Utilisation de techniques de modélisation thématiques pour la détection de nouveauté dans des flux de données textuelles.

Clément Christophe*,**, Julien Velcin*, Manel Boumghar**

*Laboratoire ERIC, Université Lumière Lyon2,
5 av. P. Mendès-France, 69676 Bron Cedex, France
Julien.Velcin@univ-lyon2.fr
**EDF R&D,
7 Boulevard Gaspard Monge, 91120 Palaiseau, France
manel.boumghar@edf.fr
cle.christophe@gmail.com

Résumé. Avec l'avènement des réseaux sociaux et la multiplication des messages produits au sujet des entreprises, mieux comprendre les retours clients est devenu un enjeu primordial. Des techniques de classification automatique et de modélisation thématique permettent d'ors déjà d'observer les principales tendances observées dans ces données. Il est intéressant, dans une optique d'anticipation, d'observer les thématiques émergentes et de les identifier avant qu'elles ne prennent de l'ampleur. Afin de résoudre cette problématique, nous avons étudié la piste de l'utilisation de modèles LDA pour détecter les documents relatifs à ces thématiques émergentes. Nous avons testé trois systèmes sur plusieurs scénarios d'arrivées de la nouveauté dans le flux de données. Nous montrons que les modèles thématiques permettent de détecter cette nouveauté mais que cela dépend du scénario envisagé.

1 Introduction

De nombreuses entreprises souhaitent être en mesure d'analyser les données qui leur parviennent chaque jour. C'est le cas de l'entreprise EDF avec laquelle ce projet a été effectué. EDF surveille l'évolution des thématiques discutées dans différents types de corpus textuels (réclamations, mails, chatbot, etc.). Un plan de classement prédéfini permet de recourir à des algorithmes de classification supervisée performants afin de placer les différents documents dans des catégories prédéfinies au fur et à mesure de leur arrivée. Cependant, de nombreux documents se retrouvent mal ou même non classés. Cela peut être dû au fait que les catégories évoluent au fil du temps et qu'il est nécessaire de réviser ces plans de classement. Être en mesure de détecter au plus tôt ces tendances nouvelles représente un atout important pour une entreprise. EDF souhaite pouvoir détecter les documents qui ont permis d'amorcer ces évolutions car ils peuvent avoir le potentiel d'anticiper la constitution de nouvelles catégories. Ils constituent une forme d'explication du changement en cours permettant une meilleure interaction avec les utilisateurs du système au sein d'EDF. Afin de surveiller ces évolutions, il est nécessaire de prendre en compte la notion de nouveauté. Dans ce contexte, l'analyse de

nouveauté à partir du flux des documents est une piste envisagée sérieusement pour mieux appréhender ces évolutions.

Dans cet article nous voulons étudier la possibilité d'utiliser des méthodes basées sur des thématiques pour améliorer la détection de nouveauté. Nous avons développé une méthode générique et trois modèles afin de capter les documents nouveaux. Les thématiques construites dans chacun des modèles sont issues de modèles LDA (Blei et al. (2003)) mais il est tout à fait envisageable d'utiliser d'autres modèles comme PLSA (Hofmann (1999)) ou NMF (Lee et Seung (1999)). En plus des modèles développés et au vu de la difficulté d'accessibilité des jeux de données annotés pour la nouveauté, nous avons mis au point une méthodologie permettant d'ajouter artificiellement de la nouveauté dans des données textuelles. Cette méthodologie nous permet de tester plusieurs scénarios d'arrivée (fréquence, volume, etc.) et de mesurer la performance de nos systèmes.

Nous commencerons par définir la notion de nouveauté dans un flux de données textuelles. Nous présenterons ensuite la méthode générique et la déclinaison en trois modèles distincts. Nous montrerons comment nous ajoutons artificiellement de la nouveauté dans notre jeu de données et enfin nous présenterons les différents résultats que nous avons obtenus. En conclusion, nous verrons dans quelle mesure et par quels moyens notre système pourrait être amélioré et dans quels cas il serait particulièrement utile.

