# De quoi parlent les patients dans les forums de santé : classification non-supervisée par LDA

Mike Donald Tapi Nzali, Sandra Bringay, Christian Lavergne, Caroline Mollevi

# De quoi parlent les patients dans les forums de santé : classification non-supervisée par LDA

Mike Donald Tapi-Nzali [1,3,4] & Sandra Bringay [2,4] & Christian Lavergne [2,3] & Caroline Mollevi [5]

[1] *Université de Montpellier, France*
[2] *Université Paul Valéry, Montpellier 3, France*
[3] *Institut Montpelliérain Alexander Grothendieck, France*
[4] *Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, France*
[5] *Institut du Cancer de Montpellier, Montpellier, France*

*Mike-Donald.Tapi-Nzali@umontpellier.fr / Sandra.Bringay@univ-montp3.fr / Christian.Lavergne@univ-montp3.fr / Caroline.Mollevi@icm.unicancer.fr*

**Résumé.** De nos jours, les médias sociaux sont de plus en plus utilisés par les patients et les professionnels de santé. Il s'agit d'une ressource textuelle riche, générée par les très nombreux échanges entre patients et, dans certains cas, professionnels de santé. Dans cet article, nous utilisons le modèle d'apprentissage non supervisé connu sous le nom de LDA (Allocation de Dirichlet Latente) afin de détecter les différents thèmes abordés sur les forums de santé et les réseaux sociaux par les patients. Notre objectif est de repérer les nouveaux thèmes directement issus des préoccupations des patientes atteintes de cancer du sein et de les comparer aux thèmes existant dans les auto-questionnaires proposés dans les essais cliniques en oncologie.

**Mots-clés.** Cancer du sein, fouille de texte, médias sociaux, apprentissage non-supervisé.

**Abstract.** Nowadays, social media is increasingly used by patients and health professionals. This is a rich text resource, generated by many exchanges between patients and in some cases health professionals. In this paper, we use unsupervised learning model known as LDA (Latent Dirichlet Allocation) to detect the different topics on health forums and social networks discussed by patients. Our main objective is to detect the different themes by patients during their accounts in social media and compare them with predefined themes existing in the questionnaires used in clinical trials. We also show pretreatments to be performed on these data for such tasks.

**Keywords.** Breast cancer, text mining, socials medias, unsupervised learning.

# 1 Introduction

Les médias sociaux comme Facebook, Twitter ou les forums dédiés à la santé ont évolué pour devenir des outils participatifs facilement accessibles pour l'échange de connaissances

et d'expériences. Ils constituent une collection structurée de données textuelles très riche. Ces médias sociaux permettent aux patients de maintenir leur anonymat tout en discutant librement avec d'autres patients ou des professionnels de santé. Alors que les communications avec les médecins et le personnel médical dans les hôpitaux peuvent principalement tourner autour de questions techniques, sur la maladie et le traitement, les médias sociaux, quand à eux, offrent l'accès à des échanges, le partage d'expériences et favorisent le soutien mutuel entre patients confrontés à une situation similaire. Ainsi, les médias sociaux peuvent être considérés comme une ressource précieuse pour l'étude de la qualité de vie (QdV) liée à la santé. Certaines études [7] montrent que l'environnement anonyme des médias sociaux facilite l'expression d'opinions comme le doute et la peur.

Dans ce travail, nous proposons une approche pour structurer et évaluer les informations cliniquement pertinentes à partir des données textuelles extraites des médias sociaux tout en mettant l'accent sur la qualité de vie des patients atteints d'un cancer du sein. Si des progrès constants de la médecine conduisent à de nouveaux traitements améliorant les chances de prolonger la quantité de vie, il est tout autant nécessaire de s'intéresser à la qualité de ce gain de survie. Dans ce contexte, la qualité de vie peut donc être considérée comme un critère d'évaluation clinique d'intérêt. En particulier, dans les situations gériatriques et palliatives afin de s'assurer en priorité du confort et du bien-être des patients.

Le but de notre démarche est multiple : 1) Accéder et exploiter des données supplémentaires à celles générées par les essais cliniques ; 2) Découvrir des opinions et des sentiments sur-exprimés par les patients en dehors de l'environnement clinique et la relation médecin-patient ; 3) