Projection de documents par topic models :
Un cas d'usage sur la détection d'experts dans le paysage médiatique

Julien Denes
julien.denes@sciencespo.fr

Rapport de stage réalisé au médialab de Sciences Po sous la supervision de Vincent Lépinay
Master 2 Informatique, spécialité ANDROIDE – Sorbonne Université
Février – juillet 2019

Résumé. Les techniques de projections de documents, ou document embeddings, se font aujourd'hui toujours plus nombreuses, toujours plus complexes, et toujours moins interprétables. Pourtant au sein de ces modélisations, une famille reste peu utilisée et peu documentée dans la littérature : les projections par topic models, ou projection par les sujets (topic embeddings). Si les topic models ne répondent initialement pas à un but de projection de documents, leur usage semble naturel puisqu'ils capturent efficacement l'information d'un document sous forme d'un vecteur de proportion d'appartenance à chaque sujet, pour n'importe quel nombre de sujets donné. Nous proposons une étude de ces modèles avec deux objectifs : documenter leur utilisation comme projection de documents et les performances de ces projections, et montrer leur avantage pour l'interprétation des résultats. Nous illustrons ces deux points grâce à une étude sur la détection des experts académiques dans le débat radiophonique français à partir de leur discours. Nous atteignons un taux de justesse d'environ 80%, une performance comparable à celle des principales autres approches de projection. Nous montrons que le principal avantage de ce type de modélisation est d'être facilement interprétable, et semble donc un outil adapté à divers usages en sciences sociales, en particulier la catégorisation automatique de documents.

1 Introduction

1.1 Projection de documents

L'une des préoccupations centrales de toute approche quantitative du langage naturel est l'adoption d'une représentation adaptée du texte sous une forme numérique. En effet, de nombreux algorithmes, au premier rang desquels les algorithmes d'apprentissage automatique, requièrent pour chaque observation une représentation sous forme de vecteur d'attributs numériques de taille fixe. Pou représenter un document, l'approche la plus ancienne et encore largement la plus utilisée est le sac de mot (ou bag of words, Harris, 1954), mais il s'agit d'une représentation très volumineuse reposant des vecteurs fortement creux. De nombreuses approches se sont développées plus récemment, avec l'idée de créer des représentations numériques des mots plus compactes et capables de capturer des propriétés sémantiques variées. Regroupées sous le nom de projections de mots, ou word embeddings, ces approches comme word2vec (Mikolov et al., 2013) ou GloVe (Pennington, Socher et Manning, 2014) se montrent performantes pour capturer dans des dimensions plus faibles le sens des mots, avec des vecteurs aux propriétés intéressantes, comme la proximité de mots semblables et même des propriétés calculatoires. Par exemple le résultat de vec("Madrid") − vec("Spain") + vec("France") est très proche de vec("Paris"). Ces approches ont été suivies par la création de méthodes de projections de documents entiers, ou document embedding, à l'utilité plus large. On pourra par exemple citer doc2vec (Le et Mikolov, 2014), très inspiré de word2vec, mais aussi des approches plus nouvelles comme Skip-Thought (Kiros et al., 2015), InferSent (Conneau, Kiela et al., 2017), Recursive Autoencoders (Socher et al., 2011) ou encore Gated ConvNet (Dauphin et al., 2017). Si leur objectif est de densifier les représentations ou de collecter une information plus utile, de nombreux problèmes accompagnent ces modèles, par exemple en terme de perte de l'interprétabilité, ou encore de performances qui

laissent à désirer par rapport au temps de calcul requis (Wieting et Kiela, 2019). La complexité de leur architecture, elle aussi croissante, rend de moins en moins possible leur utilisation par des non-spécialistes du domaine.

Il existe, d'un autre côté, une autre école du traitement du langage, plus orientée vers les méthodes appliquée, avec en apparence un objectif différent puisqu'elle cherche à révéler les sujet évoqués dans un corpus de texte à partir de texte non annoté. Ces méthodes, rassemblées sous le nom de topic models, sont à peu près aussi anciennes que les techniques de projections de texte. On pourra citer par exemple l'allocation de Dirichlet latente (LDA, Blei, Ng et Jordan, 2003), modèle le plus largement répandu, et les nombreux raffinements qui lui ont suivi, comme le Correlated Topic Model (CTM, Blei et Lafferty, 2007) pour intégrer les liens entre documents, le Structural Topic Model (STM, Roberts, Stewart, Tingley et al., 2014) pour intégrer des covariables, le Biterm Topic Model (BTM, Yan et al., 2013) pour les textes courts, etc. L'objectif pour lequel ont été créés ces modèles est de découvrir les « sujets » qui sont évoqués dans les textes d'un corpus à partir de leur vocabulaire. Nous étudierons plus en détail leur fonctionnement dans la suite, mais le principe est le suivant. Le principal paramètre d'entrée de ces modèles est le nombre de sujets recherché, appelé K. Une fois entraîné, le modèle renvoient deux objets : les mots associés à chacun des sujets sous forme de distribution des mots entre chaque sujet, et surtout la distribution des sujets sur chaque document, sous forme d'un vecteur de taille K dont les composants somment à 1, et qui indique l'importance de chaque sujet dans le document. Il y a donc un lien naturel avec les projections de documents évoquées plus tôt, puisque comme eux les topic models proposent une représentation des documents sous forme d'un vecteur de taille fixe. On appellera ces projections topic embeddings, ou projections par les sujets. Étonnamment, à notre connaissance, très peu voire aucun usage n'est fait de ces vecteurs « tels quel » comme projection des documents dans la littérature. Il existe pourtant de très nombreux topic models, chacun créé avec une volonté d'amélioration des modèles précédents ou de s'adapter à des situations particulières. L'objectif du présent travail est ainsi de réconcilier la richesse des topic models avec l'importance des projections de documents, et de montrer que ces modèles peuvent être utilisés comme des projections performantes et interprétables. Pour cela, nous centrerons notre analyse sur un cas d'usage sur la détection des experts académiques dans le débat radiophonique français.

1.2 Introduction du cas d'usage

L'importance du rôle des médias dans la perception qu'un individu aura d'une idée, d'une personne ou d'un endroit n'est pas un sujet d'étude très récent. Dès 1990, Appadurai (1990) introduit la notion de mediascape, pour décrire deux objets aussi intrinsèquement liés que différents. Outre l'ensemble des moyens qui servent à produire et à disséminer l'information, il s'agit aussi des *« images du monde produites par ces médias »*, qui tendent à être centrées autour d'images et de narratifs, et qui participent selon lui à créer des *« proto-narratifs »* de l'autre, d'autres vies possibles et fantastiques. Son idée est en fait que les médias modernes conduisent des images biaisées du monde, souvent préconçues et simplifiées, qui mélangent des représentations réelles et fantasmées de sujets, de personnes et d'endroits. La Russie, parmi tous, est sans doute l'un des pays les plus sujets à de telles représentations fantasmées. Cette image de la Russie proposée par les médias est très susceptible d'avoir un impact direct sur la perception des individus sur le sujet, puisque les médias constituent, même à l'heure de l'explosion de l'information offerte par internet, la principale source d'information. Cette analyse de l'image médiatique de la Russie sera donc au centre de ce travail. Notre intérêt se portera sur une période historique large allant de 1980 à 2016, pour capturer une plus grande quantité de discours sur la Russie, et éventuellement étudier son évolution.

Plus précisément, notre intérêt se portera sur ceux qui parlent de la Russie dans les médias. En effet, le discours médiatique n'est pas entièrement neutre, et il l'est encore moins lorsque ce ne sont pas des journalistes qui s'expriment mais des participants extérieurs, invités par exemple à partager leur analyse, leur opinion, ou à débattre. Notre intérêt se portera plutôt vers ces intervenants, dont nous chercherons à analyser le discours. En particulier, nous nous focaliserons sur la question de l'expertise. Parmi ces intervenants en effet, certains justifient d'une expérience sur le sujet et produirons des discours plus éclairés et informés, là

où d'autres au contraire participeront à propager une image naïve. Said (1997) décrit par exemple comment la perception de l'Islam et des musulmans dans la société américaine a été orientée par la vision qui en était donnée par les médias, en particulier pour construire l'idée d'une « menace civilisationnelle ». Albæk, Christiansen et Togeby (2003) s'intéressent plus avant au rôle d'expertise joué par les académiques dans la presse danoise, avec une multiplication par 7 de leur mention dans les journaux entre 1961 et 2000. De nombreux autres travaux s'attachent à comprendre cette croissance, l'expliquant notamment par la nécessité pour les journalistes d'objectiver les faits, mais aussi de légitimer leur propre idéologie ou certaines décisions politiques (Martin, 1991 ; Peters, 1995 ; Steele, 1995). Ainsi, l'importance des experts est croissante dans les médias, et participe grandement à l'image que ces derniers souhaitent véhiculer sur un sujet.

Notre étude se focalise dans les faits sur la radio française. Le choix de se restreindre à la radio est guidée par deux aspects : d'abord un facteur contraignant, puisque les bases de données dont nous disposons sont bien plus riches et facilement exploitables pour la radio que pour la télévision ou les journaux. La Section 4.1 donne plus de détails sur la base de données et ses informations. Mais il s'appuie aussi sur un choix délibéré, puisque la radio est le média qui tend à faire le plus appel à des intervenants extérieurs, et donc à des experts externes. Elle nous permet ainsi un ensemble plus diversifié d'interventions en terme de profil d'expertise, et donc une plus grande richesse d'analyse.

1.3 Questions de recherche

D'un point de vue technique, notre recherche s'intéressera à l'applicabilité des topic models comme projections de documents efficaces. Sous le terme efficace, nous entendons en fait deux aspects : **les projections par les sujets sont-elles aussi performantes que d'autres projections plus répandues ?** Mais aussi, **quelles sont les capacités des projections par les sujets à répondre aux faiblesses de ces autres approches, en terme d'interprétabilité et de complexité algorithmique ?** Dans le cadre du cas d'étude, qui viendra illustrer l'utilisation des méthodes que nous aurons décrites préalablement et nous permettra de répondre à ces questions, nous nous focaliserons sur la question de l'expertise et de son identification. Nous l'aborderons plus précisément sous la perspective du langage, avec une question simple mais qui pourtant ne semble pas avoir été traitée par le passée : est-il possible d'identifier l'expertise grâce au langage ?

La suite de ce travail s'organise ainsi : dans la Section 2, nous proposons une revue des travaux connexes sur le lien entre topic models et projections de documents ainsi que sur la recherche d'experts. Dans la Section 3, nous présentons les principaux topic models. La Section 4 présente les données du cas d'étude, et la Section 5 les résultats de l'application des projections par les sujets. La Section 6 démontre la capacité d'interprétation de ces projections. Enfin, la Section 7 propose une discussion des résultats.

2 Travaux connexes

2.1 Projections et topic models

Si de nombreux travaux existent pour développer de nouveaux modèles de topic models, en attestent les nombreuses revues des approches existantes (Blei, Carin et Dunson, 2010 ; Blei, 2012), peu de travaux s'intéressent à la possibilité d'appliquer ces techniques à d'autres usages que ceux pour lesquels elles ont étés initialement conçues, comme la projection de document. Les bénéfices entre les deux approches sont pourtant mutuels. Un aspect de cette collaboration cherche par exemple à améliorer les topic models grâce aux projections de mots. On pourra citer par exemple Jiang et al., 2016, qui, cherchant à éviter l'hypothèse sac de mot (qui postule que l'ordre des mots n'a pas d'importance) utilisée dans les topic models, intègre les topic models et les projections de mots dans un unique système nommé Latent Topic Embedding (LTE) qui les génère conjointement. De manière similaire, Moody, 2016 propose le modèle lda2vec, qui possède la particularité de s'appuyer sur un LDA pour apprendre simultanément une projection sur les mots et une sur les documents appartenant au même espace vectoriel. Très proche également, l'article de Das, Zaheer et Dyer (2015) cherche à proposer une généralisation des topic models en leur permettant de découvrir des

sujets non pas sur l'espace des mots mais sur l'espace de leur projection, pour la création d'un topic model appelé Gaussian LDA.