2 État de l'art

La détection de nouveauté peut être définie comme le fait de reconnaître des données qui sont différentes d'une certaine manière des données habituellement traitées. Il est courant de rapprocher les idées de nouveauté et de signaux faibles car les méthodes de détection sont souvent employées sur des jeux de données contenant un très grand nombre d'exemples normaux et peu de données considérées comme "anormales". À partir de cette définition, nous pouvons voir le problème de détection de nouveauté comme un problème de classification à deux classes où nous avons une classe de données "anormales" en faible volume qui doit être distinguée des autres possibilités. Au sein de la littérature, les termes *novelty detection* sont souvent rapprochés de termes comme *anomaly detection* et *outlier detection* . Le dictionnaire Merriam-Webster définit le terme *novelty* comme ceci : «qui ne ressemble à rien de ce qui a déjà été observé".»

Dans (Pimentel et al. (2014)), la détection de nouveauté est classée en cinq catégories distinctes : (1) probabiliste, (2) basée sur la distance, (3) basée sur la reconstruction, (4) basée sur le domaine, et (5) basée sur des techniques de théorie de l'information. La première méthode cherche 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. La seconde approche part du principe qu'une nouveauté va apparaître loin de ses plus proches voisins. Pour (3), il est nécessaire d'entraîner un modèle de régression et l'observation d'une erreur importante entre la prédiction et la valeur réelle donne du poids à un score de nouveauté. La quatrième méthode va utiliser des méthodes spécifiques aux domaines pour caractériser les données d'entraînement.

Généralement, elles définissent une frontière autour des données dites “normales”. L’approche (5) va calculer l’apport informationnel des données d’entraînements grâce à l’entropie ou à d’autres techniques basées sur la théorie de l’information de Shannon. Elle se base sur le principe qu’une nouveauté va modifier significativement le contenu informationnel d’un jeu de données.

Les modèles que nous allons présenter se basent sur des mesures de distances (2nde approche) pour calculer un score de nouveauté d’un document arrivant dans notre corpus. Les techniques de modélisation thématiques existantes permettent d’associer des termes et des documents qui ont des relations sémantiques et qui sont souvent utilisés au sein d’un même sujet. Afin de générer automatiquement des thématiques, nous utiliserons un modèle de *Latent Dirichlet Allocation* (Blei et al. (2003)). *LDA* est un modèle probabiliste utilisé pour décrire un corpus de D documents associés à un vocabulaire de taille V. Dans ce modèle, des variables latentes sont utilisées pour représenter des thématiques présentes dans chaque document. LDA utilise un processus génératif qui permet de simuler la création d’un document. A partir des paramètre α et β, le modèle détermine les variables cachées Z_n correspondant aux thématiques. Ces thématiques sont décrites par une distribution de probabilité des termes du vocabulaire sur les thématiques ($\phi_k = p(w_n|z_k)$, où w_n est le n-ième mot du vocabulaire) et une distribution de probabilité des thématiques sur les documents ($\theta^d = p(z_k|d)$). La distribution ϕ_k est, en partie, illustrée dans le tableau 1.

Topic 1	*algorithm (0.042), routing (0.038), dynamic (0.033), packet (0.018)*
Topic 2	*data (0.070), queries (0.033), query (0.028), optimal (0.022)*

TAB. 1 – *Exemple de thématiques LDA.*

Le modèle LDA permet la construction de thématiques à un temps donné. Il sera nécessaire, à terme, de s’intéresser à l’aspect temporel des données afin de construire des modèles pouvant évoluer dans le temps. Plusieurs méthodes ont été présentées ces dernières années et permettent de modéliser l’évolution des thématiques dans le temps, c’est le cas pour (Wang et McCallum (2006), Blei et Lafferty (2006), AlSumait et al. (2008), Wang et al. (2012), et Amoualian et al. (2016)). Ces modèles se basent sur des approches différentes pour modéliser l’évolution des thématiques. Certains utilisent des fenêtres temporelles et se basent sur les paramètres α et β pour mettre à jour le modèle (Blei et Lafferty (2006)) tandis que d’autres (AlSumait et al. (2008) lient les distributions de termes-thématiques pour déterminer les β. Il est intéressant de noter que des approches par prédiction (Wang et al. (2012)) ainsi que l’utilisation d’objets mathématiques complexes comme les copules (Amoualian et al. (2016)) ont fait leur preuve. Pour modéliser la détection au plus tôt, des modèles basés sur des mesures physiques comme la vitesse ou l’accélération sont apparus dans la littérature (Xie et al. (2016), He et Parker (2010). Ces méthodes permettent de détecter très tôt des événements qui apparaissent rapidement mais ne détectent pas les nouveautés qui peuvent apparaître progressivement.

Nous avons observé que ces articles ne proposent pas de méthodes d’évaluation quantitatives pour la détection de la nouveauté. C’est pourquoi nous avons décidé de ne pas nous baser sur ces modèles complexes mais plutôt sur des modèles LDA.

3 Les modèles

3.1 Méthode générique

Dans cette section, nous allons décrire le modèle de façon générique et expliquerons les différents moyens que nous pouvons utiliser afin de l'implémenter.

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. t_c correspond au 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 récent 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 $D_{cont} = \{(d_i, t_i) / t_c < t_i \leq t_c + w_c\}$.

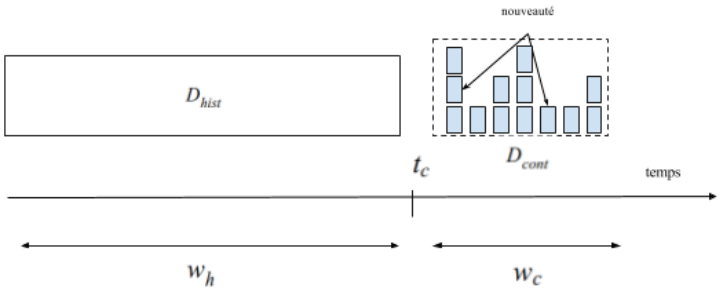

FIG. 1 – *Modèle générique.*

Au sein de cet article, nous utiliserons les notations notées dans le tableau 2

Symbole	Définition
d_i	contenu du document i
t_i	date de parution du document i
t_c	date d'observation
$\mathcal{D}$	ensemble des documents
w_h	taille de la fenêtre d'historique
w_c	taille de la fenêtre de contexte
$H = D_{hist}$	ensemble des documents dans W_h
$C = D_{cont}$	ensemble des documents dans W_c
Z_D	ensemble des thématiques *LDA* calculés sur D

TAB. 2 – *Notations utilisées dans cet article.*

3.2 Calcul de distance.

Comme nous l'avons dit précédemment, nous voulons détecter au plus tôt l'émergence de thématiques nouvelles ainsi que différents phénomènes complexes. Ces thématiques sont construites à partir des documents et c'est donc ces derniers qui induisent la nouveauté. Afin de la mesurer au niveau des thématiques, il faut aussi être capable de la mesurer au niveau des documents. Dans notre modèle, nous voulons leur associer 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 nous basons sur le fait que la nouveauté apparaît anormalement loin, en termes de distance, de ses plus proches voisins. Notre fonction de scoring $score$ doit agréger des calculs de dissimilarité des documents par rapport à l'historique et au contexte. Cette dissimilarité peut être calculée par rapport à plusieurs ensembles. Une première idée consiste à calculer la dissimilarité entre les documents du contexte et de l'historique : $score(d_i, H) = aggreg_{\substack{d' \in H \\ d_i \in C}}(diss(d, d'))$ pour tout d_i dans C.

Plusieurs fonctions sont connues dans la littérature pour calculer la dissimilarité entre documents. Si nous considérons $\overrightarrow{x}$ un vecteur de dimension n où n correspond au nombre de mots dans le vocabulaire. $\overrightarrow{x}$ peut représenter un document sous la forme de sac de mots ou bien les mots les plus probables d'une thématique LDA. Nous pouvons utiliser la divergence Cosine, la divergence de Kullback-Leibler symétrique ou encore la divergence de Jensen-Shannon :

— $cosineDiv(x_1, x_2) = 1 - \frac{\sum x_{1i}.x_{2i}}{\sqrt{\sum {x_{1i}}^2}.\sqrt{\sum {x_{2i}}^2}}$

— $KLDiv(x_1, x_2) = \frac{1}{2}\sum x_{1i}.log(\frac{x_{1i}}{x_{2i}}) + \frac{1}{2}\sum x_{2i}.log(\frac{x_{2i}}{x_{1i}})$

— $JSDiv(x_1, x_2) = \frac{1}{2}\sum x_{1i}.log(\frac{x_{1i}}{\frac{1}{2}(x_{1i}+x_{2i})}) + \frac{1}{2}\sum x_{2i}.log(\frac{x_{2i}}{\frac{1}{2}(x_{1i}+x_{2i})})$

3.3 Modèle de comparaison documents-documents

La Figure 2 représente une comparaison des termes présents dans les documents de l'historique et du 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. Pour la comparaison entre les documents, nous pouvons utiliser les mesures citées dans la partie 3.2. Nous avons choisi d'utiliser une divergence *cosine* car elle est connue pour avoir une performance élevée (Strehl et al. (2000)). Les scores associés aux termes correspondent au TF-IDF de chaque terme estimé sur le corpus. Une fois les distances calculées, nous obtenons une matrice de distance avec, en ligne, les documents du contexte (là où nous voulons trouver la nouveauté) et, en colonne, les documents de l'historique. Nous avons vu dans l'état de l'art qu'un document nouveau est anormalement loin de ses voisins. Nous ne voulons donc pas comparer un document du contexte avec tous les documents de l'historique mais seulement avec les documents qui y ressemblent fortement donc avec ses plus proches voisins. Le but est d'identifier les documents isolés dont même les plus proches voisins sont anormalement loin. Afin d'agréger les résultats, nous faisons donc une moyenne sur les distances par rapport aux k plus proches voisins. Plus ce score est élevé, plus le document est susceptible d'être nouveau.

— 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 $diss(d, d'_i)$

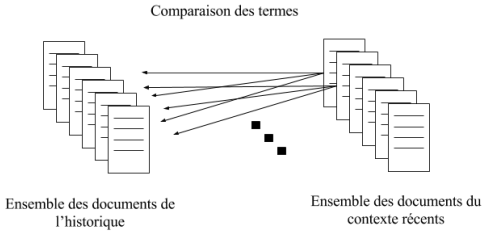

FIG. 2 – *Comparaison des documents deux à deux.*

3.4 Modèle de comparaison thématiques-documents

La Figure 3 introduit la notion de thématiques dans le modèle. Ce dernier permet de comparer les termes des documents du contexte récent avec les termes les plus probables des thématiques de l'historique. Comme nous l'avons dit en présentant le modèle LDA, une thématique est décrite grâce à sa distribution de probabilité sur les termes et sur les documents. Nous prenons les 100 termes les plus probables par thématique et nous pouvons donc comparer les documents avec les thématiques dans un même espace. Pour ce modèle nous choisissons d'utiliser la distance *cosine* en prenant, pour score, la probabilité des termes dans la thématique avec laquelle on compare. Bien que la comparaison entre un score TF-IDF et une probabilité ne soit pas mathématiquement rigoureuse, nous considérons les probabilités comme un score d'appartenance à une thématique, ce qui nous permet d'utiliser les deux mesures dans la même expression. Ce modèle détermine rapidement si un document fait partie des données «normales»ou s'il se trouve loin des thématiques de l'historique.

