Évolution du nombre d'articles publiés sous différentes catégories par jour dans le *New York Times*.

à des évènements importants, qui se voient très rapidement sur leur dynamique. Des catégories cycliques sont aussi présentes, comme celles liés au basketball ou football, dont la dynamique va correspondre aux saisons régulières. D'autres catégories ne présentent pas de dynamique facilement identifiable : elles sont soit très bruitées, soit elle présente des volumes d'articles associés très faibles. Les dynamiques de certaines de ces catégories sont présentées sur la Figure 4.1.

Au sein des données courriels EDF, nous choisissons de travailler sur 8 catégories du plan de classement. Il est originellement construit autour de 12 catégories, mais certaines ne sont pas exploitées pour ce travail, car elles sont composées de spams, de courriels vides, de courriels non classés ou de très petit volume. Certaines de ces dynamiques sont présentées sur la Figure 4.2.

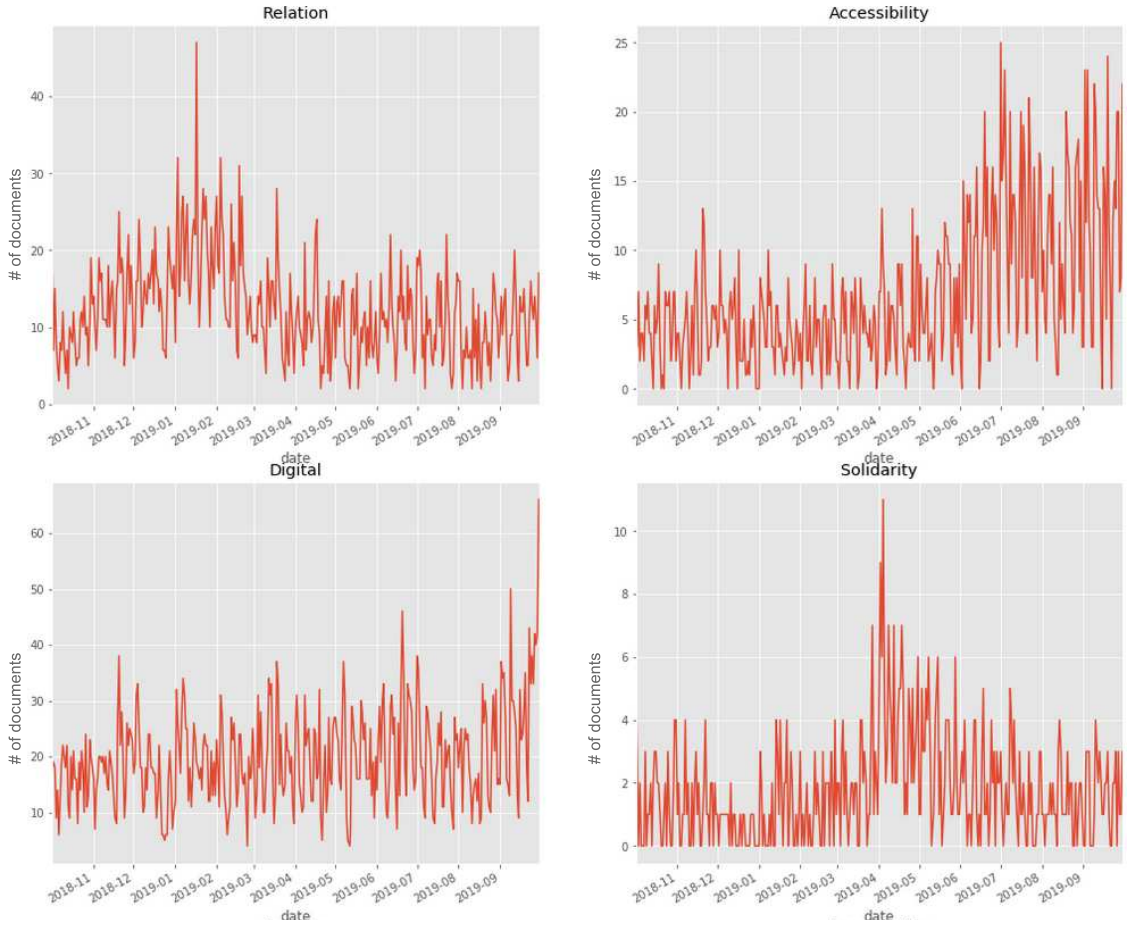


FIGURE 4.2: Évolutions du nombre de courriels classés sous différentes catégories par jour dans le corpus EDF.

4.4.2 Variables exogènes

Dans la section précédente, nous avons présenté des modèles de prédictions de séries temporelles basés sur une ou plusieurs variables. Dans le cas de la prédiction univariée, nous l'avons dit, nous prédisons la valeur de y_{t+1} seulement à partir des valeurs passées de $[y_1, y_2, ..., y_t]$. Dans le deuxième cas, nous nous basons non plus sur le signal passé, mais sur un certain nombre de signaux observés dans le contenu textuel de nos données. Nous présentons, dans cette section, les différentes variables qui ont été étudiées.

Comme nous l'avons vu en introduction, lorsque nous étudions de données textuelles, un des traitements les plus fréquents consiste à observer l'évolution de la

fréquence des mots. Au lieu de compter la fréquence des mots dans chaque document, comme pour construire une matrice TFxIDF, nous comptons simplement le nombre de mots à chaque instant, c'est-à-dire dans tous les documents d'un instant.

En plus de considérer seulement l'évolution de la fréquence brute, nous explorons aussi ses dynamiques physiques comme cela a été fait dans les travaux de [2]. Ces mesures physiques correspondent à la vitesse et à l'accélération d'un signal observées $x(t)$, elles sont définies comme :

$$\omega_{\Delta T}(t) = \sum_{t_i<t} X_i . \frac{exp((t_i - t)/\Delta T)}{\Delta T}$$

$$a(t) = \frac{\omega_{\Delta T_2}(t) - \omega_{\Delta T_1}(t)}{\Delta T_1 - \Delta T_2}$$

où X_i correspond à la fréquence d'un mot dans le i-ème instant, t_i est la date de l'instant et Δ_t la taille de la fenêtre de calcul qui permet de lisser les courbes. Pour capturer le changement dans la vitesse, l'accélération est donc définie comme la différence entre des vitesses avec des fenêtres de tailles différentes. Un exemple de ces types de signaux est illustré sur la Figure 4.3.

En plus de ces mesures physiques, nous choisissons aussi de surveiller les dynamiques de ces fréquences du point de vue de leur périodicité. En effet, nous avons remarqué que certaines catégories, et donc certains mots, présentent des dynamiques périodiques à la semaine, au mois ou à l'année. Sur les données EDF, cette périodicité est connue par les experts métiers et nous a été communiquée. En effet, il est plus courant que les changements de contrats aient lieu pendant l'été, car cela correspond au pic de déménagement. De même, les clients ont plus tendance à envoyer des courriels de réclamations le week-end. Sur les données du *New York Times*, nous avons mené une étude d'autocorrélation sur les signaux des catégories afin de vérifier leur périodicité. L'autocorrélation d'une série temporelle $x(t)$ est simplement la corrélation du signal par rapport à une version décalée dans le temps de lui-même. (rajouter une définition mathématique). Elle permet donc d'estimer le

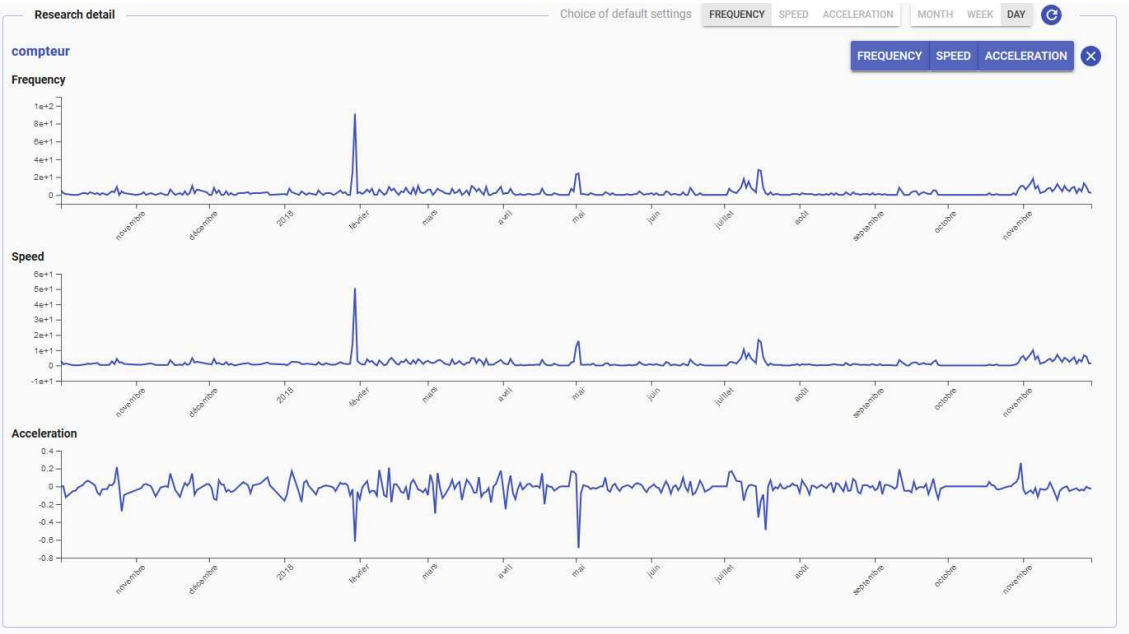


FIGURE 4.3: Exemple de signaux de fréquence brute, de vitesse et d'accélération pour le terme "Compteur" dans le jeu de données EDF.

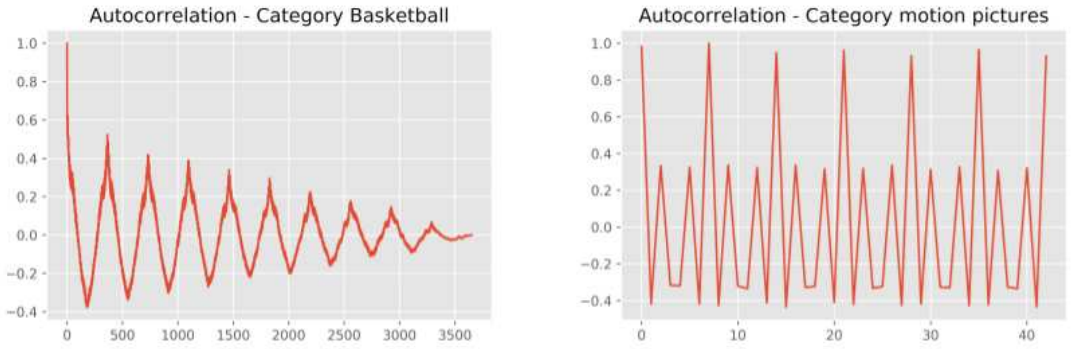


FIGURE 4.4: Autocorrélation pour une catégorie avec une périodicité annuelle (gauche) et hebdomadaire (droite).

nombre de périodes d'un signal sur un temps donné. Nous représentons, sur la Figure 4.4, l'autocorrélation pour 2 catégories du *New York Times*. Sachant que, dans notre cas d'application, un instant correspond à une journée, nous voyons que la catégorie *Basketball* présente une périodicité annuelle : en comptant le nombre de sommets, on retrouve 10 périodes sur 3650 instants. Aussi, nous retrouvons la périodicité hebdomadaire de la catégorie *Motion Pictures* : nous comptons 7 périodes sur 49 instants.

Avec cette idée en tête, nous développons des signaux de retards par rapport aux signaux originaux de fréquences. Ces signaux se définissent sous la forme :

$$l_p(t) = x(t) - x(t-p)$$

Les retards à $p = 365$, $p = 30$ et $p = 7$ permettent de lisser les signaux de fréquences par rapport à leurs périodicités annuelles, mensuelles et hebdomadaires.

En plus des fréquences de simples mots, nous choisissons d'étudier la dynamique dans les données textuelles grâce à un autre indicateur : les thématiques construites de manières non supervisées.

Comme nous l'avons vu dans les chapitres précédents, l'algorithme de *Latent Dirichlet Allocation* (LDA) [4] permet de construire des thématiques probabilistes automatiquement en observant les co-occurrences entre les mots dans les documents. Lorsque nous utilisions, auparavant, cet algorithme, nous nous basions sur les distributions de probabilités sur le vocabulaire ϕ_z, z étant l'indice de la thématique. Nous utilisons, ici, la distribution des thématiques sur les documents θ^d. Dans ce chapitre, nous voulons observer l'évolution des thématiques dans nos ensembles de documents dans le temps. Pour cela, nous entraînons un modèle LDA sur un ensemble d'entraînement $D_{train} = D_0, D_1, ...D_{t_c}$ et Z thématiques. Nous avons $|D_t|$ qui représente le nombre de documents publiés à l'instant t. Nous formons donc des signaux qui correspondent à la fréquence, par pas de temps, des thématiques dans nos corpus. Pour chaque document arrivant dans le corpus de manière continue, nous estimons le mélange de thématique θ^d qui le compose et nous formons donc un signal de type :

$$[k_{1,i}, ..., k_{t,i}, ..., k_{T,i}] \text{ avec } k_{t,i} = \frac{1}{|D_t|} \sum_{j=1}^{|D_t|} p(z_i|d_{j,t})$$

où z_i est la i-ème thématique et d_j un document dans D_t

Basketball	Justice	College Sports	Cinema	Art	US President
Game	Case	Tournament	Film	Work	State
Point	Court	College	Movie	Show	United
Team	Lawyer	Connecticut	Director	Art	American
Knicks	Judge	State	Hollywood	Artist	Clinton
Second	Trial	National	Star	Image	President
Topic 21	**Topic 12**	**Topic 7**	**Topic 29**	**Topic 6**	**Topic 17**

TABLE 4.1: Mots les plus probables pour certaines thématiques du *New York Times*.

Accessibilité	Consommation	Communication	Factures	Contrat
Fil	Équipe	Pièce	Compte	Contrat
Consommation	e.quilibre	Jointe	Suite	Demande
Accès	Consommation	Trouver	Facture	Service
Jour	Électricité	Téléphone	Service	Résiliation
Actualité	Distribuer	Ci-joint	Réclamation	Offres
Topic 8	**Topic 11**	**Topic 7**	**Topic 5**	**Topic 15**

TABLE 4.2: Mots les plus probables pour certaines thématiques des courriels EDF.

Les tableaux 4.1 et 4.2 montrent les 5 mots les plus probables pour certaines thématiques construites via un modèle LDA sur les jeux de données du *New York Times* et des courriels EDF. Les titres donnés aux thématiques ont été choisis manuellement afin d'illustrer leur contenu dans ce manuscrit. Pour la plupart, on retrouve des thématiques non supervisées qui ressemblent aux catégories construites manuellement par les annotateurs du *New York Times* et les experts métiers EDF. Sur les Figures 4.5 et 4.6, nous voyons l'évolution des fréquences de ces thématiques dans le temps. Les thématiques sont donc décrites à partir des mots qui les composent, mais on peut aussi extraire de l'information de leurs dynamiques temporelles. Par exemple sur la Figure 4.5, nous remarquons un changement de dynamique pour la thématique 17 ainsi que des dynamiques cycliques pour les thématiques 7 et 21. Nous verrons dans les sections suivants que ces dynamiques peuvent être utiles pour prédire l'évolution des catégories des plans de classements.

En plus de ces signaux de thématiques, nous voulons observer comment la co-occurrence entre ces thématiques varie dans le temps. Nous comptons comme une

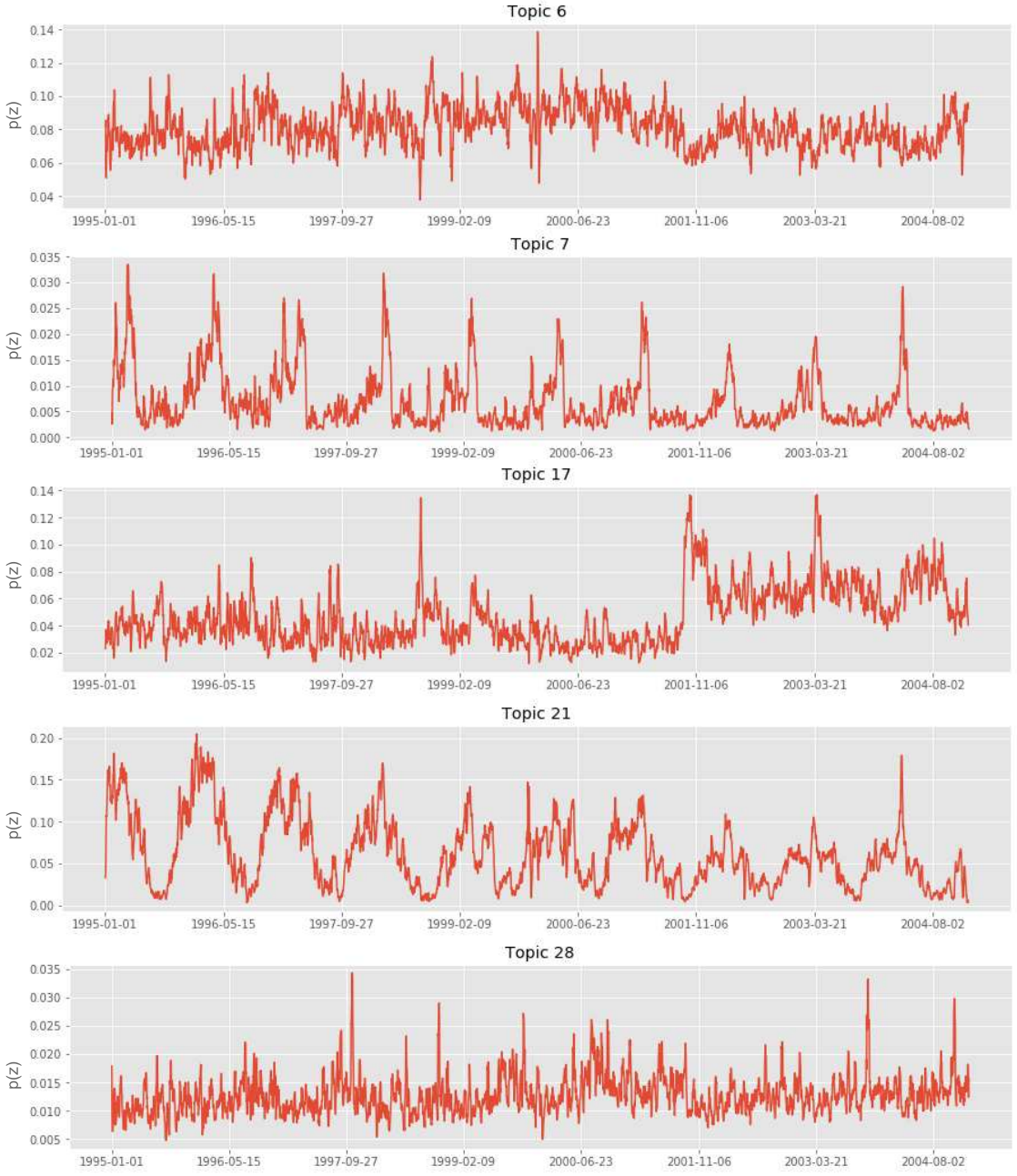


FIGURE 4.5: Évolution de la fréquence de certaines thématiques du *New York Times*.

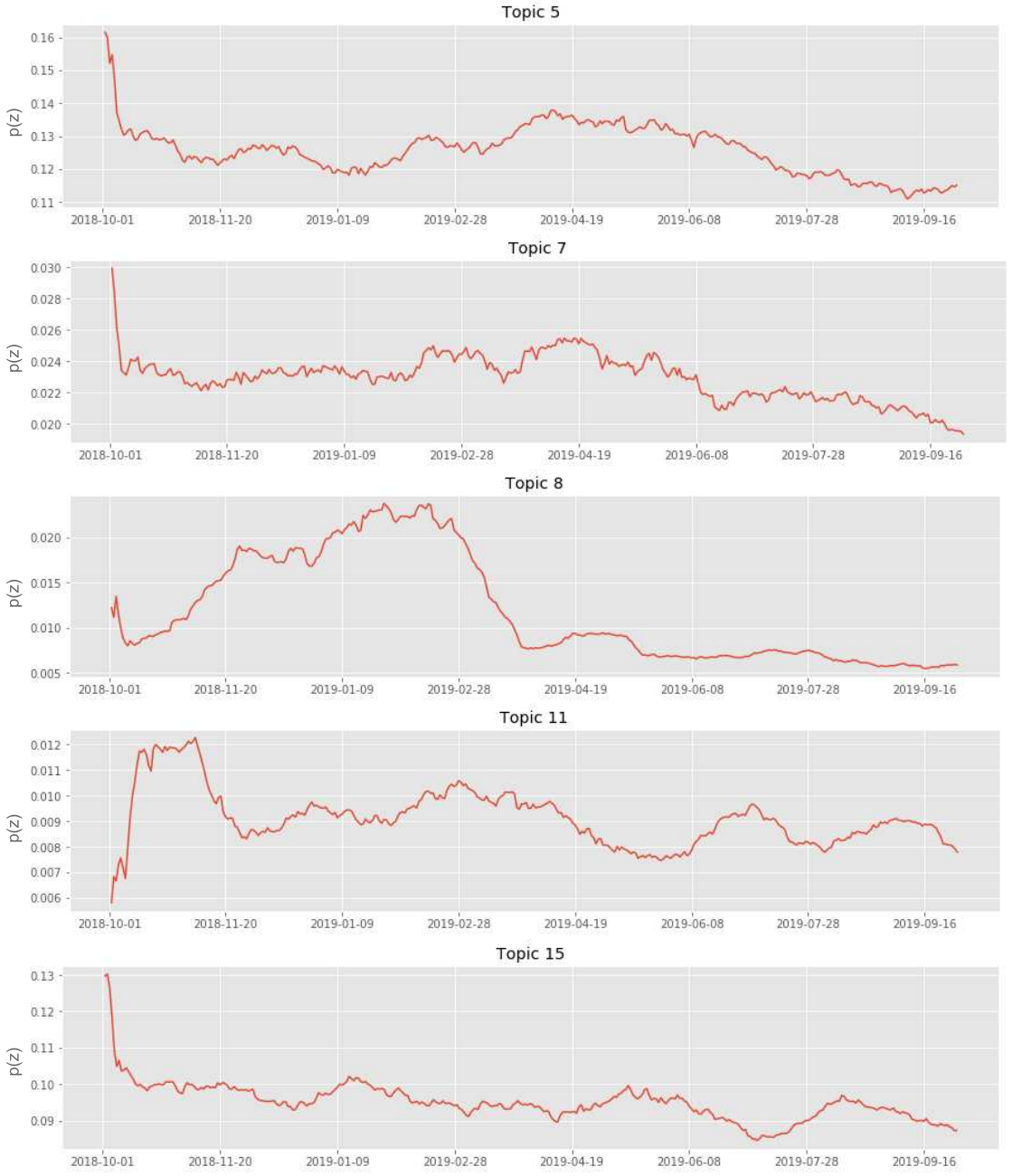


FIGURE 4.6: Évolution de la fréquence de certaines thématiques des courriels EDF.

co-occurrence dans un document, le fait que deux thématiques soient présentes à un ratio élevé. Une thématique z_i est considérée comme présente dans un document si $p(z_i|d) > 0.1$. Nous avons donc des signaux $c(z_{i,j})$ représentant l'évolution des co-occurrences entres des thématiques z_i et z_j.

Sur les signaux $k(t)$ et $c(t)$ représentant les thématiques et leurs co-occurrences, nous développons les mêmes dérivés que pour les fréquences de mots. C'est-à-dire, l'utilisation de la vitesse et de l'accélération ainsi que le développement de retards annuel, mensuel et hebdomadaire.

Une fois que toutes ces variables ont été développées, nous devons mesurer leur importance pour le système de prédiction.

4.4.3 Sélection et importance des variables

Dans la section précédente, nous avons décrit quelles variables nous utilisons pour prédire l'évolution du signal de fréquence d'une catégorie de notre plan de classement. Nous analysons, ici, quelles sont les variables les plus intéressantes pour l'algorithme de prédiction et comment nous les choisissons.

Nous avons commencé par décrire des signaux construits autour de la fréquence des mots. Comme nous voulons un algorithme final rapide, et que notre but n'est pas forcément d'avoir le modèle de prédiction parfait, nous ne pouvons pas nous baser sur des signaux basés sur l'ensemble du vocabulaire : sa taille est de plusieurs milliers de mots. Nous développons donc un système automatique pour choisir automatiquement les mots sur lesquels nous allons nous baser pour faire notre prédiction. Comme dans le chapitre précédent, notre but est d'extraire les mots les plus descriptifs de chaque catégorie. Avec la même idée en tête, nous résolvons un problème de classification avec un classifieur de type Naïve Bayes sur une matrice TFxIDF construite sur le début de nos données (de 0 à t_c). Nous n'avons pas pour but d'obtenir le meilleur classifieur possible, mais plutôt un satisfaisant, qui nous permet d'extraire

Terrorism	Art	Motion Pictures	Basketball	US Itl. Relations
attack	art	film	game	clinton
terrorist	museum	movie	knicks	united
state	painting	actor	team	state
official	artist	director	point	president
federal	work	hollywood	player	american
american	gallery	directed	basketball	official
people	show	story	season	nato
united	exhibition	character	net	palestinian
bombing	sculpture	theater	coach	administration
terrorism	new	festival	play	china

TABLE 4.3: Mots les plus discriminants pour certaines catégories du *New York Times*

les variables les plus discriminantes pour la classification. Sur le *New York Times* et les courriels EDF, nous obtenons, respectivement, 87% et 63% de précision. Ces résultats sont assez bons pour le *New York Times* mais assez décevants pour les données EDF, qui sont plus spécifiques. En effet, le vocabulaire employé dans les catégories des courriels EDF n'est pas très spécifique : il y a beaucoup de superposition de vocabulaire entre les catégories. Cela rend l'analyse plus compliquée. Nous décidons donc de nous remettre à l'expertise interne pour choisir les mots les plus discriminants. Les mots les plus discriminants pour certaines catégories du *New York Times* sont illustrés dans le tableau 4.3. Nous choisissons d'intégrer les signaux des 10 mots les plus discriminants de chaque catégorie pour aider la prédiction.

Une fois ces mots choisis, nous pouvons effectuer la prédiction avec notre modèle de type *Random Forest*. Ce modèle nous permet facilement d'observer les variables les plus importantes à la prédiction. Sur les tableaux 4.4 et 4.5, nous voyons les variables les plus importantes pour, respectivement, les catégories du *New York Times* et des courriels EDF.

Pour la plupart des catégories surveillées, la fréquence de certains mots est la variable la plus importante pour la prévision. Pour le *New York Times*, les 6 variables les plus importantes de la catégorie *Terrorism* correspondent à des fréquences de mots, surtout sur les mots "Bombing" et "Federal". En effet, la fréquence brute

apporte de l'information et l'utilisation des retards aide à la prévision. Ces retards portent l'information sur le changement de la périodicité et semblent être utiles lorsque des pics importants sont présents dans les données. Dans le même sens, la catégorie *Digital* est entièrement décrite grâce à la fréquence brute de 4 mots.

Dans certaines catégories, par exemple sur *Motion Pictures*, *Art* et, plus important encore, sur *US Itl. Relations*, l'information semble être portée par les signaux des thématiques LDA. En effet, pour cette dernière, la variable la plus importante est celle liée à la thématique 17. Lorsque l'on regarde sur le tableau 4.1, nous remarquons que la thématique 17 est celle liée aux Présidents des États-Unis et que ses mots les plus probables sont par rapport à Clinton (alors Président des États-Unis), aux mots "united", "states", "administration", "nato". Ces mots sont des bons descripteurs de la catégorie *US Itl. Relations* du *New York Times*.

Nous observons aussi, pour des catégories telles que *Basketball* où *Relation*, les signaux de co-occurrence de deux thématiques apportent de l'information pour la prévision. Pour la catégorie *Basketball*, les thématiques 7 et 12 sont, respectivement, à propos des Sports Universitaires et de la justice. Bien que la première soit très proche de la catégorie *Basketball* en termes de mots utilisés, la deuxième, où seulement le mot "Court" est facilement interprétable , est plus complexe a analyser. En fait, les thématiques du *Basketball* et de la justice ont souvent été abordées dans les mêmes documents au début des années 2000 à cause de plusieurs affaires judiciaires autour de la NBA (*National Basketball Association*).

Enfin, nous remarquons que, parmi toutes les variables importantes à la prédiction du signal de chaque catégorie, nous ne retrouvons jamais les signaux dérivés de mesures physiques : la vitesse et l'accélération. Ces observations n'aident absolument pas le modèle à bien prédire l'évolution du signal. C'est le cas aussi pour les retards mensuels : nous n'avions pas connaissance, avant cette analyse, de périodicité mensuelle dans nos données. Le fait que ces retards $p = 30$ n'aident pas à la prédiction confirme cette hypothèse. Afin de ne pas complexifier le modèle de prédiction avec

Terrorism	Basketball	Motion Pictures	Art	US Itl. Relations
bombing	game	film	art	Topic 17
bombing_365	basketball	directed	gallery	Clinton
federal_365	Topic 21	Topic 29	Topic 6	united
federal	Cooc 12 & 7	Topic 1	painting_365	united_365
federal_7	team	movie_365	gallery_365	president_365
terrorist	coach	story	art_365	state_365

TABLE 4.4: Variables les plus importantes pour la prédiction pour certaines catégories du *New York Times*. "_365" et "_7" indiquent l'utilisation des retards $p = 365$ et $p = 7$. "Cooc" indique le signal de co-occurrence de certaines thématiques.

Relation	Accessibility	Digital	Solidarity
connecter	téléphone	mail	social
Topic 8	joindre	site	Topic 5
Cooc Topics 11 & 6	numéro	numéro	Topic 15
compte	Topic 7	internet	assistant

TABLE 4.5: Variables les plus importantes pour la prédiction pour certaines catégories des courriels EDF. "Cooc" indique le signal de co-occurrence de certaines thématiques.

des variables qui ne lui sont pas utiles, nous décidons d'enlever les dérivées de vitesses, d'accélération et ces retards mensuels. Toutes les autres variables sont conservées dans le *Random Forest* dont on va pouvoir maintenant évaluer l'efficacité.

4.5 Résultats

Dans ce travail, nos objectifs principaux étaient d'apprendre la dynamique attendue des catégories d'un plan de classement et de lancer des alertes lorsque nous estimons que ces dynamiques deviennent anormales. Nous devons aussi être particulièrement réactifs, mais pas sensibles aux valeurs extrêmes. Nous allons présenter les résultats dans cette section, ils permettront de conclure si nous avons atteint nos objectifs.

4.5.1 Sélection de la méthode de prédiction

En premier lieu, nous devons choisir notre méthode de prédiction. Nous avons présenté dans la section 4.3, différentes approches : deux méthodes univariées classiques et une méthode de type *Random Forest* basée sur des variables exogènes que nous avons illustrées dans la section précédente.

Notre principale hypothèse était que les variables exogènes observées dans le texte sont utiles pour prédire l'évolution de notre signal (le volume de documents classifiés dans une catégorie). Pour justifier cette hypothèse, nous comparons l'erreur moyenne de prédiction entre les trois modèles en termes de *Root Mean Squre Error* (RMSE) :

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(\hat{y}_i - y_i)^2}{N}}$$

avec $\hat{y}_i$ étant la valeur prédite de la série temporelle au moment i, y_i la vraie valeur et N la taille de la série temporelle.

Comme nous avons un modèle de prédiction par catégorie, nous calculons la moyenne des RMSE sur le jeu de données complet. Cette moyenne pourrait camoufler des comportements différents entre les signaux de chaque catégorie comme ils ont des volumes différents. Nous avons observé que les différences en RMSE entre les modèles sont plutôt proportionnelles.

Les résultats sont présentés dans le tableau 4.6. Nous observons, en général, une meilleure prédiction lorsque le modèle de type *Random Forest* est utilisé. Il est important de noter que la combinaison optimale d'hyperparamètres a été utilisée pour chaque modèle après plusieurs expérimentations. Par exemple, le *Random Forest* comporte un hyperparamètre n qui correspond au nombre d'arbres dans la forêt. Ici, la valeur $n = 500$ a été choisie après $k = 10$ validation croisée. Comme le *Random Forest* est le meilleur modèle de prédiction, nous présenterons les résultats

Algorithm	Mean RMSE on NYTAC	Mean RMSE on EDF
KNN-Prediction	4.15 ± 1.22	13.08 ± 1.88
ARIMA	3.97 ± 1.08	10.41 ± 1.52
Random-Forest	2.41 ± 0.82	7.87 ± 1.36

TABLE 4.6: RMSE de différents algorithmes de prédictions. KNN et ARIMA sont univariées et le *Random-Forest* utilise les variables exogènes décrites dans la section 4.4.2.

d'alertes lancées avec ce modèle et nous le comparerons avec d'autres méthodes de la littérature.

4.5.2 Baselines

Afin d'évaluer les performances de notre approche, nous devons la comparer avec des méthodes existantes de la littérature. Nous choisissons deux algorithmes, que nous dénommerons ici nos "baselines". La première est adaptée de [1] où une méthode pour détecter et suivre de nouvelles thématiques dans le temps est présentée. Cette méthode est basée sur la simple statistique de TFxIDF qu'on a présentée dans les chapitres précédents. Cette statistique est surveillée et permet de lancer des alertes lorsque sa valeur passe au-dessus d'un seuil prédéfini. Le deuxième algorithme est [2] que l'on a décrit dans le chapitre précédent. Il est, lui aussi, basé sur une valeur de seuil prédéfinie.

Le modèle CDPred lève des alertes non pas sur des mots, comme les deux baselines choisis, mais sur des signaux de catégories. Nous devons donc ajouter une étape de sélection des alertes qui correspondent à des mots clairement liés à notre catégorie. Pour cela, nous sélectionnons 100 mots par catégorie avec la même méthode présentée en section 4.4.3. Les mots sélectionnés semblent être discriminants aussi bien qualitativement (Tableau 4.3) que quantitativement (Tableau 4.4).

Dans nos deux baselines TopicSkecth [2] et TF-IDF [1], les valeurs des seuils sont déterminées manuellement. Nous avons réalisé plusieurs expériences avec des valeurs différentes et nous présentons les résultats optimaux en termes de nombre d'alertes lancées. C'est-à-dire que nous minimisons le ratio entre le nombre d'alertes intéressantes et le nombre de fausses alertes. Les alertes lancées par ces baselines sont illustrées dans la figure 4.7.

Sur cette figure, nous voyons clairement que les alertes lancées par [1] sont placées assez aléatoirement chronologiquement parlant par rapport au signal observé. Nous notons un seulement un grand nombre d'alertes concentrées autour du pic de la catégorie *US Itl. Relations* mais au-delà de ça, nous pouvons conclure que cette méthode ne semble pas du tout efficace pour résoudre notre tâche.

L'algorithme TopicSketch [2] semble lancer des fausses alertes dans des catégories telles que *Theater* et *Art* bien que celles-ci n'aient pas de dynamiques anormales à première vue. Les alertes lancées sur les catégories *Basketball* et *US Itl. Relations* ont du sens par rapport aux dynamiques de ces signaux. En effet, dans *Basketball*, TopicSketch semble lancer des alertes à chaque pic (partie haute de la période), c'est-à-dire à chaque fois que la saison de NBA bat son plein. Du côté de *US Itl. Relations*, cette méthode lance des alertes à chaque petit pic, c'est-à-dire à chaque valeur un peu extrême. Bien qu'elle semble être une bonne méthode pour détecter des changements brefs et abrupts dans des données textuelles, elle ne parvient pas à détecter des changements sur le long terme. Autrement dit, elle n'a pas de mémoire de la dynamique normale du signal surveillée.

4.5.3 Alertes lancées par CDPred sur le *New York Times*.

La figure 4.8 illustre les alertes lancées par notre modèle CDPred sur quatre catégories du *New York Times* : *Terrorism*, *US International Relations*, *Basketball* et *Politics and Government*. Nous avons choisi de nous concentrer sur ces catégories, car

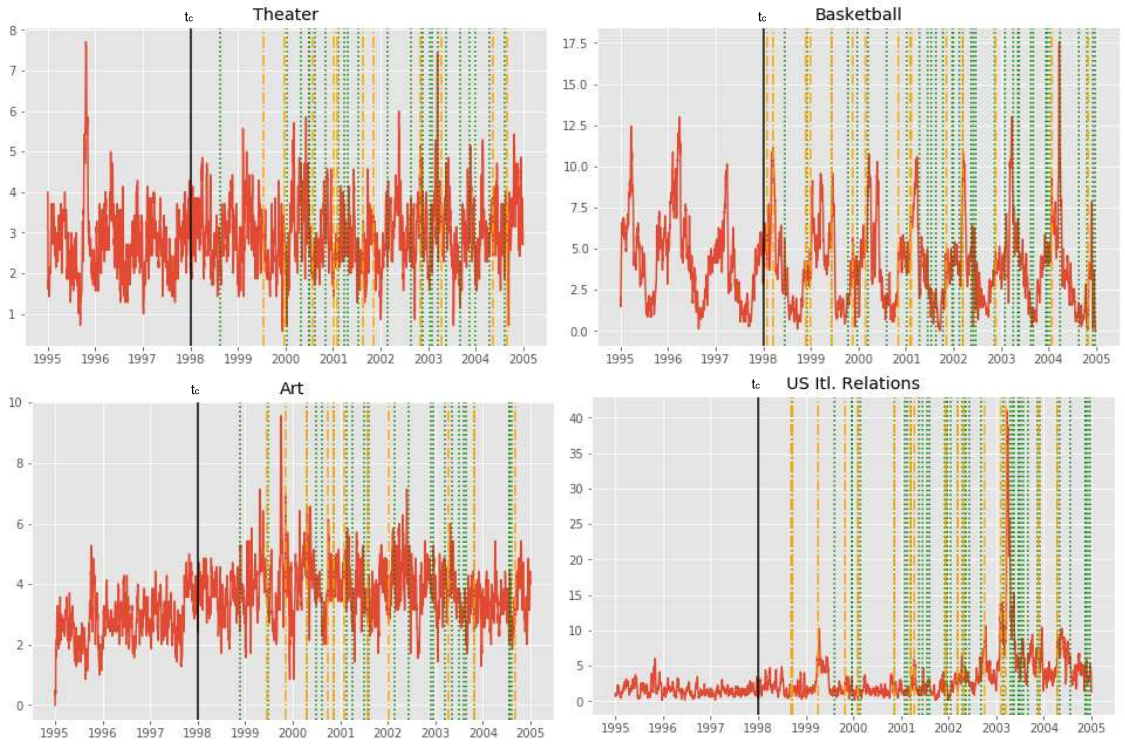


FIGURE 4.7: Alertes lancées par nos baselines TF-IDF ([1]) (vert pointillé) et TopicSketch ([2]) (orange) sur 4 catégories du *New York Times*.

elles présentent une dynamique temporelle intéressante. Les deux premières (*Terrorism* et *US International Relations*) sont liées à des évènements importants qui ont considérablement impacté la ligne éditoriale du journal. Ces évènements sont facilement identifiables en regardant l'aspect temporel de ces catégories. Nous avons identifié quatre pics qui peuvent être considérés comme des évolutions inattendues de la catégorie :

— **Premier pic, 21 août 1998** : le bombardement par l'armée américaine de cibles en Afghanistan et au Soudan a pour effet le renforcement de la sécurité dans les grandes villes américaines et particulièrement à New York.

— **Deuxième pic, 11 septembre 2001** : attaques terroristes sur le *World Trade Center*.

— **Troisième pic, mars et avril 1999** : entrée en guerre des forces de l'OTAN au Kosovo.

— **Quatrième pic, mars 2003**, invasion de l'Irak par l'armée américaine.

Pour les pics 2 et 4, nous observons que les alertes lancées par CDPred sont bien groupées autour de ces dates. Pour *US International Relations*, nous avons, par rapport aux alertes lancées par nos baselines dans la figure 4.7, des alertes plus cohérentes avec les évènements réels. Pour le troisième pic, même s'il est beaucoup moins important, en termes de volume, que le pic numéro 4, il a considérablement modifié la proportion d'articles associés à la catégorie *US International Relations* durant environ deux mois. Nous observons que notre système a correctement détecté ce pic assez tôt. Le premier pic est important en termes de volume dans la catégorie *Terrorism* mais ne provoque pas de changement à long terme dans la classification : c'est une valeur anormale et ce pic ne dure qu'un seul jour. On pourrait donc le classer comme fausse alerte, mais si le volume associé y est très anormal.

Les catégories en rapport avec *Basketball* et *Politics and Government* sont aussi intéressantes à surveiller. Le signal de *Basketball* est cyclique, vu qu'il suit les saisons de NBA et NCAA aux États-Unis. Nous voyons, dans la figure 4.7 que notre baseline TopicSketch lance une alerte à chaque pic bien qu'il apparaisse chaque année. Notre modèle CDPred, ne lance, au contraire, aucune alerte sur cette catégorie. Nous pouvons donc conclure que notre algorithme a correctement appris sa dynamique normale. Le signal associé à catégorie *Politics and Government* semble plutôt constant dans le temps bien qu'il soit très bruité. En l'observant, il est difficile de constater qu'un changement important est apparu. Notre modèle CDPred a lancé une seule alerte à la date suivante : 2 janvier 2002. Cette date correspond à l'inauguration du nouveau maire de New York Michael Bloomberg et a bien fait l'objet d'une anomalie importante dans le nombre d'articles associé à la catégorie en question. Bien que cette anomalie soit particulièrement grande, cette alerte ne constitue pas un changement de dynamique.

Sur la figure 4.9, nous présentons d'autres catégories annotées dans le *New York Times* : *Art*, *Theater*, *Automobiles* et *Dancing*. Ces catégories ont toutes un signal

Category	Event	t_{change}	Δ_{tfidf} [1]	Δ_{Sketch} [2]	Δ_{CDPred}
Terrorism	Security tightening	1998-08-21	33	19	**0**
Terrorism	*WTC* terrorist attacks	2001-09-11	3	1	**1**
US Itl. Relations	Kosovo War	1999-03-23	251	13	**10**
US Itl. Relations	Iraq War	2003-03-20	12	5	**0**

TABLE 4.7: Valeur absolue du Δ (en jour) entre la date réelle du changement et la date de l'alerte.

assez constant, mais bruité : pas de changements de dynamiques sont observés. Nous voyons que, par rapport aux alertes lancées par nos baselines sur le figure 4.7, notre modèle CDPred ne lance pas ou peu d'alertes. Seulement quatre sont considérées comme de fausses alertes sur ces quatre catégories. Nous comparons, dans le tableau 4.8, le nombre d'alertes lancées sur ces catégories par nos baselines et par notre modèle CDPred. Nous voyons que notre modèle est bien moins enclin à lancer de fausses alertes.

Enfin, un de nos objectifs de départ était de construire un modèle réactif : c'est-à-dire qui prend peu de temps à identifier un réel changement de dynamique dans une catégorie. Nous mesurons cet aspect dans le tableau 4.7 par rapport aux quatre pics que nous avons étudiés précédemment. Nous mesurons la différence entre la date de l'évènement et la date de la première alerte lancée par notre modèle après cet évènement. Nous avons donc $\Delta_{method} = t_{change} - t_{alert}$. Nous remarquons tout d'abord que pour les gros changements comme les attaques du 11 septembre et le début de la guerre en Irak, les trois modèles sont capables de lancer des alertes assez rapidement. CDPred est quand même plus réactif. Par rapport à l'entrée en guerre de l'OTAN au Kosovo, nous voyons que, bien que cet évènement a un effet à long terme sur la catégorie, notre modèle est plus rapide que TopicSketch. La baseline TF-IDF ne détecte jamais ce changement.

Dans cette section, nous avons observé les résultats de notre modèle CDPred sur différentes tâches. Bien qu'il soit difficile d'évaluer ce type d'approche, nous avons pu conclure que CDPred est efficace pour détecter des pics ayant des effets à long terme sur les dynamiques des catégories. Nous avons vu que CDPred est moins enclin a

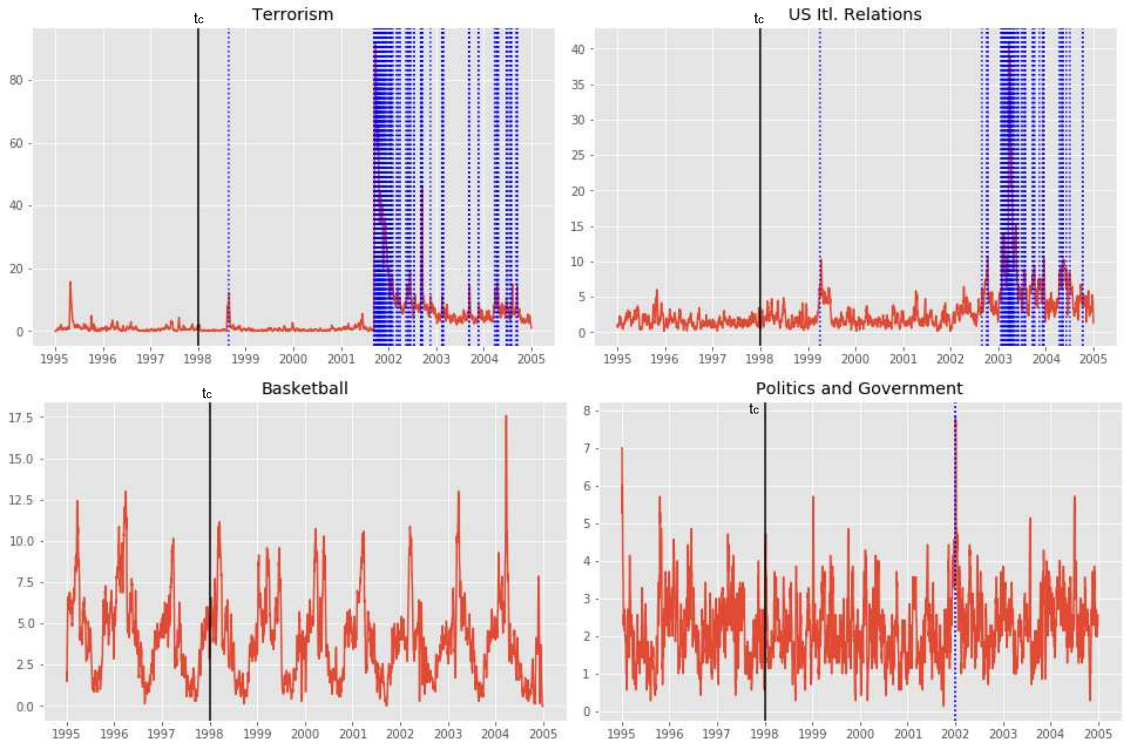


FIGURE 4.8: Alertes lancées par CDPred (bleu pointillé) sur 4 catégories de NYTAC.

Categories	N_{Sketch} [2]	N_{tfidf} [1]	N_{CDPred}
Art	10	32	0
Murders	6	10	0
Theater	11	24	1
Travel	4	18	1
Dancing	2	8	1
Automobiles	5	10	2

TABLE 4.8: Nombre de fausses alertes N lancées par chaque modèle pour différentes catégories constantes.

lancer des alertes sur des pics très courts (qui ont lieu que sur une journée) bien que cela arrive sur des volumes très importants. CDPred lance moins de fausses alertes sur des signaux qui ne changent pas de dynamiques et semble être plus réactif que nos baselines sur les vrais changements. Enfin, CDPred présente l'avantage d'avoir une mémoire à long terme des dynamiques des signaux qu'il surveille. Cela permet la surveillance de signaux cyclique sans lancer d'alertes.

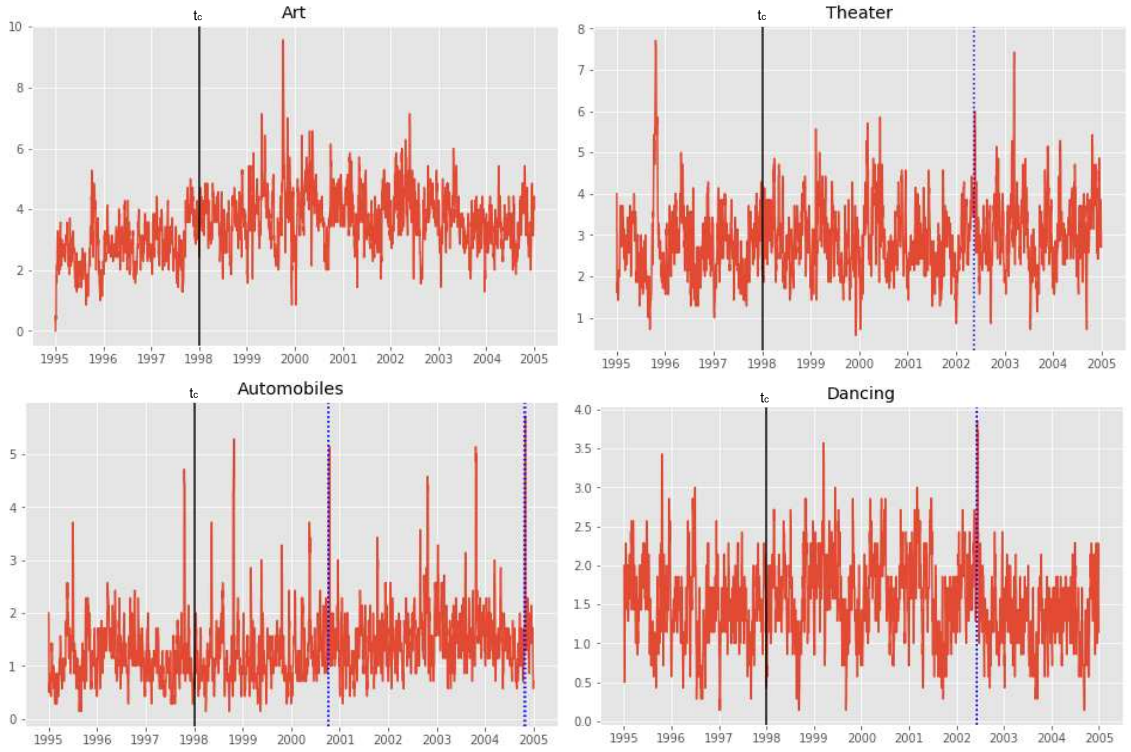


FIGURE 4.9: Fausses alertes lancées par CDPred (bleu pointillé) sur 4 catégories de NYTAC.

4.5.4 Applications industrielles sur les données EDF.

La motivation principale de ce travail, et de ce manuscrit de thèse en général, est d'être capable de détecter au plus tôt des changements de dynamiques. Nous avons donc testé notre approche CDPred sur des données de courriels clients EDF. Ce jeu de données a été décrit dans la section 4.4.1. Nous présentons ici quelques résultats obtenus sur différentes catégories de ce dernier. Les catégories sont les suivantes : "Relation avec EDF", "Accessibilité", "Digitale" et "Solidarité". Nous observons sur la figure 4.5.4 que des alertes sont lancées sur ces catégories.

— Pour la catégorie "Accessibilité", nous voyons que le changement de dynamique qui a lieu autour du mois de juin 2019 est clairement détecté par notre modèle et que beaucoup d'alertes sont toujours lancées après ce changement. Cela est dû au fait que la nouvelle dynamique n'est pas encore enregistrée

comme dynamique normale, car elle n'est pas encore entrée dans la partie entraînement du modèle de prédiction.

— Pour la catégorie "Solidarité", nous observons un pic qui dure à peu près deux mois (entre avril et juin 2019). Il est clairement détecté par notre modèle. Il est intéressant de noter que ce changement est détecté même si le volume initial de documents classés dans cette catégorie est très faible : un ou deux documents par jour.

— Pour la catégorie "Relation avec EDF", le changement de dynamique semble apparaître au début de notre observation. il est même en partie pris en compte dans la partie entraînement du modèle de prédiction. La fin de ce changement est quant à elle bien détectée au cours du mois de janvier 2019.

— Pour la catégorie "Digitale", trois alertes sont lancées, une au cours du mois de juin qui ne semble pas correspondre à un changement de dynamique et deux autour d'un pic important. Ces alertes peuvent être considérées comme fausses. Nous voyons qu'un changement important de dynamique commence à apparaître à la fin de notre observation. Il est légitime de se demander si, avec quelques observations de plus, nous aurions pu être capables de détecter ce changement.

Par rapport aux données du *New York Times*, le jeu de données de courriels EDF contient moins de valeurs extrêmes sur un instant et quelques catégories font l'objet de vrais changements qui doivent être détectés. Nous avons vu que CDPred est capable de détecter ces changements. Comme ces données sont particulièrement sensibles et que les raisons expliquant ces changements pourraient porter atteinte à EDF, nous ne sommes pas en mesure de fournir des explications.

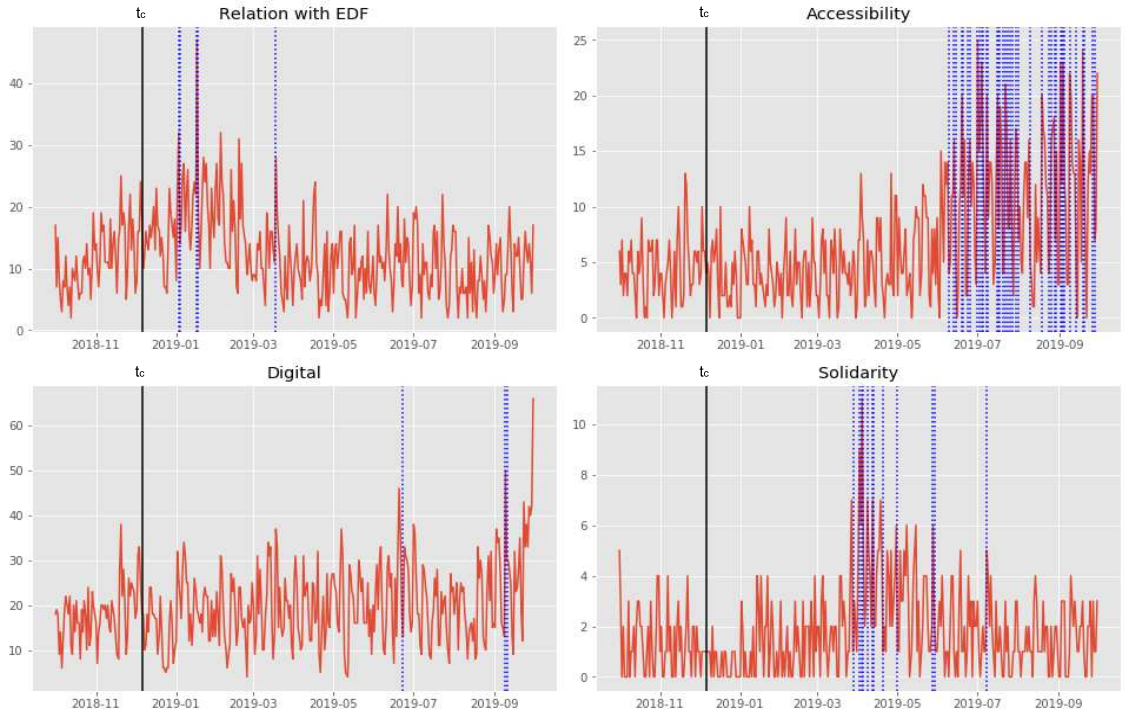


FIGURE 4.10: Alertes lancées par CDPred (bleu pointillé) sur 4 catégories du jeu de données EDF

4.6 Conclusion

Dans ce chapitre, nous avons présenté un algorithme capable de lancer des alertes lorsqu'un changement inattendu est détecté dans un flux de données textuelles. Pour cela, nous avons transformé nos données textuelles en séries temporelles et nous nous sommes basés sur les changements de volumes des catégories d'un plan de classement préétabli. Nous avons basé notre approche sur l'hypothèse qu'un changement inattendu dans un signal est lié à une erreur de prédiction plus importante. Au lieu d'utiliser des méthodes traditionnelles univariées, nous avons choisi d'utiliser les informations contenues dans le texte. En nous basant sur des méthodes classiques d'analyses de données textuelles et de modélisation thématiques, nous avons montré que nous sommes capables d'améliorer la prédiction du signal sans ajouter énormément de complexité. L'objectif principal de ce travail n'était pas d'obtenir un modèle de prédiction parfait, mais plutôt d'en avoir un qui nous permettrait

d'expliquer les prédictions. C'est pourquoi nous avons choisi un modèle de type *Random Forest* qui, grâce à sa grande explicabilité, nous permet de nous raccrocher aux informations textuelles.

Nous avons ensuite lié ce modèle de prédiction à une méthode d'analyse séquentielle traditionnellement utilisée dans la surveillance de processus industriel. En appliquant la méthode CUSUM sur notre erreur de prédiction, nous sommes capables d'évaluer la stabilité et de lancer des alertes dès que cette dernière sort de sa zone de confort. Nous avons démontré que notre modèle est plus performant que d'autres méthodes de la littérature pour détecter des pics dans les données. Comme nous nous basons sur un algorithme de prédiction, notre modèle présente l'avantage de ne pas être sensible aux catégories cycliques. Enfin, CDPred a montré une sensibilité moins importante aux bruits, c'est-à-dire qu'il lance moins de fausses alertes sur des évènements très courts et sur des catégories constantes, mais bruitées.

Nous avons développé un certain nombre de variables tirées du texte de nos documents, mais nous pourrions en imaginer d'autres. En effet, il serait intéressant d'ajouter des informations telles que le nombre de verbes, noms et adjectifs puisque cela devrait être variable d'une catégorie à une autre. Comme développé dans le chapitre précédent, nous pourrions ajouter des informations quant aux mouvements des mots dans des espaces de représentations construites avec Word2Vec ([74]). Comme nous l'avons dit, une partie de nos choix est motivée par l'aspect explicabilité, nous pourrions analyser l'importance de chaque variable pour identifier les mots ou les documents responsables d'une alerte afin que celle-ci ne soit plus liée qu'à un moment. Aussi, une extension possible réside dans le fait de pouvoir valider une alerte lancée par le système. Ceci permettrait de remettre à zéro le système et de ne pas accumuler les alertes liées aux mêmes évènements comme cela a été le cas dans les catégories *Terrorism*, *US Itl. Relations* et *Accessibilité*.

Chapitre 5

Conclusion et perspectives

Pour conclure ce manuscrit de thèse, nous résumons les travaux que nous avons présentés puis nous nous concentrons sur les perspectives de recherche qui s'inscrivent dans la continuité de ces travaux.

5.1 Conclusion

Dans cette thèse, nous avons étudié la détection de nouveauté dans des flux de données textuelles. Nous avions pour objectifs de développer des méthodes qui peuvent être utilisées par l'entreprise EDF.

Nous avons commencé par étudier le concept de nouveauté dans la littérature. Nous avons vu que la nouveauté est étudiée dans tous les domaines et dans tous les types de données, mais elle est définie différemment dans chaque cas d'application. Nous avons énuméré les différentes familles de méthodes de détections qui sont utilisées dans des domaines différents. Après cet état de l'art général, nous nous sommes concentrés sur l'aspect textuel. En effet, les données textuelles ont des caractéristiques spécifiques et leur étude nécessite des algorithmes dédiés. Nous avons vu que, pour ce type de données, la nouveauté pouvait se trouver à plusieurs niveaux de granularité ce qui rend la définition plus complexe. Nous avons défini les tâches que nous voulons résoudre et développé une méthode de comparaison entre différentes approches existantes en unifiant les mesures d'évaluation. Cela nous a permis de mettre en lumière certaines approches plus prometteuses.

Nous avons ensuite exploré l'utilité des méthodes de modélisation thématiques pour la détection de la nouveauté. En développant trois approches sur des cas simples, nous avons vu que ce type de méthode ne permet pas de détecter de la nouveauté complexe, proche de ce qui existait auparavant. Nous nous sommes ensuite concentrés sur les modèles de plongements de mots et nous avons exploité une observation que nous avons faite au niveau des mouvements dans ce type d'espace. Ceci nous a permis de développer le modèle CEND qui se base sur la corrélation entre les mouvements d'un mot dans un espace en grande dimension et la dynamique de sa fréquence. En utilisant des jeux de données annotés par catégorie, nous avons pu simuler l'émergence de thématique et évaluer notre approche sur une tâche de détection de mots. Nous avons montré que notre méthode CEND fonctionne bien sur des articles de presses où le texte est correctement structuré et où les champs

lexicaux sont précis. En revanche, elle n'a pas fonctionné sur les données courriels d'EDF car ces derniers ne sont pas écrits par des spécialistes. Le choix du jeu de données sur lequel nous appliquons ce type de méthode est donc primordial.

Enfin, nous nous sommes intéressés à la surveillance de signaux connus. Ils sont importants, surtout dans un environnement industriel, car EDF a déjà des plans de classements en place qui permettent de catégoriser automatiquement les courriels clients. Des variations importantes dans la dynamique de ces catégories pourraient être un signe de problème ou de changement dans les systèmes de l'entreprise. Nous avons donc développé un modèle CDPred dont l'hypothèse de départ est qu'un changement a lieu sur un signal connu lorsque nous nous trompons dans la prédiction de son évolution. Nous développons donc un modèle de prédiction basé sur des variables exogènes tirées du contenu textuel de nos données. Nous analysons son erreur de prédiction avec une méthode d'analyse séquentielle de type CUSUM. Ce modèle CDPred nous permet de lancer des alertes au plus tôt sur des changements importants dans les dynamiques de classifications.

Ce travail de thèse a été rendu difficile par l'aspect complexe de la tâche que nous voulions résoudre. En effet, nous avons dû conceptualiser cette dernière et définir les objectifs et les attentes. L'aspect temporel des données textuelles a été également très important. En effet, nous avons dû expérimenter avec des jeux de données textuelles temporelles où celles-ci sont associées à des catégories afin de nous permettre de construire une référence à détecter. La construction de cette référence a été un point crucial de ce travail de thèse, car elle nous a permis de comprendre le type d'évolution que nous voulions détecter. C'est en échangeant avec les experts métiers que nous avons pu faire la distinction entre les nouveautés de volume et de structure et ainsi séparer les deux types d'approches : une où nous connaissons ce que nous surveillons et une où nous devons modéliser nous même nos données de la meilleure des façons.

Les différentes structures du langage ont mis en valeur un des enseignements majeurs de ce travail : la manière dont la donnée est pensée à sa création a énormément d'influence sur la performance de la détection. En effet, un article de presse ou une publication scientifique sont écrits avec une grande précision autour des termes et des champs lexicaux choisis alors que des courriels clients ou des réclamations se basent sur un vocabulaire beaucoup plus classique et où les mots sont utilisés dans une grande variété de contexte.

Pour la suite, nous présentons nos perspectives de travaux qui s'inscrivent dans la continuité de cette thèse.

5.2 Perspectives

Dans les conclusions des différents chapitres, nous avons soulevé plusieurs axes d'améliorations. L'un de ces axes concerne le traitement des nouveautés plus complexes. Nous commençons par décrire les cas complexes que nous pourrions traiter. Nous continuons en décrivant les améliorations que nous pouvons imaginer pour un modèle de type CEND. Enfin, nous imaginons des améliorations qui permettraient une meilleure intégration de CDPred dans les systèmes de traitement d'EDF.

5.2.1 Type de nouveautés complexes

Nous avons vu dans l'introduction et dans le chapitre 2 qu'il existait plusieurs types de nouveautés et nous nous sommes principalement intéressés à l'émergence de nouvelles thématiques.

En effet, l'émergence d'une thématique dans le temps ne représente qu'un type de nouveauté. C'est celui qui est le plus important dans une optique d'anticipation des nouveaux problèmes clients chez EDF, mais il ne permet pas de comprendre, à lui seul, l'ensemble des dynamiques d'un jeu de données textuelles. Nous avons vu

que des nouveautés du type “disparition”, “fusion” et “séparation” de thématiques pouvaient apparaître. Des approches thématiques non paramétriques telles que les *Hierarchical Dirichlet Process* [105] pourraient permettre d'effectuer ce type d'observation dans un système similaire à celui développé en section 3.3. La création et la séparation des thématiques dans le temps ressemblent à un processus de restaurant chinois [106] et nous pourrions imaginer une version temporelle d'un modèle thématique hiérarchique développé dans [107].

5.2.2 Représentation contextuelle et justification théorique

Nous avons utilisé, dans le chapitre 3 des modèles de représentations de type Word2Vec et SVD pour représenter le sens des mots dans un espace en grande dimension. Récemment, des approches comme BERT [76] ou ELMO [77] ont permis d'ajouter des informations contextuelles dans la représentation des termes. En effet, un des inconvénients de Word2Vec est que chaque terme possède qu'une seule représentation dans l'espace alors qu'il possède plusieurs sens qui dépendent du contexte autour de celui-ci. Par exemple les phrases “il regarde à travers ses jumelles” et “les jumelles sont dans la même classe” montrent que le mot “jumelles” a plusieurs sens. Les modèles BERT et ELMO produisent des vecteurs dépendant du contexte pour chaque mot. Cela permet de désambiguïser le sens de ces mots. Nous pourrions imaginer une version temporelle de ce type de modèle et nous baser sur les mouvements des mots dans ces espaces pour détecter de la nouveauté.

Dans le chapitre 3, nous avons basé notre modèle sur l'observation des corrélations entre les dynamiques de mouvement et de fréquence. Bien que nous ayons observé cela sur l'ensemble de nos catégories et sur des modèles de type SGNS et SVD, nous n'avons pas de justification théorique expliquant ce comportement. Nous avons aussi des différences de distributions de corrélations entre SGNS et SVD. La corrélation majoritairement positive représente une caractéristique connue de SGNS : plus un mot est utilisé dans un même contexte, plus la norme de son vecteur est grande [95].

La différence de comportement dans les corrélations peut, elle, être expliquée par le fait que la SVD est plus stable et moins biaisée envers les nouvelles observations que SGNS. Cette conclusion est supportée dans, [89] mais mériterait plus d'analyses dans notre cas d'application spécifique.

5.2.3 Groupement automatique de mots et d'alertes

Dans le chapitre 3, nous lançons des alertes sur des mots uniques qui peuvent être signe de l'émergence d'une nouvelle thématique. Nous avons vu que ces mots-là ont tendance à faire partie du champ lexical d'une même thématique et nous avons vu que notre modèle CEND a tendance à les détecter plus souvent : leur comportement est plus marqué. Afin d'améliorer l'intégration de ce type de système dans les structures d'EDF, il faudrait ajouter une étape permettant de grouper automatiquement ces mots afin de diminuer le nombre d'alertes et d'alerter sur des thématiques entières.

Dans le chapitre 4, nous lançons des alertes quand notre erreur de prédiction est grande. Nous avons vu que, lors de grands évènements, notre modèle a tendance a lancer une grande quantité d'alertes à la suite. Afin de, là aussi, faciliter l'intégration dans les systèmes d'EDF, il serait intéressant d'ajouter une possibilité de remettre à zéro le système d'alerte ainsi que l'apprentissage de la dynamique normale une fois qu'une alerte a pu être validée par un superviseur humain.

Bibliographie

[1] James Allan, Victor Lavrenko, Daniella Malin, and Russell Swan. Detections, bounds, and timelines : Umass and tdt-3. In *Proceedings of topic detection and tracking workshop*, pages 167–174. sn, 2000.

[2] Wei Xie, Feida Zhu, Jing Jiang, Ee-Peng Lim, and Ke Wang. Topicsketch : Real-time bursty topic detection from twitter. *IEEE Transactions on Knowledge and Data Engineering*, 28(8) :2216–2229, 2016.

[3] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan) :993–1022, 2003.

[5] Lasse L Molgaard, Jan Larsen, and Cyril Goutte. Temporal analysis of text data using latent variable models. In *2009 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2009.

[6] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[7] Sean Gerrish and David M Blei. A language-based approach to measuring scholarly impact. In *ICML*, volume 10, pages 375–382. Citeseer, 2010.

[8] Lu Ren, David B Dunson, and Lawrence Carin. The dynamic hierarchical dirichlet process. In *Proceedings of the 25th international conference on Machine learning*, pages 824–831, 2008.

[9] Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99 :215–249, 2014.

[10] Ian Soboroff and Donna Harman. Overview of the trec 2003 novelty track. In *TREC*, pages 38–53. Citeseer, 2003.

[11] Julie Beth Lovins. Development of a stemming algorithm. *Mech. Transl. Comput. Linguistics*, 11(1-2) :22–31, 1968.

[12] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 1972.

[13] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185, 2014.

[14] Francois Role and Mohamed Nadif. Handling the impact of low frequency events on co-occurrence based measures of word similarity. In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (KDIR-2011). Scitepress*, pages 218–223, 2011.

[15] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296, 1999.

[16] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[17] Rob J Hyndman and George Athanasopoulos. *Forecasting : principles and practice*. OTexts, 2018.

[18] Evan Sandhaus. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12) :e26752, 2008.

[19] Elina Hiltunen. The future sign and its three dimensions. *Futures*, 40(3) : 247–260, 2008.

[20] Elina Hiltunen et al. *Weak signals in organizational futures learning*. Helsinki School of Economics, 2010.

[21] Osmo Kuusi, Elina Hiltunen, and Hannu Linturi. Heikot tulevaisuussignaalit : Delfoi-tutkimus. *Futura 19 (2000) : 2*, 2000.

[22] P Rossel. Meta framing : The art of putting weak signals in perspective. In *Proceedings of the COSTA22 Conference*, 2007.

[23] Mika Mannermaa. *Heikoista signaaleista vahva tulevaisuus*. Wsoy, 2004.

[24] Markos Markou and Sameer Singh. Novelty detection : a review—part 1 : statistical approaches. *Signal processing*, 83(12) :2481–2497, 2003.

[25] Markos Markou and Sameer Singh. Novelty detection : a review—part 2 : : neural network based approaches. *Signal processing*, 83(12) :2499–2521, 2003.

[26] Clayton Scott and Robert Nowak. A neyman-pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11) :3806–3819, 2005.

[27] Aarti Singh, Robert Nowak, and Jerry Zhu. Unlabeled data : Now it helps, now it doesn't. In *Advances in neural information processing systems*, pages 1513–1520, 2009.

[28] Régis Vert and Jean-Philippe Vert. Consistency and convergence rates of one-class svms and related algorithms. *Journal of Machine Learning Research*, 7 (May) :817–854, 2006.

[29] Ran El-Yaniv and Mordechai Nisenson. Optimal single-class classification strategies. In *Advances in Neural Information Processing Systems*, pages 377–384, 2007.

[30] Lambert Pépin. *Fouille exploratoire de messages publiés sur Twitter pour l'aide à la décision*. PhD thesis, University of Nantes, 2015.

[31] Fabrizio Angiulli and Clara Pizzuti. Fast outlier detection in high dimensional spaces. In *European conference on principles of data mining and knowledge discovery*, pages 15–27. Springer, 2002.

[32] E Knorr and R Ng. Algorithms for mining distance-based outliers in very large databases. In *VLDB Conference*, 1998.

[33] Girish Keshav Palshikar. Distance-based outliers in sequences. In *International Conference on Distributed Computing and Internet Technology*, pages 547–552. Springer, 2005.

[34] Yufeng Kou, Chang-Tien Lu, and Dechang Chen. Spatial weighted outlier detection. In *Proceedings of the 2006 SIAM international conference on data mining*, pages 614–618. SIAM, 2006.

[35] Dongil Kim, Pilsung Kang, Sungzoon Cho, Hyoung-joo Lee, and Seungyong Doh. Machine learning-based novelty detection for faulty wafer detection in semiconductor manufacturing. *Expert Systems with Applications*, 39(4) : 4075–4083, 2012.

[36] Ashok N Srivastava and Brett Zane-Ulman. Discovering recurring anomalies in text reports regarding complex space systems. In *2005 IEEE aerospace conference*, pages 3853–3862. IEEE, 2005.

[37] Ashok N Srivastava. Enabling the discovery of recurring anomalies in aerospace problem reports using high-dimensional clustering techniques. In *2006 IEEE Aerospace Conference*, pages 17–pp. IEEE, 2006.

[38] James C Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media, 2013.

[39] Raghuram Krishnapuram and James M Keller. A possibilistic approach to clustering. *IEEE transactions on fuzzy systems*, 1(2) :98–110, 1993.

[40] Dantong Yu, Gholamhosein Sheikholeslami, and Aidong Zhang. Findout : finding outliers in very large datasets. *Knowledge and information Systems*, 4(4) :387–412, 2002.

[41] Sugato Basu, Mikhail Bilenko, and Raymond J Mooney. A probabilistic framework for semi-supervised clustering. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 59–68, 2004.

[42] MF Augusteijn and BA Folkert. Neural network classification and novelty detection. *International Journal of Remote Sensing*, 23(14) :2891–2902, 2002.

[43] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9) : 1464–1480, 1990.

[44] Khaled Labib and Rao Vemuri. Nsom : A real-time network-based intrusion detection system using self-organizing maps. *Networks and Security*, 21(1), 2002.

[45] Manikantan Ramadas, Shawn Ostermann, and Brett Tjaden. Detecting anomalous network traffic with self-organizing maps. In *International Workshop on Recent Advances in Intrusion Detection*, pages 36–54. Springer, 2003.

[46] Da Deng and Nikola Kasabov. On-line pattern analysis by evolving self-organizing maps. *Neurocomputing*, 51 :87–103, 2003.

[47] Stephen Marsland, Jonathan Shapiro, and Ulrich Nehmzow. A self-organising network that grows when required. *Neural networks*, 15(8-9) :1041–1058, 2002.

[48] Tirthankar Ghosal, Vignesh Edithal, Asif Ekbal, Pushpak Bhattacharyya, George Tsatsaronis, and Srinivasa Satya Sameer Kumar Chivukula. Novelty goes deep. a deep neural solution to document level novelty detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2802–2813, 2018.

[49] Mark Kliger and Shachar Fleishman. Novelty detection with gan. *arXiv preprint arXiv :1802.10560*, 2018.

[50] Haimonti Dutta, Chris Giannella, Kirk Borne, and Hillol Kargupta. Distributed top-k outlier detection from astronomy catalogs using the demac system. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 473–478. SIAM, 2007.

[51] Mei-Ling Shyu, Shu-Ching Chen, Kanoksri Sarinnapakorn, and LiWu Chang. A novel anomaly detection scheme based on principal component classifier. Technical report, MIAMI UNIV CORAL GABLES FL DEPT OF ELECTRICAL AND COMPUTER ENGINEERING, 2003.

[52] Anukool Lakhina, Mark Crovella, and Christophe Diot. Mining anomalies using traffic feature distributions. *ACM SIGCOMM computer communication review*, 35(4) :217–228, 2005.

[53] Heiko Hoffmann. Kernel pca for novelty detection. *Pattern recognition*, 40 (3) :863–874, 2007.

[54] Nojun Kwak. Principal component analysis based on l1-norm maximization. *IEEE transactions on pattern analysis and machine intelligence*, 30(9) :1672–1680, 2008.

[55] Yingchao Xiao, Huangang Wang, Wenli Xu, and Junwu Zhou. L1 norm based kpca for novelty detection. *Pattern Recognition*, 46(1) :389–396, 2013.

[56] Kevin Faust, Quin Xie, Dominick Han, Kartikay Goyle, Zoya Volynskaya, Ugljesa Djuric, and Phedias Diamandis. Visualizing histopathologic deep learning classification and anomaly detection using nonlinear feature space dimensionality reduction. *BMC bioinformatics*, 19(1) :173, 2018.

[57] David MJ Tax and Robert PW Duin. Support vector domain description. *Pattern recognition letters*, 20(11-13) :1191–1199, 1999.

[58] Colin Campbell and Kristin P Bennett. A linear programming approach to novelty detection. In *Advances in neural information processing systems*, pages 395–401, 2001.

[59] Trung Le, Dat Tran, Wanli Ma, and Dharmendra Sharma. An optimal sphere and two large margins approach for novelty detection. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2010.

[60] Yi-Hung Liu, Yan-Chen Liu, and Yen-Jen Chen. Fast support vector data descriptions for novelty detection. *IEEE Transactions on Neural Networks*, 21(8) :1296–1313, 2010.

[61] Bernhard Schölkopf, Robert C Williamson, Alex J Smola, John Shawe-Taylor, and John C Platt. Support vector method for novelty detection. In *Advances in neural information processing systems*, pages 582–588, 2000.

[62] Volker Roth. Outlier detection with one-class kernel fisher discriminants. In *Advances in Neural Information Processing Systems*, pages 1169–1176, 2005.

[63] Volker Roth. Kernel fisher discriminants for outlier detection. *Neural computation*, 18(4) :942–960, 2006.

[64] Andrew B Gardner, Abba M Krieger, George Vachtsevanos, and Brian Litt. One-class novelty detection for seizure analysis from intracranial eeg. *Journal of Machine Learning Research*, 7(Jun) :1025–1044, 2006.

[65] Junshui Ma and Simon Perkins. Time-series novelty detection using one-class support vector machines. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 3, pages 1741–1745. IEEE, 2003.

[66] Paul Hayton, Simukai Utete, Dennis King, Steve King, Paul Anuzis, and Lionel Tarassenko. Static and dynamic novelty detection methods for jet engine health monitoring. *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, 365(1851) :493–514, 2007.

[67] Zengyou He, Shengchun Deng, and Xiaofei Xu. An optimization model for outlier detection in categorical data. In *International Conference on Intelligent Computing*, pages 400–409. Springer, 2005.

[68] Zengyou He, Shengchun Deng, Xiaofei Xu, and Joshua Zhexue Huang. A fast greedy algorithm for outlier mining. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 567–576. Springer, 2006.

[69] Shin Ando. Clustering needles in a haystack : An information theoretic analysis of minority and outlier detection. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 13–22. IEEE, 2007.

[70] Eamonn Keogh, Stefano Lonardi, and Chotirat Ann Ratanamahatana. Towards parameter-free data mining. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 206–215, 2004.

[71] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5 :135–146, 2017.

[72] Jiajia Huang, Min Peng, and Hua Wang. Topic detection from large scale of microblog stream with high utility pattern clustering. In *Proceedings of the 8th Workshop on Ph. D. Workshop in Information and Knowledge Management*, pages 3–10, 2015.

[73] Min Peng, Shuang Ouyang, Jiahui Zhu, Jiajia Huang, Hua Wang, and Jianming Yong. Emerging topic detection from microblog streams based on emerging pattern mining. In *2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD))*, pages 259–264. IEEE, 2018.

[74] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[75] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove : Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[76] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*, 2018.

[77] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv :1802.05365*, 2018.

[78] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635, 2015.

[79] William L Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1489–1501, 2016.

[80] Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. Outta control : Laws of semantic change and inherent biases in word representation models. In

Proceedings of the 2017 conference on empirical methods in natural language processing, pages 1136–1145, 2017.

[81] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6) :391–407, 1990.

[82] Shoaib Jameel, Wai Lam, and Lidong Bing. Nonparametric topic modeling using chinese restaurant franchise with buddy customers. In *European Conference on Information Retrieval*, pages 648–659. Springer, 2015.

[83] Donald Metzler, Congxing Cai, and Eduard Hovy. Structured event retrieval over microblog archives. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 646–655. Association for Computational Linguistics, 2012.

[84] Jey Han Lau, Nigel Collier, and Timothy Baldwin. On-line trend analysis with topic models :\# twitter trends detection topic model online. *Proceedings of COLING 2012*, pages 1519–1534, 2012.

[85] Rui Long, Haofen Wang, Yuqiang Chen, Ou Jin, and Yong Yu. Towards effective event detection, tracking and summarization on microblog data. In *International Conference on Web-Age Information Management*, pages 652–663. Springer, 2011.

[86] Gerard Salton. Automatic text processing : The transformation, analysis, and retrieval of. *Reading : Addison-Wesley*, 169, 1989.

[87] Stephen E Robertson and K Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3) :129–146, 1976.

[88] Zellig S Harris. Distributional structure. *Word*, 10(2-3) :146–162, 1954.

[89] Maria Antoniak and David Mimno. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6 :107–119, 2018.

[90] Loulwah AlSumait, Daniel Barbará, and Carlotta Domeniconi. On-line lda : Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 3–12. IEEE, 2008.

[91] Yu Wang, Eugene Agichtein, and Michele Benzi. Tm-lda : efficient online modeling of latent topic transitions in social media. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 123–131. ACM, 2012.

[92] Hesam Amoualian, Marianne Clausel, Eric Gaussier, and Massih-Reza Amini. Streaming-lda : A copula-based approach to modeling topic dependencies in document streams. In *SIGKDD*, 2016.

[93] Xuerui Wang and Andrew McCallum. Topics over time : a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433, 2006.

[94] Chong Wang, David Blei, and David Heckerman. Continuous time dynamic topic models. *arXiv preprint arXiv :1206.3298*, 2012.

[95] Adriaan MJ Schakel and Benjamin J Wilson. Measuring word significance using distributed representations of words. *arXiv preprint arXiv :1508.02297*, 2015.

[96] Shendy M. El-Shal and Alan S Morris. A fuzzy expert system for fault detection in statistical process control of industrial processes. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 30 (2) :281–289, 2000.

[97] Luc Noyez. Control charts, cusum techniques and funnel plots. a review of methods for monitoring performance in healthcare. *Interactive cardiovascular and thoracic surgery*, 9(3) :494–499, 2009.

[98] John F MacGregor and Theodora Kourti. Statistical process control of multivariate processes. *Control Engineering Practice*, 3(3) :403–414, 1995.

[99] André Eugênio Lazzaretti, David Martinus Johannes Tax, Hugo Vieira Neto, and Vitor Hugo Ferreira. Novelty detection and multi-class classification in power distribution voltage waveforms. *Expert Systems with Applications*, 45 : 322–330, 2016.

[100] Alexander G Tartakovsky, Aleksey S Polunchenko, and Grigory Sokolov. Efficient computer network anomaly detection by changepoint detection methods. *IEEE Journal of Selected Topics in Signal Processing*, 7(1) :4–11, 2012.

[101] Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2) :100–115, 1954.

[102] Leo Breiman. Random forests. *Machine learning*, 45(1) :5–32, 2001.

[103] Manish Kumar and M Thenmozhi. Forecasting stock index movement : A comparison of support vector machines and random forest. In *Indian institute of capital markets 9th capital markets conference paper*, 2006.

[104] Ron S Kenett, Shelemyahu Zacks, and Daniele Amberti. *Modern Industrial Statistics : with applications in R, MINITAB and JMP*. John Wiley & Sons, 2013.

[105] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476) :1566–1581, 2006.

[106] David J Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer, 1985.

[107] Thomas Griffiths, Michael Jordan, Joshua Tenenbaum, and David Blei. Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, 16 :17–24, 2003.