Les travaux les plus fréquentes sont cependant ceux qui cherchent à incorporer de l'information fournie par les topic models dans la conception de projections. On pourra par exemple citer pour la projection de mots Liu et al., 2015 qui développe une nouvelles méthode de projections combinant des informations contextuelles à la manière de word2vec avec le vecteur de distribution des mots entre chaque sujets fourni par un LDA, nommée Topical Word Embeddings (TWE). Des travaux existent également pour la projection de documents. Li et al., 2016 vise spécifiquement à créer un nouveau type de projection qui incorpore l'information apportée par les sujets au sein du modèle génératif des projections des mots, nommée TopicVec, et dont les performances sont similaires à celles des projections les plus avancées. Il existe un réel l'intérêt autour de l'information apportée par les sujets, comme en témoigne Yu et al., 2017 qui propose un modèle de projection basé sur l'utilisation d'un réseau LSTM-RNN cherchant à extraire des sujets dans les documents sans utiliser de topic model. Le point commun de tous ces modèles est qu'ils cherchent d'utiliser de l'information fournie par les topic models dans une nouvelle projection, plutôt que d'utiliser le vecteur d'allocation des topic models comme projection « tel quel ».

Les travaux centrés sur l'utilisation de topic models comme projections en eux-mêmes sont donc assez rares, mais pas inexistants. Un travail central sur le sujet est celui de He, Hu et al. (2017). Les auteurs mettent en avant le fait que de nombreux de topic models où un coût de calcul très élevé, et résistent mal à la scalabilité dans de grandes dimensions. Ils s'intéressent en particulier au Correlated Topic Model (CTM) et proposent un modèle conçu spécifiquement pour être extraire un très grand nombre de sujets (jusqu'à 10 000 dans l'article). Le but assumé est de proposer une projection des documents efficaces pour des tâches de classifications, à l'image des tests sur différents jeux de données qu'ils réalisent. Ils sont d'ailleurs l'un des rares travaux à faire explicitement mention de topic embeddings.

2.2 Recherche et quantification de l'expertise

Plusieurs travaux se sont attachés à définir des mesures et des modèles d'expertise. On pourra par exemple citer le travail de Fang et al. (2013), qui, à partir d'une matrice représentant les compétences auto-reportées d'un ensemble d'employés, cherche à modéliser objectivement leurs aires d'expertises, y compris celles dont ils ne sont pas conscients. L'aspect prédictif des modèles d'expertises, dans lequel nous nous inscrivons, est en effet assez répandu dans ces travaux, à l'image de Cameron et al., 2010 qui cherche à identifier, dans un réseau d'académiques en informatique, l'expertise sectorielle que certains auteurs posséderaient mais que les organisateurs de conférences n'auraient pas identifié, pour les aider à trouver de nouveaux intervenants qualifiés. Cet article nous permet d'ailleurs de remarquer que la plupart des approches d'identification des experts utilisent des modélisations par les graphes, qui sont par exemple très utiles lorsqu'ils reflètent un échange d'information. Ils sont cependant assez peu applicable au monde des médias comme dans notre cas, et une approche par le traitement du langage naturel semble s'affirmer comme une bonne alternative.

Une question très évoquée dans la littérature, et qui nous intéresse évidement grandement, est celle de définir de manière objective et cohérente ce qu'est vraiment un « expert », et principalement en termes métriques. Dans notre cas, ce sont avant tout des métriques quantitatives qui nous intéressent. Pour le champs académique, Cameron et al., 2010 développe une métrique basée sur l'homogénéité des sujets évoqués dans les publications en terme de taxonomie des champs de l'informatique, et sur l'homogénéité des sujets de ses co-auteurs. Une concentration sur un petit nombre de sujets précis s'avère ainsi être un bon marqueur d'expertise, bien meilleur que les naïves métriques de citation. C'est également l'idée de Hofmann et al., 2008, qui fixe l'objectif de trouver de nouveaux experts étant donnée une base d'experts avérés. Leurs résultats montrent que c'est avant tout le sujet évoqué qui permet une bonne prédiction, avant même les facteurs contextuels comme la place dans la structure sociale ou l'influence sociale. Herling, 2000 propose une revue des différentes perspectives de l'expertise, mais illustre que la définition de l'expertise se base bien souvent sur une notion de « connaissance » peu applicable dans notre cas.

3 Méthodes de projection de documents par les sujets

3.1 Principe général des topic models

Le principe général des topic models est assez standard, et décrit par exemple dans Blei, Carin et Dunson, 2010. On définit formellement un sujet comme une distribution sur un vocabulaire fixé V. Par exemple, un sujet sur le sport est une distribution pour laquelle les mots “hockey”, “équipe” ou “match” sont associés à des probabilités fortes, au contraire de “biologie” ou d’“immigration”. Comme leur nom l’indique, les topic models reposent sur une hypothèse d’existence d’un modèle génératif pour les sujets, les documents et les mots, avant même de prendre en compte les données. Ce processus procède en deux phases. Dans une première phase, on suppose que les sujets préexistent à la génération des documents et on les génère aléatoirement. Dans un second temps, les documents sont générés selon le processus suivant :

1. Pour chaque document d_i, choisir aléatoirement une distribution sur les sujets (somme donc à 1)
2. Pour chaque mot $w_{i,j}$ du document d_i,
   (a) Choisir aléatoirement un sujet selon la distribution des sujets de l’étape 1,
   (b) Choisir aléatoirement un mot selon la distribution de ce sujet.

Ce modèle reflète ainsi l’intuition que chaque document appartient à plusieurs voire à tous les sujets, dans des proportions différentes (1). Chaque mot du document « affirme » l’appartenance du document à l’un des sujets, et chaque document est ainsi affirmé en proportion de son appartenance à ce sujet (2). En pratique, l’existence de ce modèle stochastique génératif permet d’écrire la vraisemblance de tout jeu d’observations, en fonction des paramètres du modèle, et ainsi d’approximer les meilleurs paramètres pour reconstruire une modélisation. Les principales variations entre topic models portera donc sur la nature des deux lois de probabilités sous-jacentes (la distribution des sujets sur les mots et la distribution des sujets) et sur l’intégration de variables supplémentaires.

Dans la suite, la convention de notation suivante est systématiquement utilisée. D désigne le corpus constitué d’un ensemble de n documents, notés d et indexés par i : $D = (d_1, d_2, ..., d_i, ..., d_n)$. Chaque document d_i est de taille m_i et est composé d’un ensemble de mots, appelé w_i et dont les mots sont indexés par j, soit $w_i = (w_{i,1}, ..., w_{i,j}, ..., w_{i,m_i})$. On note V le vocabulaire, K le nombre de sujets, et β un sujet (i.e. une distribution sur V). Finalement, nous utiliserons fréquemment la notation $X_{1:n}$ pour signifier $X_1, X_2, ..., X_n$.

3.2 Allocation de Dirichlet latente (LDA)

Cette modélisation, introduite dans Blei, Ng et Jordan, 2003, pose la base des topic models modernes. Elle s’appuie sur l’hypothèse centrale suivante : tout sujet β est une distribution de Dirichlet sur le vocabulaire, paramétré par un vecteur de paramètres multinomial η (i.e. un vecteur dont les composants somment à 1), soit : $\beta_k \sim \text{Dir}_V(\eta)$ pour tout $k \in K$. Le processus génératif des documents est le suivant :

1. Pour chaque document d_i, on tire la proportion de chaque sujet une loi symétrique de Dirichlet : $\theta_i \sim \text{Dir}_K(\alpha)$.
2. Pour chaque mot $w_{i,j}$ du document d_i,
   (a) On tire aléatoirement un sujet selon une loi multinomiale sur la proportion des sujets dans le document : $z_{i,j} \mid \theta_i \sim \text{Mult}(\theta_i)$.
   (b) On tire aléatoirement un mot selon une loi multinomiale sur la distribution de ce sujet : $y_{i,j} \mid z_{i,j}, \beta_{1:K} \sim \text{Mult}(\beta_{z_{i,j}})$.

Nous proposons également le modèle graphique probabiliste associé à ce processus dans la Figure 1, reproduite de Blei, Carin et Dunson, 2010. Les variables en gris foncé correspondent aux variables observées (i.e. l’occurrence des mots $y_{i,j}$) et les variables en gris clair les paramètres ou valeurs fixées par l’observation

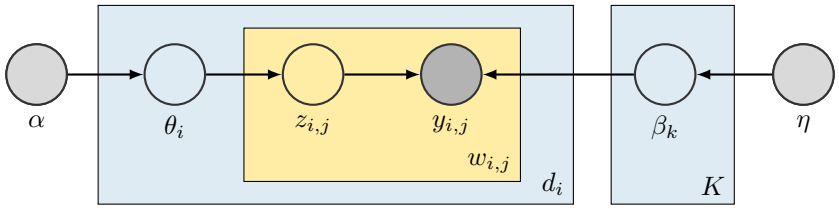

FIGURE 1 – Modèle graphique du LDA

des données. Les autres sont des variables cachées. Il résulte de ce processus que la distribution jointe des variables observées et cachées correspondant au modèle est la suivante :

$$p(\beta_{1:K}, \theta_{1:n}, z_{1:n}, y_{1:n}) = \prod_{k=1}^{K} p(\beta_k) \prod_{i=1}^{n} p(\theta_i) \left(\prod_{j=1}^{m_i} p(z_{i,j} \mid \theta_i)\, p(w_{i,j} \mid \beta_{1:K}, z_{i,j}) \right) \tag{1}$$

Il est possible d'établir ce que l'on appellera un postérieur, c'est à dire la distribution des variables cachées conditionnellement aux variables observées. Ici, on aura assez simplement :

$$p(\beta_{1:K}, \theta_{1:n}, z_{1:n} \mid y_{1:n}) = \frac{p(\beta_{1:K}, \theta_{1:n}, z_{1:n}, y_{1:n})}{p(y_{1:n})} \tag{2}$$

Le numérateur correspond à la distribution jointe de toutes les variables, qui peut être assez facilement calculé pour n'importe quelle configuration des variables cachées à partir de l'Équation 1. Le dénominateur correspond à la probabilité marginale des observations, c'est à dire la probabilité de rencontrer le corpus observé sous n'importe quel configuration de sujets. Le nombre de telles configurations est exponentiellement grand, et cette somme est donc impossible à calculer puisqu'il est nécessaire de considérer toutes les manières possibles d'assigner chaque mot observé à chaque sujet (voir Blei, Ng et Jordan, 2003). Il s'agit d'un problème classique des statistiques bayésiennes, pour lesquelles des solutions d'approximations existent. La première, connue sous le nom d'échantillonage de Gibbs, consiste à construire une chaîne de Markov d'observations afin d'estimer les probabilités utilisées dans le postérieur. Steyvers et Griffiths, 2007 propose une explication complète du processus pour le LDA. Une autre méthode sont les algorithmes variationnels, qui, en résumé, s'appuient sur une famille de distribution pré-déterminée pour la structure cachée et tente de trouver le membre de cette famille le plus proche du postérieur, transformant le problème d'inférence en problème d'optimisation. Hoffman, Blei et Bach, 2010 propose par exemple une méthode d'estimation rapide pour LDA. Dans les deux cas, les paramètres du topic model optimisé comprennent en particulier la famille de distribution la plus vraisemblable $\theta_{1:n}$, qui à chaque document associe un vecteur de taille K correspondant à la distribution des sujets sur ce document. C'est ce vecteur que nous utiliserons comme projection du document.