3.5 Modèle de comparaison thématiques-thématiques

La Figure 4 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. Nous agrégeons les résultats de la même manière que dans la partie 3.3. Afin de sélectionner seulement les thématiques vraiment nouvelles, nous fixons un seuil qui correspond à : $threshold = \mu(score) + \sigma(score)$. Ce seuil traduit la notion

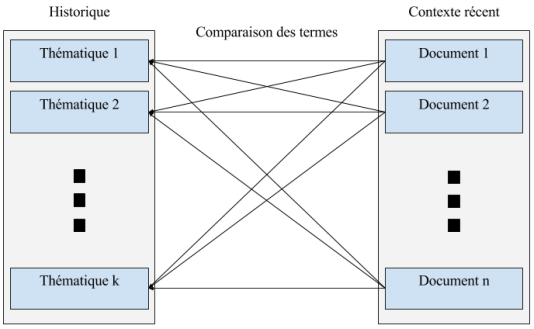

FIG. 3 – *Comparaison des documents avec les thématiques de l'historique*

d'anormalement distant des thématiques précédentes. Une fois les thématiques nouvelles identifiées, nous allons utiliser les documents les plus probables de celles-ci et comparer 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. Ensuite, nous allons comparer les termes des documents les plus probables au sein de ces thématiques nouvelles avec les termes des thématiques de l'historique (étape (b)). Cela permet de déterminer quels sont les documents responsables de la nouveauté de la thématique.

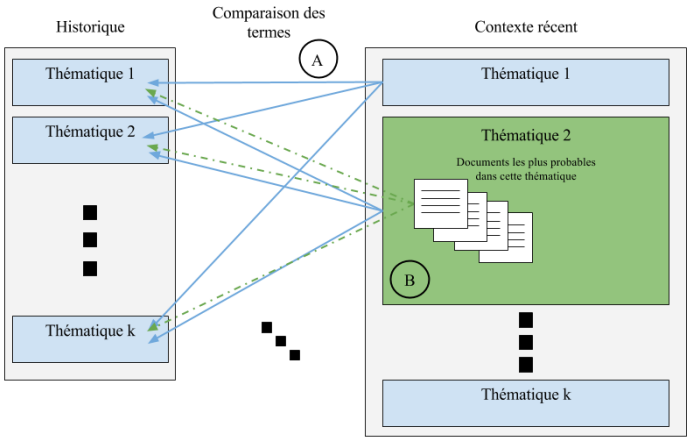

FIG. 4 – *Comparaison des documents des thématiques nouvelles avec les thématiques de l'historique*

Détection de nouveauté avec LDA.

4 Expérimentation

4.1 Méthodologie

Il n'existe pas de jeux de données facilement accessible où la nouveauté est annotée par document. Autrement dit, nous n'avons pas de vérité terrain qui permettrait d'observer si les systèmes classent bien les documents annotés comme nouveaux. On a donc dû constituer notre propre vérité terrain. Pour cela, nous avons utilisé un jeu de données où chaque texte est déjà associé à une catégorie. En utilisant ce type de jeu de données, nous pouvons simuler artificiellement de la nouveauté. Pour cela, nous avons constitué nos documents de l'historique en enlevant la totalité d'une catégorie. Pour la suite nous introduisons, dans nos documents de contexte, une partie des articles de la catégorie enlevée précédemment. Cela nous permet de contrôler la quantité de nouveauté que nous introduisons et nous pouvons donc étudier l'impact que cela a sur nos modèles.

4.2 Jeu de données

Les données de l'entreprise EDF sont des données sensibles, donc nous avons choisi un jeu de données public (Tang et al. (2012)) [1] pour tester nos modèles. Ces données correspondant à des résumés d'articles scientifiques publiés entre 1990 et 2005 et associés à cinq catégories : theory, database, datamining, visu et medical.

Nous avons appliqué notre méthodologie à ce jeu de données et nous avons constitué une fenêtre d'historique contenant les articles publiés entre 1990 et 1999 en enlevant chacune des catégories. Nous avons donc une fenêtre de contexte contenant les articles publiés entre 2000 et 2005 dans laquelle nous avons introduit un certain nombre de documents correspondant à la catégorie enlevée précédemment. Le but de ces expériences était de voir comment les techniques de modélisation thématiques pouvaient être utiles à la détection de nouveauté : nous avons étudié l'impact des différentes catégories et de la quantité de nouveauté introduite. Nous avons donc 5 jeux de données d'historique : correspondant aux 5 catégories que nous avons enlevées, et, pour chaque catégorie, 4 jeux de données de contexte. En effet, nous avons ajouté 1, 5, 20 et 100 documents nouveaux pour vérifier l'impact de la quantité. Pour information, les ensembles d'historiques et de contextes sont composés d'environ 4000 documents chacun. Nous avons répété l'opération 100 fois pour avoir des documents différents à détecter et donc des résultats plus stables. Les résultats de nos mesures sont commentés dans la partie suivante.

4.3 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. Les articles cités dans l'état de l'art utilisent une mesure de perplexité

1. `https://aminer.org/collaboration`

pour mesurer la qualité de leur thématique mais ne propose pas de mesure quantitative pour la détection de la nouveauté. 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 des 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 encore ces changements et cela peut expliquer les scores de détection pas à la hauteur de nos attentes.

Au niveau des mesures d'AUC, il est intéressant de constater qu'à partir de seulement 5 nouveaux documents introduits dans l'ensemble de contexte, on peut observer des différences dans les tendances de détections entre les modèles. En effet, dans la Table 3, quand on ajoute 5 nouveaux documents, on voit qu'il est plus difficile de détecter des catégories comme *datamining* et *medical*. Au contraire, les catégories *theory* et *visu* ont l'air d'être assez faciles à détecter ($AUC > 0.7$). Lorsqu'on utilise des modèles thématiques, que ce soit pour résumer l'historique (Table 4) ou pour la fenêtre de contexte courant (Table 5), nous remarquons que la tendance s'inverse 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 commun 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 (Table 5), la première étape identifie les nouvelles thématiques assez bien car on retrouve bien les nouveaux documents insérés dans leur liste des 100 documents les plus probables.

Nb of docs	1	5	20	100
database	0.81	0.68	0.67	0.66
datamining	0.47	0.59	0.54	0.57
medical	0.71	0.61	0.65	0.66
theory	0.73	0.82	0.80	0.80
visu	0.67	0.75	0.72	0.71

TAB. 3 – *AUC moyennes du modèle de comparaison documents-documents*

Nb of docs	1	5	20	100
database	0.73	0.68	0.66	0.65
datamining	0.33	0.47	0.43	0.44
medical	0.74	0.73	0.74	0.74
theory	0.75	0.75	0.76	0.75
visu	0.66	0.61	0.60	0.60

TAB. 4 – *AUC moyennes du modèle de comparaison thématiques-documents*

Nb of docs	1	5	20	100
database	0.43	0.64	0.73	0.60
datamining	0.25	0.68	0.53	0.55
medical	0.62	0.71	0.69	0.69
theory	0.33	0.85	0.81	0.83
visu	0.63	0.57	0.53	0.65

TAB. 5 – *AUC moyennes du modèle de comparaison thématiques-thématiques*

Les résultats présentés dans la Table 6 montrent le nombre moyen de documents nouveaux que l'on retrouve dans les 100 premiers documents classés par score de nouveauté (2.5% du classement). C'est, en quelque sorte, un zoom sur le début de la courbe ROC. Ces 100 premiers documents sont destinés à être présentés à des experts métiers pour interprétation. La première colonne montre le niveau de détection lorsque l'on compare deux à deux les documents du contexte et de l'historique par rapport aux termes qu'ils utilisent. On peut voir 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'effets sur la détection. La deuxième colonne présente les mêmes résultats une fois que l'on a résumé nos documents de l'historique sous forme de thématique LDA. Pour cela nous avons utilisé le modèle présenté en partie 3.4. On peut voir 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 l'on résume notre fenêtre de contexte sous forme de thématiques (partie 3.5), on observe la même tendance que précédemment, à part pour la catégorie *visu* qui semble rester assez 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 par thématiques nouvelles identifiées). Aussi, les modèles thématiques permettent de manipuler des matrices plus légères (en comparant 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

TAB. 6 – *Comparaison des mesures de précision à 100 des différents modèles*

5 Conclusion et perspectives

Dans cet article, nous avons commencé par apporter des précisions sur la définition de la nouveauté. Nous avons listé des articles de l'état de l'art sur des techniques qui nous ont inspirées ou nous inspireront pour la suite. Nous avons ensuite évalué la capacité des modèles thématiques à détecter de la nouveauté dans des flux de données textuelles en développant une méthodologie qui, à partir de données classées, nous permet d'introduire artificiellement de la nouveauté. Cette méthodologie pourrait être améliorée en utilisant d'autres jeux de données où les catégories n'évolueraient pas dans le temps : cela permettrait d'éviter la détection de nouveauté intra-classes. 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. Nous n'avons pas étudié l'influence des différents paramètres utilisés. Il serait intéressant d'observer à quel point le nombre de plus proches voisins utilisés ou encore le nombre de thématiques influent sur la précision du modèle. Pour la suite de ce travail, il est question de se baser sur des méthodes thématiques temporels présentés dans AlSumait et al. (2008), Amoualian et al. (2016), Blei et Lafferty (2006), et Wang et al. (2012) afin de pouvoir détecter la nouveauté dans des thématiques qui évoluent dans le temps en étudiant, par exemple, l'évolution des paramètres α et β. Aussi, afin d'aborder la question de la détection au plus tôt, nous aimerions nous baser sur des systèmes présentés dans Xie et al. (2016) et He et Parker (2010) et ainsi adapter des mesures d'accélération aux thématiques.

Références

AlSumait, L., D. Barbará, et C. Domeniconi (2008). 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*, pp. 3–12. IEEE.

Amoualian, H., M. Clausel, E. Gaussier, et M.-R. Amini (2016). Streaming-lda : A copula-based approach to modeling topic dependencies in document streams. In *SIGKDD*.

Blei, D. M. et J. D. Lafferty (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120. ACM.

Blei, D. M., A. Y. Ng, et M. I. Jordan (2003). Latent dirichlet allocation. *Journal of machine Learning research 3*(Jan), 993–1022.

He, D. et D. S. Parker (2010). Topic dynamics : an alternative model of bursts in streams of topics. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 443–452. ACM.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pp. 289–296. Morgan Kaufmann Publishers Inc.

Lee, D. D. et H. S. Seung (1999). Learning the parts of objects by non-negative matrix factorization. *Nature 401*(6755), 788.

Pimentel, M. A., D. A. Clifton, L. Clifton, et L. Tarassenko (2014). A review of novelty detection. *Signal Processing 99*, 215–249.

Strehl, A., J. Ghosh, et R. Mooney (2000). Impact of similarity measures on web-page clustering. In *Workshop on artificial intelligence for web search (AAAI 2000)*, Volume 58, pp. 64.

Tang, J., S. Wu, J. Sun, et H. Su (2012). Cross-domain collaboration recommendation. In *KDD'2012*.

Wang, X. et A. McCallum (2006). 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*, pp. 424–433. ACM.

Wang, Y., E. Agichtein, et M. Benzi (2012). 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*, pp. 123–131. ACM.

Xie, W., F. Zhu, J. Jiang, E.-P. Lim, et K. Wang (2016). Topicsketch : Real-time bursty topic detection from twitter. *IEEE Transactions on Knowledge and Data Engineering 28*(8), 2216–2229.

Summary

With the advent of social networks and the multiplication of product messages about companies, better understanding of customer feedback has become a key issue. Clustering techniques and thematic modeling already allow to observe the main trends observed in this data. It is interesting, from an anticipatory perspective, to observe the emerging themes and to identify them before they grow in size. To solve this problem, we studied the use of LDA models to detect documents related to these emerging themes. We tested three systems on several novelty arrival scenarios in the data stream. We show that the thematic models allow to detect this novelty but that it depends on the scenario considered.