3.3 Correlated Topic Model (CTM)

Introduit dans Blei et Lafferty, 2005 et Blei et Lafferty, 2007, le Correlated Topic Model (CTM) cherche à enrichir le modèle LDA en le rendant plus expressif en lui permettant de prendre en compte une covariance entre les sujets. En effet, il est plus réaliste d'imaginer que la présence de certains sujets va être corrélée avec la présence d'autres, ou au contraire avec l'absence de certains autres. Comme le résume la Figure 2, le processus génératif est le suivant :

1. Pour chaque document d_i, on tire la proportion de chaque sujet selon une loi logistique normale : $\theta_i \mid \mu, \Sigma \sim \text{LogNorm}(\mu, \Sigma)$, avec μ un vecteur de taille K et Σ une matrice de covariance.
2. Pour chaque mot $w_{i,j}$ du document d_i,
   (a) Choisir aléatoirement un sujet selon la proportion des sujets : $z_{i,j} \mid \eta_i \sim \text{Mult}(\theta_i)$.
   (b) Choisir aléatoirement un mot selon la distribution de ce sujet : $y_{i,j} \mid z_{i,j}, \beta_{1:K} \sim \text{Mult}(\beta_{z_{i,j}})$.

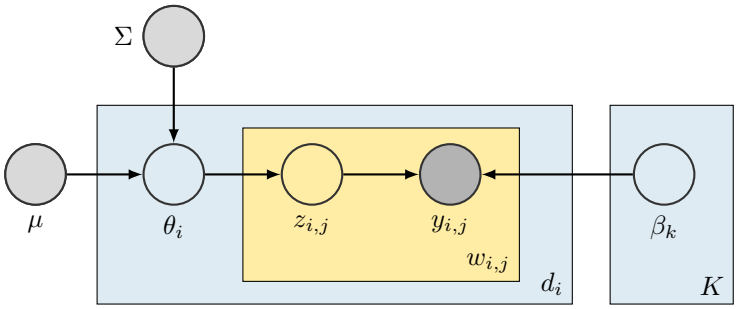

FIGURE 2 – Modèle graphique du CTM

Les changements par rapport au LDA sont assez minimes, si ce n'est que la proportion de sujet est maintenant tirée selon une loi logistique normale. Cette loi est une distribution de paramètre multinomiaux, et présente une alternative à la loi de Dirichlet en permettant des formes de corrélations plus complexes entre les composants du vecteur de paramètres. Celle-ci s'obtient de la manière suivante : dans un premier temps, on tire une variable intermédiaire selon une loi normale multidimensionnelle : $\eta_i \sim \mathcal{N}(\mu, \Sigma)$. Puis on normalise le vecteur ainsi obtenu par un softmax : $\theta_i = \exp(\eta_i)/\sum_k \exp(\eta_{i,k})$. C'est la variable cachée Σ, appelée matrice de covariance, qui permet de prendre en compte la corrélation entre les proportions des différents sujets dans un document, contrairement à la loi de Dirichlet utilisée dans le LDA pour lequel le tirage de chaque composant du vecteur de proportion est tiré selon les mêmes paramètres. Cette complexité supplémentaire introduite par LDA a cependant des inconvénients. En effet, le postérieur étant lui aussi impossible à calculer analytiquement, il faut à nouveau se tourner vers des méthodes d'approximation. Mais l'introduction de la loi logistique normale ne permet plus d'utiliser les méthodes d'échantillonage par les méthodes de Metropolis-Hastings permises uniquement par la simplicité des distributions du LDA, à cause de la forte dimensionalité des données. Il faut donc recourir à des algorithmes d'inférence variationnelle. Nous ne les détaillons pas ici.

3.4 Dynamic Topic Model (DTM)

Introduit dans Blei et Lafferty, 2006, le Dynamic Topic Model (DTM) cherche à prendre en compte le fait que les sujets peuvent changer avec le temps. Il postule que le corpus peut être découpé en époques, c'est à dire des blocs de documents de la même date. Chaque époque dispose de ses sujets propres, et pour une période fixée le processus génératif est très similaire à celui du LDA, comme le résume la Figure 3. DTM postule également qu'un sujet à une période donné dérive du même sujet dans la période précédente ; il définit donc également le modèle d'évolution d'un sujet entre les époques. Ce modèle séquentiel s'appuie sur une distribution logistique normale, cette fois-ci non pas pour établir une corrélation entre la proportion des sujets dans les documents θ comme pour CTM, mais entre un même sujet aux temps t et $t+1$. Comme résumé dans la Figure 3, le processus génératif pour un pas de temps t fixé est le suivant :

1. Tirer les K sujets selon : $\beta_{k,t} \sim \text{LogNorm}(\beta_{k,t-1}, \sigma^2 I)$
2. Pour chaque document d_i, on tire la proportion de chaque sujet une loi symétrique de Dirichlet : $\theta_{i,t} \sim \text{Dir}_K(\alpha)$.
3. Pour chaque mot $w_{i,j}$ du document d_i,
   (a) On tire aléatoirement un sujet : $z_{i,j,t} \mid \theta_{i,t} \sim \text{Mult}(\theta_{i,t})$.
   (b) On tire aléatoirement un mot selon une loi sur la distribution de ce sujet : $y_{i,j,t} \mid z_{i,j,t}, \beta_{1:K,t} \sim \text{Mult}(\beta_{z_{i,j,t}})$.

Grâce à l'utilisation de la loi logistique normale, le tirage des paramètres du sujet k de l'époque t est centré sur la moyenne de ce sujet à la période $t-1$, et la matrice de covariance est diagonale (I désigne la matrice identité, σ un paramètre de dérive entre les époques) pour permettre l'indépendance des composants.

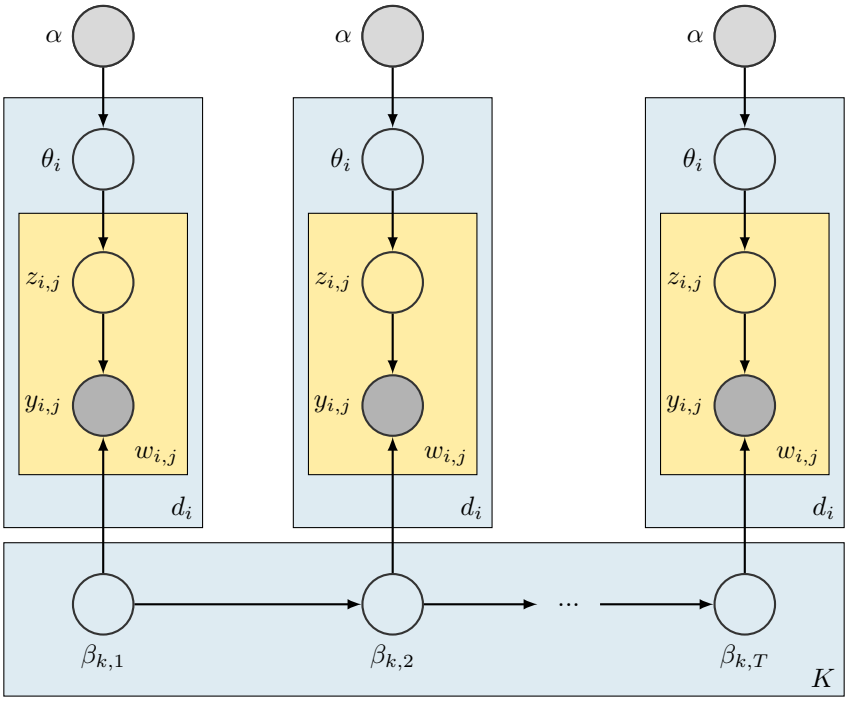

FIGURE 3 – Modèle graphique du DTM

3.5 Structural Topic Models (STM)

Introduit dans Roberts, Stewart, Tingley et al. (2014), et décrit en détail dans Roberts, Stewart et Airoldi, 2016, le Structural Topic Models (STM) propose l'innovation centrale de pouvoir s'appuyer sur un certain nombre de covariables pour les distributions des proportions de sujets θ et des sujets β. Au lieu de faire l'hypothèse que l'importance de chacun des sujets dans un document, et que le contenu des sujets (en terme de mots utilisés) sont uniformes pour toutes les observations comme LDA, ou à travers les époques comme DTM, STM permet l'incorporation de covariables afin de créer des groupes uniformes (i.e. avec les mêmes valeurs de covariables) au lieu d'une uniformité globale. Deux nouvelles variables servent à cette segmentation : X_i, un vecteur de covariables sur la distribution des sujets pour le document i, et χ_i, le vecteur de covariables sur la probabilité d'un mot d'être associé à un sujet pour le document i. En pratique, X_i va permettre une segmentation au niveau du document : tous les documents d'une année, tous les documents d'un pays, etc. χ_i va quand à lui servir à segmenter le vocabulaire au niveau de tous les documents du groupe, en intégrant dans β des paramètres reflétant l'écart de distribution du vocabulaire de ce groupe avec l'ensemble du corpus. Ce sont les variables κ présentées par la suite, dont nous ne détaillons pas le calcul. De ce fait il existe maintenant une distribution β par groupe de documents, β_i, pour permettre cette différentiation. Le processus génératif de STM, présenté graphiquement dans la Figure 4, peut ainsi être résumé de la sorte :

1. Pour chaque document d_i, tirer la distribution des sujets selon une loi logistique normale basée sur le vecteur de covariables du document : $\theta_i \mid X_i \cdot \gamma, \Sigma \sim \text{LogNorm}(X_i \cdot \gamma, \Sigma)$ avec $\gamma \sim \mathcal{N}(0, \sigma_k^2 I)$ un coefficient de régularisation propre au sujet et σ et Σ des hyperparamètres fixés.
2. Pour chaque sujet k, on tire la distribution (propre au document) sur les mots en utilisant la distribution basique m, la déviation propre au sujet κ_k, la déviation propre au groupe de covariable κ_{g_i} et le terme d'interaction des deux $\kappa_{(k,g_i)}$: $\beta_{i,k} = \exp(m + \kappa_k + \kappa_{g_i} + \kappa_{(k,g_i)}) / \sum_v \exp(m_v + \kappa_{k,v} + \kappa_{g_i,v} + \kappa_{(k,g_i),v})$.
3. Pour chaque mot $w_{i,j}$ de d_i :
   (a) Tirer aléatoirement un sujet selon la distribution des sujets du document : $z_{i,j} \mid \theta_i \sim \text{Mult}(\theta_i)$.

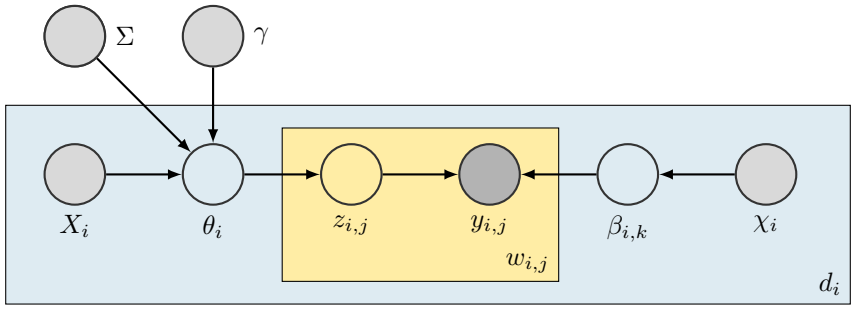

FIGURE 4 – Modèle graphique du STM

(b) Choisir aléatoirement un mot selon la distribution de ce sujet : $y_{i,j} \mid z_{i,j}, \beta_{i,z_{i,j}} \sim \text{Mult}(\beta_{z_{i,j}})$.

La postérieur qui en découle, comme pour les autres topic models, ne peut être calculer analytiquement et est donc approximé par un algorithme EM variationnel (VEM).

3.6 Hierarchical Dirichlet Process (HDP)

La conception de la famille des topic models hiérarchiques répond à l'idée que le nombre de sujets K devrait, comme les autres paramètres du modèle, être déduit des données plutôt que d'être fourni par l'utilisateur. Le plus connu d'entre eux est sûrement le Hierarchical Dirichlet Process (HDP), introduit dans Teh et al., 2006. La particularité de HDP est notamment qu'elle détermine le nombre de sujets au cours de l'exploration du corpus (il s'agit d'une méthode bayésienne non-paramétrique). En particulier, de nouveaux documents peuvent « créer » de nouveaux sujets. Notons que HDP implique un processus de Dirichlet (DP). Ce processus fournit une distribution à partir d'un espace arbitraire, et est dénoté $G \sim \text{DP}(\alpha, G_0)$ où G est donc un distribution sur un espace, α est un scalaire correspondant à la précision, et G_0 une distribution connue sur le même espace que G. Dans le modèle HDP, puisque l'on ne dispose pas du nombre K de sujets, on remplace la variable θ_i correspondant à la proportion de chaque sujet dans d_i, par une distribution sur les sujets G_i sur les sujets, tirée selon un processus de Dirichlet. Les atomes de G_i sont des sujets, c'est-à-dire on le rappelle des distributions multinomiales sur le vocabulaire V. Cependant, afin de respecter l'hypothèse du LDA que les sujets sont communs aux différents documents, il est nécessaire que la distribution de base G_0 soit commune à tous les documents. On ajoute simplement l'hypothèse que G_0 suive également un processus de Dirichlet, soit $G_0 \sim \text{DP}(\gamma, H)$ où H est une loi symétrique de Dirichlet sur le vocabulaire V. Finalement, le processus génératif de HDP, résumé dans la Figure 5, est le suivant :

1. Tirer la distribution de base sur les sujets $G_0 \sim \text{DP}(\gamma, H)$
2. Pour chaque document d_i, tirer sa distribution propre sur les sujets $G_i \sim \text{DP}(\alpha, G_0)$
3. Pour chaque mot $w_{i,j}$ de d_i :
   (a) Choisir aléatoirement un sujet pour le mot : $\beta_{i,j} \sim G_i$.
   (b) Choisir aléatoirement un mot selon la distribution de ce sujet : $y_{i,j} \mid \beta_{1:K} \sim \text{Mult}(\beta_{i,j})$.

De la même manière que pour le LDA, l'inférence exacte du postérieur pour HDP n'est pas possible à résoudre. Le même type de techniques, par échantillonage ou par des approches variationnelles sont utilisées ; nous ne les détaillerons pas.

3.7 Pseudo-Document-Based Topic Model (PTM)

Introduit dans Zuo et al., 2016, Pseudo-Document-Based Topic Model (PTM) appartient à la famille plus large des Short Texts Topic Models (STTM). Ces modèles cherchent à contourner le constat fait que les topic models classiques performent assez mal sur les textes courts, pourtant répandus par exemple sur les réseaux sociaux comme Facebook ou Twitter La principale raison identifiée est le manque d'information de co-occurrence de mots dans ces textes courts, qui ne conviennent pas aux techniques d'échantillonage

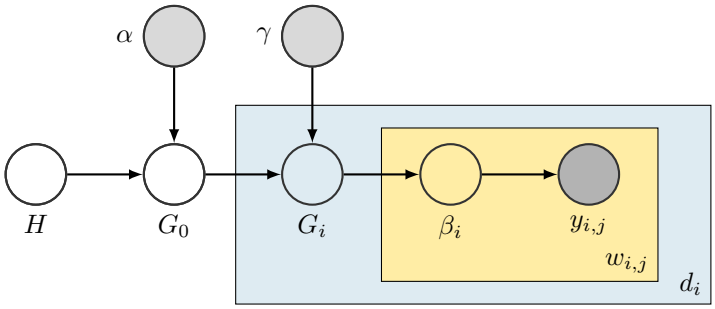

FIGURE 5 – Modèle graphique du HDP

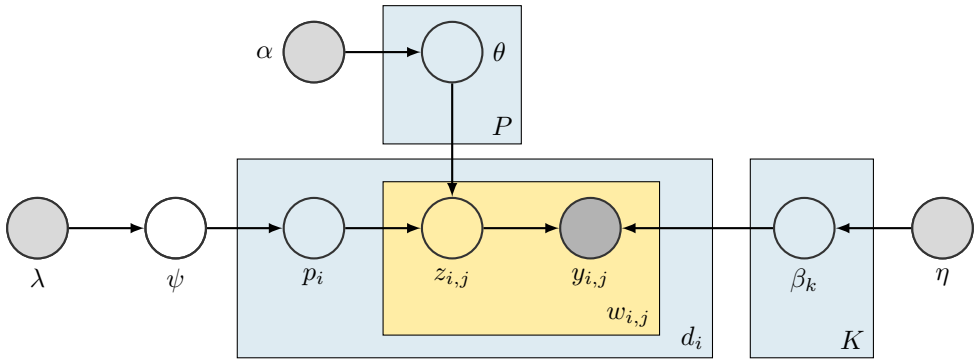

FIGURE 6 – Modèle graphique du PTM

des topic models classiques La solution proposée par PTM est d'utiliser une stratégie d'agrégation des documents. L'hypothèse sur laquelle s'appuie PTM est que le très grand nombre de textes courts disponibles sont en fait générés à partir de documents latents plus larges et de taille assez régulière, appelés pseudo-documents. L'idée est ainsi d'apprendre les sujets sur ces pseudo-documents dont le nombre est fixé à P. Résumé dans la Figure 6, le processus génératif global est le suivant :

1. On tire une distribution sur les pseudo-documents : $\psi \sim \text{Dir}_P(\lambda)$
2. On tire chaque sujet k selon une distribution de Dirichlet sur le vocabulaire : $\beta_k \sim \text{Dir}_V(\eta)$
3. Pour chaque pseudo-document p, on tire le vecteur de proportion des sujets : $\theta_p \sim \text{Dir}_K(\alpha)$
4. Pour chaque document d_i :
   (a) On tire le pseudo-document dont il est extrait : $p_i \sim \text{Mult}(\psi)$
   (b) Pour chaque mot $w_{i,j}$ du document d_i :
      i. On tire un sujet selon la proportion des sujets du pseudo-document : $z_{i,j} \mid \theta, p_i \sim \text{Mult}(\theta_{p_i})$
      ii. On tire aléatoirement le mot selon cette distribution : $w_{i,j} \mid z_{i,j}, \beta_{1:K} \sim \text{Mult}(\beta_{z_{i,j}})$

L'appartenance de chaque document à un pseudo-document n'est pas une variable observée, et participe donc à l'accroissement de la complexité du modèle.

4 Données

4.1 Prises de parole dans les médias

Le jeu de données sur lequel nous travaillons dans ce cas d'usage provient de l'Institut national de l'audiovisuel (INA), institution en charge de collecter et d'archiver toute diffusion audiovisuelle en France. La base sur laquelle nous travaillons recense l'ensemble des programmes radio qui, entre 1980 et 2016, ont porté sur la Russie. Pour la construire, nous avons récupéré l'ensemble des entrées pour lesquelles les termes

“russe*”, “russie”, “soviet*” ou “urss” apparaissent dans leurs champs descriptifs (titre, résumé, mots-clés, etc.). Au total, cela représente 120 882 entrées. La compréhension que nous avons de ce qu'est une émission portant sur la Russie est assez vaste, mais ce qui pourrait être une limitation est en réalité justifiée : toute évocation de la Russie participe à construire l'image du pays dans les esprits, et ce large ensemble nous permettra de disposer d'une palette plus large de thèmes et de types d'interventions dans lesquelles interviennent experts et non experts. Il est important de noter que nous ne disposons, pour chaque émission, que d'un résumé de ce qui y a été dit, et non une retranscription exacte du contenu. Cette caractéristique rend le travail de classification plus complexe, mais aussi plus intéressante puisque l'efficacité de l'extraction d'information est encore plus importante.

Un important nettoyage de la base a été nécessaire avant son utilisation. D'abord, nous n'avons conservé que les émissions pour lesquelles un résumé était fourni, soit seulement 25 439 entrées, 21% du jeu de données original. Nous avons ensuite choisi de ne conserver que les émissions n'étant pas des journaux d'informations, puisque ceux-ci se sont avérés n'avoir presque aucun intervenant (mis à part le présentateur), et nous avons également retiré toutes les émission sans aucun intervenant. Nous qualifions d'intervenant toute personne qui intervient dans une émission en tant qu'invités, c'est-à-dire en excluant journalistes et présentateurs. Cela nous laisse avec un total de 10 857 émissions. Pour chaque émission, nous avons ensuite procédé à une transformation consistant à attribuer à chaque intervenant le « segment » d'émission qui lui correspond. A partir de la liste d'intervenant, et grâce à la structuration des résumés, il a en effet été possible de découper une émission en différents segments et de les associer au bon intervenant. Cette dernière présentation de la base dite « d'interventions » dispose de 20 376 entrées.

Avant toute utilisation du texte des résumés, celui-ci est nettoyé selon une procédure standard (suppression des stopwords et des mots apparaissant moins de 2 fois dans le corpus, des caractères non alphabétiques, conversion en minuscule, lemmatisation puis tokenisation). Puisque notre but est de travailler uniquement à l'aide du texte pour tenter de prédire le degré d'expertise de l'intervenant, nous avons également supprimé toute mention du nom de ce dernier dans le texte, ce qui se produit souvent puisque celui-ci est un résumé de l'intervention.

4.2 Définition exogène de l'expertise des intervenants

La seconde phase de préparation nécessite de créer des étiquettes pour ces interventions, entre experte (annoté 1) ou non experte (resp. 0), et ce d'une manière exogène ne dépendant pas du texte, pour pouvoir ensuite soumettre à un classifieur la tâche de retrouver l'expertise à partir de ce dernier. En apparence assez simple, cette tâche d'étiquetage s'inscrit dans un débat important de la définition de l'expertise. De nombreuses écoles s'opposent, chacune avec ses catégories et ses critères de classification. Collins et Evans (2007) expliquent par exemple qu'il ne faut pas définir les experts par le biais de la possession de connaissance (qu'ils qualifient d'*« expertise réelle et substantive »*), mais de manière relationnelle : l'expert est celui qui est reconnu comme tel par les autres experts établis. Cette définition est caractérisée par un processus social, c'est-à-dire qu'on est expert grâce à un statut social plus que grâce à une expertise. Une piste simple que nous choisissons de suivre est celle de considérer comme expert les personnes membre du corps académique. Celle-ci répond partiellement à la définition de Collins et Evans puisqu'il faut avoir été reconnu par d'autres académiques pour en devenir un, par le biais de la soutenance d'une thèse. Albæk, Christiansen et Togeby (2003) proposent dans leur article trois arguments forts qui appuient cette décision :

> *The first problem encountered in investigating whether experts appear more often in news coverage today is that the meaning of the term “expert”, both in the social science literature and in the view of the general public, has changed over the years to include a broader range of actors. In order to address this problem, we have included in our study only a very specific subset of experts – namely, scientific researchers working at independent research institutions – and we have done so for three reasons : (1) irrespective of whether the narrow, classical or the broader, modern definition of “expert” is used, few would disagree that scientific researchers,*

provided that they work within intellectually independent institutions, constitute a proper subset of experts ; (2) the definition of this subset (i.e., of a scientific researcher) will be the same for the entire forty-year period for which we collected data ; and (3) this subset is empirically easy to operationalize.

Le troisième argument se vérifie en effet en pratique. Nous utilisons ainsi la règle suivant pour déterminer si un intervenant (et donc, par extension, ses intervention) est expert : toute personne qui (a) est renseignée comme telle par la nomenclature de l'INA ou (b) a écrit ou dirigé une thèse, un article scientifique, ou un livre, qui porte sur le thème de la Russie. Le critère (a) s'appuie sur un recensement partiel que réalise l'INA des intervenants : il résume à part la qualité de l'intervenant, c'est à dire son métier et sa nationalité. Nous considérons comme académiques les métiers de "chercheur", "universitaire", "économiste", "historien", etc. Le critère (b) est établi par une exploration de différentes bases de données externes, dont par exemple theses.fr, le catalogue de la Bibliothèque nationale de France (BnF), et de systèmes d'archivages de publications en ligne. Cette convention nous permet d'étiqueter toutes les interventions en fonction du statut, expert ou non, de son intervenant. Au total, nous obtenons environ 23% d'interventions expertes, et le reste est étiqueté non-expert par défaut. Ce déséquilibre des étiquettes requiert un processus d'équilibrage que nous décrivons dans l'Appendice A.

5 Résultats

5.1 Modèles de référence

Afin d'analyser la performance des topic embeddings que nous venons de détailler, nous les soumettons à une comparaison avec quatre autres importantes approches de projections de document existantes. Le modèle le plus simple, et pourtant toujours le plus répandu, est celui du sac de mots, ou bag of words (**BoW**, Harris, 1954). Plutôt que d'utiliser le compte brut, nous utilisons une normalisation par TF-IDF, pour prendre en compte le fait que les mots très fréquents dans tous les documents apportent une information plus réduite. Puisque nous souhaiterons fixer la taille de la projection à n'importe quel K comme pour les autres approches, nous garderons uniquement les K mots les plus fréquents. Le second modèle est la projection **doc2vec** (Le et Mikolov, 2014). Le troisième (**Pool**) est une architecture simple mais qui s'est montrée performante dans la littérature : on obtient d'abord une projection des mots du corpus, puis, pour chaque document, on construit sa projection en agrégeant les projections des mots qui le composent. Nous utiliseront la moyenne comme fonction d'agrégation et word2vec comme projection des mots (Mikolov et al., 2013). Enfin, nous utiliserons également le modèle **LSA** (Deerwester et al., 1990). LSA repose sur une décomposition en valeurs singulières (SVD) de la matrice d'occurrence des mots (où les lignes correspondent aux documents et les colonnes aux mots) afin de réduire sa dimensionalité à K. Il s'agit en quelque sorte de l'ancêtre des topic models, puisque les mots (colonnes) vont être regroupés sous forme de concepts en fonction de leur proximité d'utilisation, en quelque sorte équivalents aux sujets des topic models. La principale différence est que LSA ne repose pas sur l'hypothèse d'un modèle stochastique génératif.

Le code de l'ensemble de modélisation et des tests réalisés pour obtenir les qui suivent est accessible en ligne. [1] Si les approches les plus simples reposent sur les implémentations de la librairie gensim, [2] nous utilisons plusieurs autres langages et librairies pour pré-calculer les projections ensuite utilisées en test.

5.2 Évaluation des classifications

Nous utilisons 3 métriques pour déterminer la performance de nos projections, calculées sur un sous ensemble des données sur lequel l'algorithme n'a pas été entraîné (dites données de test), qui compte environ 3000 observations. Les trois mesures sont définies en fonction des 4 classes possibles pour chaque observation : vrais positifs (VP), faux positifs (FP), vrais négatifs (VN) et faux positifs (FP). La justesse

1. https://github.com/jdenes/ina
2. https://radimrehurek.com/gensim

est définie par $(\mathrm{VP}+\mathrm{VN})/(\mathrm{VP}+\mathrm{FP}+\mathrm{VN}+\mathrm{FN})$ et donne une mesure globale de la capacité de l'algorithme à correctement identifier experts et non-experts. La précision donne une mesure de la capacité du modèle à ne pas faussement identifier des personnes qui ne seraient pas expertes comme expertes, soit $\mathrm{VP}/(\mathrm{VP}+\mathrm{FP})$. Le rappel mesure lui plutôt la capacité à retrouver le plus grand nombre possible d'experts, soit $\mathrm{VP}/(\mathrm{VP}+\mathrm{FN})$. La Table 1 présente les résultats obtenus (en pourcentage) avec une régression logistique comme classifieur, et la Table 2 ces mêmes résultats avec une machine à vecteurs de support (SVM). Notons que l'Appendice B détaille la procédure de choix des classifieurs et de sélection des tailles de projections, calibrées pour maximiser les performances.

Modélisation	Taille du vecteur	Justesse	Précision	Rappel
BoW	3000	**75.97**	**74.23**	77.91
Doc2Vec	2000	67.64	65.45	75.37
Pool	2000	68.66	67.81	71.03
LSA	2000	75.94	73.33	**81.12**
LDA	2000	71.49	69.21	74.06
HDP	150	54.16	52.40	74.26
DTM	2000	00.00	00.00	00.00
CTM	2000	00.00	00.00	00.00
STM	1000	70.35	70.31	70.91
PTM	1000	63.31	61.18	67.84

Note : les valeurs 00.00 indiquent que la modélisation n'est pas encore prête.

TABLE 1 – Performances par projection pour une régression logistique

Modélisation	Taille du vecteur	Justesse	Précision	Rappel
BoW	3000	**85.09**	**88.25**	80.90
Doc2Vec	2000	72.45	67.90	**85.63**
Pool	2000	73.49	72.29	76.17
LSA	2000	75.51	71.46	84.50
LDA	2000	80.41	79.23	81.65
HDP	150	54.51	52.31	85.00
DTM	2000	00.00	00.00	00.00
CTM	2000	00.00	00.00	00.00
STM	1000	72.13	70.80	75.76
PTM	1000	66.77	63.34	75.71

TABLE 2 – Performances par projection pour une SVM

6 Analyse

6.1 Interprétation de la classification

Notre méthodologie s'inscrivant dans un objectif d'utilisation par et pour les sciences sociales, il est central à nos yeux de permettre une interprétabilité de la classification pour permettre ensuite aux chercheurs de mieux comprendre les structures sous-jacentes, qu'elles soient sociales, politiques, et même linguistiques. Nous proposons une analyse selon la méthodologie suivante : nous entraînons une régression logistique, puis nous identifions quels sont les dimensions dont les coefficients sont les plus élevés (resp. les plus faibles). Cela signifie que ces dimensions jouent un fort rôle dans la reconnaissance d'une intervention

experte (resp. non experte). Or ces dimensions correspondent, pour les projections par topic models, à un sujet précis. Nous montrons ainsi la facilité d'interprétation des résultats, qui permettre de comprendre la classification contrairement à des approches de projections très abstraites comme doc2vec ou l'agrégation de projections de mots.

La Table 3 montre les 5 dimensions thématiques possédant les coefficients les plus positifs, dans une régression logistique sur un vecteur de taille 2000 construit par LDA. Ce sont dont les dimensions qui caractérisent le plus l'expertise. Les mots affichés correspondent aux termes mots plus significativement associés à ce sujet. De même, la Table 4 propose les mêmes informations pour les dimensions aux coefficients les plus négatifs, i.e. qui caractérisent le plus la non-expertise.

Dimension	**266**	**123**	**209**	**1208**	**1653**
Valeur du coef.	4.549	3.905	3.465	2.984	2.568
Mots associés	russi	carrer	est	syri	politiqu
	russ	politiqu	puissanc	bachar	pologn
	rappel	soviétiqu	occidental	assad	varsovi
	poutin	problèm	guerr	international	est
	syri	union	russi	islamiqu	épreuv
	obama	pai	conflit	etat	pari
	syrien	roug	daech	daech	europ
	pari	gorbatchev	international	recherch	franc
	attentat	société	iran	avec	anné
	international	urss	ryad	vedrin	lien

TABLE 3 – Dimensions aux coefficients les plus positifs et mots associés

Dimension	**1373**	**402**	**984**	**594**	**1645**
Valeur du coef.	−5.918	−5.093	−4.991	−4.780	−4.399
Mots associés	équip	film	brésil	personnag	footbal
	basket	cinéma	piano	film	perch
	dopag	est	prochain	ell	saut
	médaill	cann	musiqu	vie	olympiqu
	athlèt	réalisateur	commerc	actric	ligu
	natat	jour	disqu	travail	athlétism
	interview	franck	interprété	souvient	bastia
	féminin	plage	invité	lui	donetsk
	tenni	allemand	pianist	strada	monaco
	franc	festiv	chanson	acteur	heur

TABLE 4 – Dimensions aux coefficients les plus négatifs et mots associés

On identifie ainsi facilement les pistes de classifications suivies par l'algorithme : si l'intervention parle des relations entre Russie et États-Unis (dim. 266), de la fin de l'URSS (dim. 123), de conflits au Moyen-Orient (dim. 209 et 1208), ou encore de politique polonaise (dim. 1653), le document est très probablement une intervention experte. Au contraire, si elle concerne le sport (dim. 1373 et 1645), le cinéma (dim. 402 et 594) ou la musique (dim. 984), elle est plus probablement non-experte.

D'autres projections permettent une interprétation rapide des composants du textes qui permettent ou non d'affirmer l'expertise de l'intervenant. On pense par exemple à la représentation par sac de mot, où chaque dimension correspond à un mot du dictionnaire. Comme on le constate dans la Table 5, l'interpré-

tation est bien moins facile que pour le LDA. Les mots sont isolés de tout contexte, et ne forment pas un vocabulaire cohérent pour comprendre instinctivement la logique de classification. Le classifieur semble « tricher » en extrayant les mots décrivant explicitement la qualité de l'intervenant ("historien", "économiste") mais pas les thèmes évoqués. L'approche par topic models permet une plus grande lisibilité des dimensions et ainsi de réellement comprendre la cohérence thématique de la classification.

Positifs	historien	économist	russi	spécialist	carrer
	sibéri	soviétiqu	novembr	goulag	astronaut
	russ	politologu	stratégiqu	économiqu	tsar
Négatifs	président	équip	concour	respons	rencontr
	concert	cinéast	instrument	match	nobel
	ministr	musical	discour	propos	sida

TABLE 5 – Mots associés aux coefficients les plus positifs et négatifs (sac de mot)

6.2 Utilisation pour une classification automatique des intervenants

Au delà de la simple interprétation, nous souhaiterions également étudier à quel point notre modèle est généralisable, et plus performant qu'un l'étiquetage arbitraire que nous avons initialement produit. En effet, qu'est ce qui garantit qu'un invité qui n'est pas un académique ne fasse pas une intervention de la qualité de celle d'un expert ? Ou au contraire, qu'un expert académique vienne en réalité partager ses impressions sur un match de football ? Les individus sur lesquels nous ne disposons pas d'informations sont également classés comme non experts par défaut, et nous souhaiterions par exemple pouvoir les caractériser de manière plus efficace. Nous proposons ainsi d'étudier les faux positifs et les faux négatifs produits par le modèle de classification, afin de d'étudier la capacité de nos projections à proposer une classification automatique performante. Notre méthode consiste simplement à produire une étude au niveau micro des faux positifs avec le score le plus haut (i.e. ceux pour lequel l'algorithme est le plus sûr que la classification que nous lui proposons est erronée), et de même les faux négatifs pour lesquels le score est le plus faible. Nous étudions la personnalité de l'intervenant, mais aussi l'intervention en elle-même pour tenter de comprendre pourquoi leur discours ne colle pas à leur rôle.

Nous nous basons toujours sur le résultat d'une régression logistique appliquée à une projection par LDA de taille 2000, sur lequel on a obtenu une justesse de 71.77%. Plutôt que de travailler sur les données de test, qui contiennent des observations synthétiques dues au sur-échantillonnage, nous travaillons sur les 20 376 vraies observations, y compris celles utilisées en entraînement. On obtient alors une justesse de 69.72%. La Table 6 nous montre quels sont les intervenants les plus souvent mal classés. Les trois premières colonnes correspondent aux faux positifs, avec le compte d'interventions mal classées et le compte total d'interventions, et les trois dernières donnent les mêmes informations pour les faux négatifs. On constate d'abord que pour les intervenants dont les interventions sont mal classées de nombreuses fois, la part d'interventions mal classées est souvent assez haute (FP/Total ou FN/Total proche de 1). Cela signifie que c'est la personne toute entière qui est mal catégorisée, et non certaines de ces interventions, puisque celles-ci tombent systématiquement dans la mauvaise catégorie.

Détaillons les faux positifs : elle contient de nombreux « connaisseurs » non académiques de la Russie, ayant acquis une expérience grâce à leur rôle politique. Ainsi François Mitterand est intervenu de nombreuses fois en tant que Premier secrétaire du Parti socialiste (1971-1981) pour commenter les évènements liés à la chute de l'URSS ; de même pour Georges Marchais en tant que Secrétaire général du Parti communiste français (1972-1994). Ou encore Hubert Védrine, Bernard Kouchner, Michel Jobert ou encore Laurent Fabius, tous anciens ministres des Affaires étrangères. On notera cependant que la présence de Jacques Chirac, Vladimir Poutine ou encore de Boris Eltsine illustre le bruit présent inhérent à l'utilisation de données « naturelles », puisque les personnes dont on entend les voix dans des interviews ou des images

sonores ont tendance à être référencées comme intervenants dans les émissions par l'INA. De manière plus intéressante, on découvre des personnalités dont l'expertise semble insoupçonnée selon nos critères mais facilement vérifiable : Richard Bakis était ainsi le représentant des États baltes en France lors de l'éclatement de l'URSS, et Jean-Pierre Haigneré est un spationautes français ayant œuvré à collaboration spatiale russo-européenne.

Intervenant	**Faux positifs**	**Total**	**Intervenant**	**Faux négatifs**	**Total**
Mitterrand François	48	150	Lischke André	38	63
Védrine Hubert	26	39	Dumas Roland	25	49
Chirac Jacques	26	65	Troyat Henri	22	33
Marchais Georges	16	43	Winock Michel	17	51
Poutine Vladimir	16	31	Fedorovski Vladimir	16	35
Fabius Laurent	15	25	Ferro Marc	13	31
Haigneré Jean Pierre	15	20	Sarkozy Nicolas	13	21
Eltsine Boris	15	24	Kahane Martine	12	12
Jobert Michel	13	20	Vitez Antoine	12	14
Bakis Richard	12	14	Adler Alexandre	11	40

TABLE 6 – Fréquence des intervenants les plus souvent mal classés

Les faux négatifs nous donnent également des informations intéressantes. André Lischke par exemple est considéré comme un expert puisqu'il est musicologue spécialisé sur la musique russe ; cependant il semble que notre algorithme se soit spécialisé sur les experts en géopolitique et il est donc souvent mal classé. Cette analyse est soutenue par le fait que c'est plutôt le vocabulaire de la géopolitique qui apparaît dans l'analyse des dimensions thématiques. De même, l'analyse des interventions de Henri Troyat et de Michel Winock révèle qu'elles se concentrent sur leurs ouvrages littéraires, hors du champ d'expertise politique. Il en va de même pour Martine Kahane, archiviste paléographe, ou Antoine Vitez, acteur et metteur en scène. D'autres montrent la limite de notre définition de l'expertise : si Nicolas Sarkozy et Roland Dumas sont considérés comme experts parce qu'ils ont écrit un livre ou un article sur la Russie (respectivement), ils sont avant tout des hommes politiques et leurs interventions sont naturellement plutôt classées comme non expertes. D'autres comme Vladimir Fedorovski, Marc Ferro ou encore Alexandre Adler sont difficilement classables puisqu'ils sont réellement à la frontière entre politique et académique.

Cette analyse simple et rapide illustre les capacités d'interprétations liées à l'utilisation des topic models. Nous avons ainsi pu comprendre quels sont les sujets associés aux interventions expertes (majoritairement liées à des sujets géopolitiques), mais aussi proposer une analyse critique des faux positifs et faux négatifs fréquents pour permettre une critique de la méthode scientifique utilisée pour l'étiquetage. La méthode de classification par topic models démontre ainsi être un outil puissant au service des sciences sociales. Dans le cas précis du projet d'étude de la Russie dans lequel ce travail s'inscrit, nous avons ainsi pu proposer une distinction efficace à des chercheurs en sciences sociales entre interventions expertes et non expertes, afin qu'il réalise un travail plus en profondeur sur la place de ces experts dans le débat radiophonique, par exemple en terme de changements en réaction aux évènements en lien avec la Russie. Avec au maximum 20% d'erreur, nous évitons à ces chercheurs un classement à la main, tout en leur permettant de vérifier rapidement les biais de classification grâce à des dimensions interprétables.

7 Discussion

7.1 Limites

Malgré des performances sur des tâches de classification prometteuses, et une capacité d'interprétation que nous venons de démontrer, la principale limite de l'utilisation des topic models comme projection

provient du temps de calcul qu'ils nécessitent. En effet, les principaux algorithmes de projections possèdent aujourd'hui des implémentations efficaces et optimisées, à l'image des projections doc2vec ou word2vec qui sont calculées en quelques minutes par la librairie gensim. Au contraire, les projections les plus complexes comme STM ou CTM nécessitent plusieurs jours de calcul simplement pour $K = 1000$. Deux facteurs peuvent expliquer la lenteur de l'optimisation des topic models. Le premier est bien entendu lié à leur complexité algorithmique : l'optimisation des paramètres des modèles repose bien souvent sur des algorithmes variationnels, à la complexité très forte, et dont le nombre d'itérations avant convergence n'est a priori pas connu. Mais le second est aussi lié au manque de publicité dont ces modèles disposent. En effet, si l'on considère par exemple le modèle LDA, celui-ci a été tellement étudié et utilisé que des efforts ont été faits pour optimiser son temps de résolution, à la fois de manière algorithmique avec l'utilisation d'un échantillonage de Gibbs (Hoffman, Blei et Bach, 2010) et dans le code, avec de nombreuses implémentations parallélisant les calculs. [3] Par conséquent, le modèle LDA converge lui aussi en quelques minutes sur gensim. Mais cette qualité de l'implémentation ne s'applique pas aux autres modèles, qui ne jouissent pas de la même renommée. Il y a donc encore des efforts à faire pour fournir des implémentations rapides de tous les modèles, dans l'idéal sur une plateforme centralisée où de nombreux modèles seraient disponibles. [4] Il faut en effet encore aujourd'hui jongler entre langages et librairies selon les modèles, ce qui pose un frein majeur à leur utilisation par des non-spéciales du domaine. Cela pourrait être un prolongement de ce travail.

Le cas d'usage présente lui aussi des limites qu'il faudra garder à l'esprit avant d'utiliser ses résultats. La première provient du type de contenu que nous avons utilisé : en effet, travailler sur des résumés est une tâche totalement différente de celle de travailler sur le texte original. Si l'on pourrait naturellement considérer qu'il s'agit d'une tâche plus complexe, puisque les différences de style et de langages auront tendance à être lissées, il serait nécessaire de tester cette théorie pour une plus grande robustesse des résultats. Une autre limite évidente provient de notre méthode d'étiquetage, et dont nous avons discuté plus tôt. Un étiquetage réalisé par des connaisseurs des études russes aurait été optimal, malheureusement infaisable pour notre base de donnée à plus de 20 000 entrées.

7.2 Travaux futurs

Un approfondissement naturel et nécessaire de ce travail serait évidemment de soumettre les projections par topic models à des tâches de classifications plus variées. En restant sur notre cas d'étude par exemple, nous pourrions les soumettre à une classification multiclasse en introduisant d'autres types de rôles ou d'expertises : les experts journalistiques, diplomatiques, politiques, etc. Mais il serait également intéressant de se détacher de cette tâche d'identification de l'expertise pour ce tourner vers des jeux de tâches plus standards dans la littérature. Conneau et Kiela, 2018 présente dans ce sens un ensemble de tâches standards, appelé SentEval, dont le but est de soumettre les modèles à des tâches variées pour en étudier tous les aspects. On pourrait ainsi évaluer la performance des topic models dans une situations où il n'est pas forcément attendu que les sujets entre étiquettes soient différents, contrairement à l'identification d'interventions expertes. Enfin, un autre approfondissement intéressant pourrait être d'étudier quelles sont les caractéristiques du langage capturées par ces projections, cette fois en terme syntaxique ou sémantique. Dans ce but, il serait intéressant de les soumettre aux tâches proposée dans Conneau, Kruszewski et al., 2018 conçues dans ce but.

Les nouvelles applications que nous espérons fournir grâce à ce travail relèvent principalement de l'utilisation des projections par topic models à des problématiques de sciences sociales, et sont à nos yeux nombreuses. L'une des principales que nous imaginons serait la réalisation d'un étiquetage automatique de données observées du « monde réel ». Aujourd'hui encore, une grande part de la recherche en sciences sociales doit recourir à une annotation à la main de documents, parce que les outils sont souvent trop complexes et trop peu compréhensibles pour les chercheurs. Comme pour l'étiquetage des experts dans le cas d'usage que nous avons présenté, nous espérons permettre une annotation simple et interprétable de docu-

3. https://radimrehurek.com/gensim/models/ldamulticore.html
4. Des initiatives semblent en chemin : https://github.com/RaRe-Technologies/gensim/issues/1038

ments. On pourra également penser à des usages plus larges, comme la détection automatique de fake news, l'analyse de la polarisation politique d'un tweet, la reconnaissance de la discipline d'un article à partir de son résumé, etc. Si ces tâches peuvent déjà être réalisées avec des projections plus classiques, les projections par topic models offrent la possibilité aux chercheurs de vérifier et de comprendre les contenus discriminants. Mais la tâche de classification peut également être vue comme un moyen d'enrichir l'information des topic models, puisqu'elle elle permet en effet de mettre en évidence les sujets les plus significatifs et propose ainsi une pondération de l'importance d'un sujet en fonction du contexte de la tâche.

8 Conclusion

Au cours de ce travail, nous nous sommes attachés à répondre à un objectif central, celui de documenter les projections par topic models et de présenter leur performances dans une tâche de classification de texte. Nous avons montré que ces projections atteignent, en terme de métriques de justesse, des performances semblables aux projections de documents les plus répandues aujourd'hui. Mais nous nous sommes également attachés à montrer que leurs performances en terme d'interprétabilité des résultats surpassent largement ces projections de références, qui ne sont que peu ou pas interprétables.

A plusieurs égards, nous avons cherché à proposer des nouveautés dans le champs de l'analyse du langage naturel. D'abord, nous avons proposé une nouvelle tâche de classification du langage, la détection d'expertise, qui pourrait être généralisée à d'autres types de corpus (journaux, tweets, textes politiques, etc.) et qui ouvre la porte à une analyse quantitative de la qualité d'un discours en terme d'expertise de son contenu. Nous avons également travaillé avec des résumés plutôt qu'avec des transcriptions exactes de prises de paroles, ajoutant un degré supplémentaire de difficulté à la tâche. Enfin, nous nous sommes attachés à proposer une analyse linguistique des résultats obtenus. Cet aspect se détache des travaux habituels de traitement du langage, qui favorisent plus souvent la complexité à l'explicabilité. La perspective dans laquelle s'inscrit ce travail, entre informatique et sciences sociales, nous a poussé sur le chemin de cette interprétabilité dans l'espoir de favoriser, entre autre, une meilleure compréhension des modélisations utilisées et des connaissances que les modèles informatiques peuvent révéler à d'autres disciplines.

Références

[1] Albæk, E., Christiansen, P. M. et Togeby, L. « Experts in the Mass Media : Researchers as Sources in Danish Daily Newspapers, 1961–2001 ». *Journalism & Mass Communication Quarterly* 80.4 (2003), p. 937–948. DOI : 10.1177/107769900308000412.

[2] Appadurai, A. « Disjuncture and Difference in the Global Cultural Economy ». *Theory, Culture & Society* 7.2-3 (1990), p. 295–310. DOI : 10.1177/026327690007002017.

[3] Blei, D. M. « Probabilistic Topic Models ». *Communication of the ACM* 55.4 (2012), p. 77–84. DOI : 10.1145/2133806.2133826.

[4] Blei, D. M., Carin, L. et Dunson, D. « Probabilistic Topic Models ». *IEEE Signal Processing Magazine* 27.6 (2010), p. 55–65. DOI : 10.1109/MSP.2010.938079.

[5] Blei, D. M. et Lafferty, J. D. « A correlated topic model of Science ». *The Annals of Applied Statistics* 1 (2007). DOI : 10.1214/07-AOAS114.

[6] Blei, D. M. et Lafferty, J. D. « Correlated Topic Models ». In : *Proceedings of the 18th International Conference on Neural Information Processing Systems*. NIPS'05. 2005, p. 147–154.

[7] Blei, D. M. et Lafferty, J. D. « Dynamic Topic Models ». In : *Proceedings of the 23rd International Conference on Machine Learning*. ICML'06. 2006, p. 113–120. DOI : 10.1145/1143844.1143859.

[8] Blei, D. M., Ng, A. Y. et Jordan, M. I. « Latent Dirichlet Allocation ». *Journal of Machine Learning Research* 3 (2003), p. 993–1022.

[9] Bowyer, K. W. et al. « SMOTE : Synthetic Minority Over-sampling Technique ». *Journal of Artificial Intelligence Research* 16 (2002), p. 321–357. DOI : 10.1613/jair.953.

[10] Cameron, D. et al. « A Taxonomy-Based Model for Expertise Extrapolation ». In : *2010 IEEE Fourth International Conference on Semantic Computing*. 2010, p. 333–340. DOI : 10.1109/ICSC.2010.27.

[11] Collins, H. et Evans, R. *Rethinking expertise*. University of Chicago Press, 2007. DOI : 10.7208/chicago/9780226113623.001.0001.

[12] Conneau, A. et Kiela, D. « SentEval : An Evaluation Toolkit for Universal Sentence Representations ». In : *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. 2018.

[13] Conneau, A., Kiela, D. et al. « Supervised Learning of Universal Sentence Representations from Natural Language Inference Data ». In : *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, p. 670–680. DOI : 10.18653/v1/D17-1070.

[14] Conneau, A., Kruszewski, G. et al. « What you can cram into a single $& !#* vector : Probing sentence embeddings for linguistic properties ». In : *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. 2018, p. 2126–2136.

[15] Das, R., Zaheer, M. et Dyer, C. « Gaussian LDA for Topic Models with Word Embeddings ». In : *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*. 2015, p. 795–804. DOI : 10.3115/v1/P15-1077.

[16] Dauphin, Y. N. et al. « Language Modeling with Gated Convolutional Networks ». In : *Proceedings of the 34th International Conference on Machine Learning*. PMLR 70. 2017.

[17] Deerwester, S. et al. « Indexing by latent semantic analysis ». *Journal of the American Society for Information Science* 41.6 (1990), p. 391–407. DOI : 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[18] Fang, D. et al. « Quantifying and Recommending Expertise When New Skills Emerge ». In : *2013 IEEE 13th International Conference on Data Mining Workshops*. 2013, p. 672–679. DOI : 10.1109/ICDMW.2013.33.

[19] Harris, Z. S. « Distributional Structure ». *Word* 10.2-3 (1954), p. 146–162. DOI : 10.1080/00437956.1954.11659520.

[20] He, H., Bai, Y. et al. « ADASYN : Adaptive synthetic sampling approach for imbalanced learning ». In : *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 2008, p. 1322–1328. DOI : 10.1109/IJCNN.2008.4633969.

[21] He, J., Hu, Z. et al. « Efficient Correlated Topic Modeling with Topic Embedding ». In : *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD'17. 2017, p. 225–233. DOI : 10.1145/3097983.3098074.

[22] Herling, R. W. « Operational Definitions of Expertise and Competence ». *Advances in Developing Human Resources* 2.1 (2000), p. 8–21. DOI : 10.1177/152342230000200103.

[23] Hoffman, M. D., Blei, D. M. et Bach, F. « Online Learning for Latent Dirichlet Allocation ». In : *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*. NIPS'10. 2010, p. 856–864.

[24] Hofmann, K. et al. « Integrating Contextual Factors into Topic-centric Retrieval Models for Finding Similar Experts ». In : *Proceedings of the SIGIR 2008 Workshopon Future Challenges in Expertise Retrieval (fCHER)*. 2008, p. 29–36.

[25] Jiang, D. et al. « Latent Topic Embedding ». In : *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*. 2016, p. 2689–2698.

[26] Kiros, R. et al. « Skip-Thought Vectors ». In : *Advances in Neural Information Processing Systems 28 (NIPS 2015)*. 2015, p. 3294–3302.

[27] Le, Q. et Mikolov, T. « Distributed Representations of Sentences and Documents ». In : *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. ICML'14. 2014, p. 1188–1196.

[28] Li, S. et al. « Generative Topic Embedding : a Continuous Representation of Documents ». In : *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. 2016, p. 666–675. DOI : 10.18653/v1/P16-1063.

[29] Liu, Y. et al. « Topical Word Embeddings ». In : *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015, p. 2418–2424.

[30] Martin, S. E. « Using Expert Sources in Breaking Science Stories : A Comparison of Magazine Types ». *Journalism Quarterly* 68.1-2 (1991), p. 179–187. DOI : 10.1177/107769909106800119.

[31] Mikolov, T. et al. « Distributed Representations of Words and Phrases and their Compositionality ». In : *Advances in Neural Information Processing Systems 26*. 2013, p. 3111–3119.

[32] Moody, C. E. « Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec ». *CoRR* 1605.02019 (2016). arXiv : 1605.02019.

[33] Pennington, J., Socher, R. et Manning, C. « Glove : Global Vectors for Word Representation ». In : *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, p. 1532–1543. DOI : 10.3115/v1/D14-1162.

[34] Peters, H. P. « The interaction of journalists and scientific experts : co-operation and conflict between two professional cultures ». *Media, Culture & Society* 17.1 (1995), p. 31–48. DOI : 10.1177/016344395017001003.

[35] Roberts, M. E., Stewart, B. M. et Airoldi, E. M. « A Model of Text for Experimentation in the Social Sciences ». *Journal of the American Statistical Association* 111.515 (2016), p. 988–1003. DOI : 10.1080/01621459.2016.1141684.

[36] Roberts, M. E., Stewart, B. M., Tingley, D. et al. « Structural Topic Models for Open-Ended Survey Responses ». *American Journal of Political Science* 58.4 (2014), p. 1064–1082. DOI : 10.1111/ajps.12103.

[37] Said, E. W. *Covering Islam : How the Media and the Experts Determine how We See the Rest of the World*. Vintage Books, 1997, p. 2000.

[38] Socher, R. et al. « Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection ». In : *Proceedings of the 24th International Conference on Neural Information Processing Systems*. NIPS'11. 2011, p. 801–809.

[39] Steele, J. E. « Experts and the Operational Bias of Television News : The Case of the Persian Gulf War ». *Journalism & Mass Communication Quarterly* 72.4 (1995), p. 799–812. DOI : 10.1177/107769909507200404.

[40] Steyvers, M. et Griffiths, T. « Probabilistic Topic Models ». In : *Latent Semantic Analysis : A Road to Meaning*. Sous la dir. de T. Landauer et al. Laurence Erlbaum, 2007, p. 427–448.

[41] Teh, Y. W. et al. « Hierarchical Dirichlet Processes ». *Journal of the American Statistical Association* 101.476 (2006), p. 1566–1581. DOI : 10.1198/016214506000000302.

[42] Wieting, J. et Kiela, D. « No Training Required : Exploring Random Encoders for Sentence Classification ». *CoRR* 1901.10444 (2019). arXiv : 1901.10444.

[43] Yan, X. et al. « A biterm topic model for short texts ». In : *International World Wide Web Conference (WWW)*. 2013.

[44] Yu, J. et al. « Topic embedding of sentences for story segmentation ». In : *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 2017, p. 1602–1607. DOI : 10.1109/APSIPA.2017.8282280.

[45] Zuo, Y. et al. « Topic Modeling of Short Texts : A Pseudo-Document View ». In : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD'16. 2016, p. 2105–2114. DOI : 10.1145/2939672.2939880.

Appendices

A Préparation des données

A.1 Deux présentations différentes de la base

Une fois les émissions divisées en segments, chaque segment correspondant à l'intervention d'un unique invité, un choix se présente entre deux modèles de présentation de la base de données. Une première mise en forme consiste à garder la base comme telle, et à avoir donc une entrée par intervention. On étiquette ensuite chaque segment en fonction du statut d'expertise de son intervenant. Le problème qui se pose avec une telle procédure est que l'étiquette ainsi obtenue pour l'intervention en elle-même ne reflète peut-être pas bien son contenu. Quid par exemple de l'académique venu présenter son livre sur un sujet totalement autre que la Russie, et qui en plus se retrouve à commenter les résultats du match de football de la veille dans la conversation lancées par les autres invités ? Pour éviter ce cas de figure, une autre possibilité de présentation serait de construire une méta-intervention pour chaque intervenant, qui concaténerait l'ensemble de ses interventions. On peut ainsi espérer que des cas de figure comme ceux que nous venons de décrire seraient « noyés » dans un ensemble d'interventions bien plus représentatif de l'intervenant et correspondant donc bien mieux à son étiquette. Les textes, en outre plus longs, seraient plus faciles à classifier. La contrepartie d'une telle présentation est cependant la diminution du nombre d'entrées de la base : 11 092 pour le premier format, contre 20 376 pour le premier. Enfin, ces deux présentations impliquent en réalité des tâches de classifications différentes : l'une consiste à attribuer une expertise à une intervention unique, l'autre à une personne (i.e. à la concaténation de ses interventions).

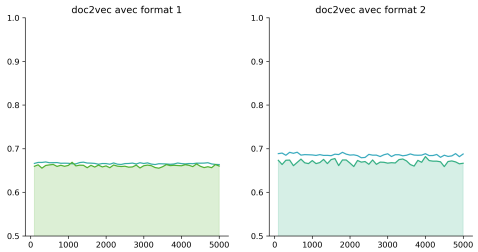

FIGURE A1 – Performances d'une régression logistique sur les deux formats

Face aux avantages et aux désavantages de ces deux présentations, le seul moyen de réellement les départager est de tester leurs performances en classification. La Figure A1 nous montre les performances pour les deux formats, avec une projection par doc2vec. L'axe des abscisses correspond à la taille de la projection, celui des ordonnées à la justesse. On remarque que les deux formats présentent des résultats similaires (66.87% et 68.18% en test respectivement). Le format 2 est cependant plus sujet au sur-apprentissage, puisque l'erreur en entraînement est plus faible mais ne se répercute pas en test. Il semble de plus plus intéressant de travailler sur une tâche d'attribution d'expertise à des interventions plutôt qu'à des individus. C'est pourquoi nous choisissons de travailler avec le premier format.

A.2 Équilibrage des données

Une fois la question de la présentation de la base de données réglée, se pose une nouvelle question sur l'équilibrage des observations. En effet, les interventions étiquetées comme expertes ne représentent que 22.65% des observations. Pour permettre une évaluation correcte des représentations par les algorithmes d'apprentissage automatique, il est nécessaire de leur fournir un jeu de données équilibré, avec à peu près autant d'observations de chacun des deux types d'étiquettes. Deux solutions sont possibles : des solutions de sous-échantillonage, où l'on retirerait des observations étiquetées non-expert, et des solutions de sur-échantillonage où l'on rajoute des observations étiquetées expert. Pour cette dernière solution, deux

procédures sont principalement utilisées : SMOTE, défini dans Bowyer et al., 2002, et ADASYN proposé dans He, Bai et al., 2008. Ces deux procédures créent de fausses observations de la classe minoritaire (dans notre cas la classe experte), appelées observations synthétiques. Si les procédures de sous-échantillonnage ont l'avantage d'être rapides et de ne pas biaiser l'apprentissage, elles réduisent cependant le nombre d'observations utilisables. Au contraire, les procédures de sur-échantillonnage l'augmentent, mais au prix d'un plus grand temps de calcul et surtout en créant de « fausses » observations. Un biais possible serait alors que l'algorithme d'apprentissage apprenne en fait à reconnaître ces observations synthétiques pour détecter l'étiquette minoritaire.

La Figure A2 nous montre le résultat d'un test placebo effectué pour détecter un tel biais des algorithmes de sur-échantillonnage. Le test est construit de la sorte : on remplace dans les données les vrais étiquettes d'expertise par des étiquettes générées aléatoirement. En abscisse, on fait varier la proportions de « faux experts » aléatoirement générés, et par conséquent le nombre d'observations synthétiques qui seront générées de manière inversement proportionnelle. On pourra ainsi évaluer si l'algorithme est biaisé par ces procédures et reconnaît plutôt les observations synthétiques que l'expertise. Si l'exactitude reste autour de 50%, alors il n'y a logiquement pas de biais puisque les étiquettes sont effectivement aléatoires et il n'y a donc aucun moyen de les reconnaître. Si, au contraire, l'exactitude augmente avec le nombre d'étiquettes à générer (c'est à dire une proportion d'étiquettes 1 très forte ou très faible), alors l'algorithme détecte en fait les étiquettes synthétiques. La Figure A2 nous montre ainsi que la régression logistique est assez peu sujette à un tel biais, de même que la SVM dans la Figure A3. En revanche, la Fig. A4 nous montre que la forêt d'arbres décisionnels est très sujette à ce biais, et nous pousse à éviter ce modèle dans la suite. Notons que dans l'ensemble des figures suivantes, l'exactitude sur les données de test est représentée par une courbe avec remplissage, et l'exactitude sur les données d'entraînement par la courbe sans remplissage.

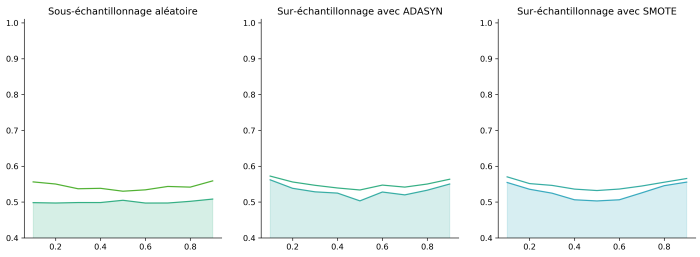

FIGURE A2 – Test placebo pour les méthodes de sur-échantillonnage avec une régression logistique

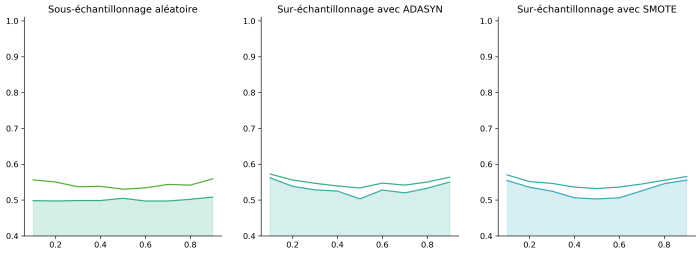

FIGURE A3 – Test placebo pour les méthodes de sur-échantillonnage avec une SVM

Enfin, la question d'intérêt plus central est de déterminer si les procédures de sur-échantillonnage permettent réellement d'obtenir de meilleurs résultats. On pourra, dans le cas contraire, s'en tenir à un sous-échantillonnage. La Figure A5 nous montre l'exactitude d'un modèle logit sur un vecteur sac de mots, avec en abscisse la taille du vecteur et en ordonnée d'exactitude, pour les trois méthodes d'échantillonage évoquées. Comme on le constate, les performances sont similaires sur les données d'entraînement, ce qui nous laisse penser que le modèle n'est pas biaisé par les procédure de sur-échantillonage. Mais le modèle entraîné par sous-échantillonnage est sujet au sur-apprentissage, et donc des performances inférieures sur les données de test, probablement à cause du nombre trop faible d'observations dont elle dispose. Face à

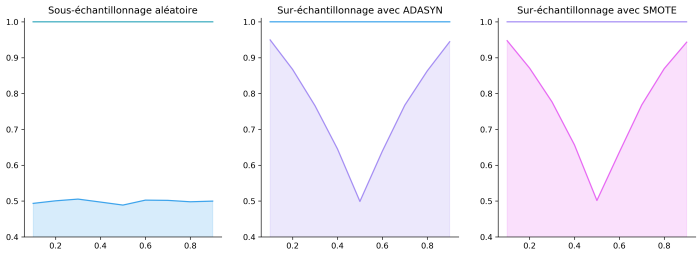

FIGURE A4 – Test placebo pour les méthodes de sur-échantillonnage avec une forêt aléatoire

cette observations, et au fait que les procédures SMOTE et ADASYN ont des performances similaires, nous choisissons d'utiliser dans l'ensemble de ce travail un sur-échantillonnage par SMOTE, plus rapide.

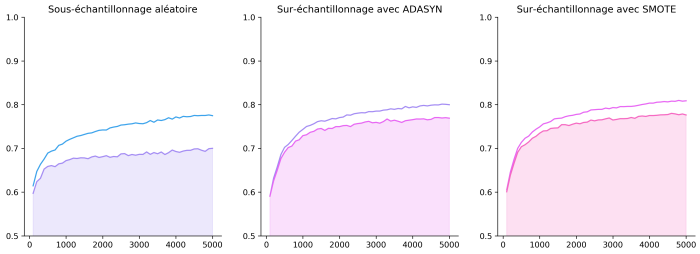

FIGURE A5 – Performance d'une régression logistique pour différentes méthodes de sur-échantillonnage

B Optimisation des modèles

B.1 Choix du modèle

Afin d'obtenir les meilleurs résultats, nous soumettons au test un modèle par grande famille de classifieur. On teste ainsi une régression logistique avec une pénalisation L2 (**logit**), un classifieur naïf de Bayes avec hypothèse de distribution gaussienne (**Bayes**), un k plus proches voisins avec $k = 3$ (**KNN**), un arbre de décision (**Arbre**) de profondeur maximale 100, un algorithme **AdaBoost** basé sur 100 arbres de décision de profondeur 1, une machine à vecteurs de support (**SVM**) doté d'un kernel RBF, et finalement un réseau neuronal artificiel (**RNA**) à trois couches de taille respectives 1000, 500 et 50 avec fonction d'activation RELU. La Table B1 nous présente la justesse de chacun de ces classifieur dans une validation croisée à 5 échantillons pour certaines projections de documents.

	BoW	doc2vec	LSA	LDA	DTM	STM	CTM
Logit	76.11	67.12	75.33	71.42	00.00	70.44	00.00
Bayes	76.05	58.81	62.77	60.85	00.00	67.15	00.00
KNN	81.45	**78.87**	75.35	71.56	00.00	**74.50**	00.00
Arbre	79.79	73.78	73.91	77.28	00.00	71.67	00.00
AdaBoost	76.44	70.34	71.99	70.01	00.00	71.57	00.00
SVM	**85.17**	70.92	75.35	**80.31**	00.00	73.07	00.00
RNA	81.09	76.62	**83.22**	73.76	00.00	72.39	00.00

TABLE B1 – Performances par algorithme de classification

Face à ces résultats, nous choisissons de présenter dans le corps de l'article les résultats obtenus avec une régression logistique, nécessaire puisque ses coefficients sont interprétables, et avec une SVM, dont les résultats sont les meilleurs pour le plus grand nombre de projections.

B.2 Taille des vecteurs

Le dernier paramètre à optimiser est celui de la taille des vecteurs. Ils correspondent en particulier pour nos projections par topic models au paramètre K. Le problème de l'optimisation de ce paramètre est qu'il est très coûteux d'un point de vu calculatoire, et il n'est pas réalisable pour les projections des topic models les plus complexes dont une seule estimation nécessite plusieurs jours. Ainsi, nous présentons les résultats d'optimisation pour les approches par sac de mot, doc2vec et par agrégation de projections de mots. Nous les présentons également pour LDA, et nous estimerons que le paramètre optimal pour cette projection sera également optimal pour toutes les projections par topic models et pour LSA. Nous présentons les résultats de cette optimisation pour les deux classifieurs retenus, la régression logistique (Fig. B1) et la SVM (Fig. B1).

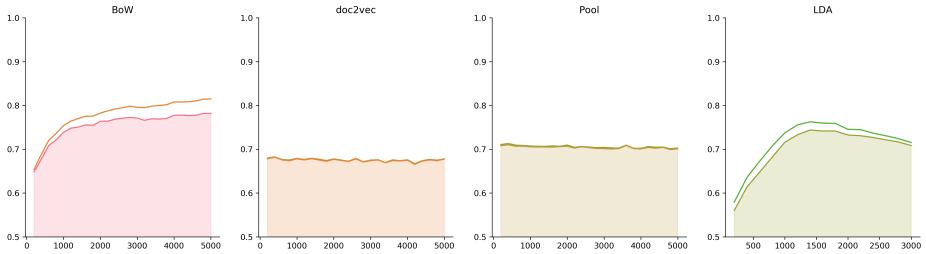

FIGURE B1 – Performances d'une régression linéaire en fonction de la taille des vecteurs

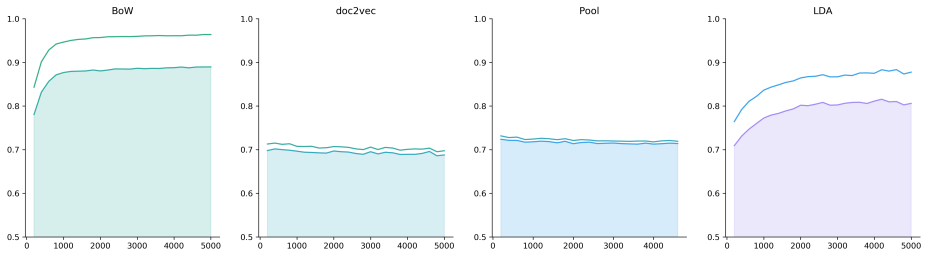

FIGURE B2 – Performances d'une SVM en fonction de la taille des vecteurs

Résumés dans la Table B2, les résultats appellent les conclusions suivantes : si l'approche par sac de mots bénéficie d'un vecteur aussi grand que possible, le gain d'information est très faible passé $K = 1000$. Pour les modélisations par doc2vec ou word2vec au contraire, la justesse est décroissante passé $K = 400$. Enfin le LDA présente une courbe singulière pour la régression logistique, puisque la justesse atteint un pic pour $K = 1400$. Pour la SVM, elle bénéficie en revanche d'avoir un vecteur aussi large que possible. La dernière ligne de la Table B2 résume notre prise de décision pour la taille des vecteurs utilisée dans le corps de ce travail. Le choix pour les topic models est avant tout dicté par une contrainte de temps de calcul.

Modélisation	**BoW**	**Doc2Vec**	**Pool**	**LDA**	**Autres TM**
Taille optimale (logit)	4800	400	400	1400	
Performance (logit)	78.17	68.19	71.02	74.39	
Taille optimale (SVM)	5000	400	200	4200	
Performance (SVM)	88.96	70.16	72.38	81.55	
Taille retenue	5000	400	400	2000	1000

TABLE B2 – Taille de vecteur optimale par modélisation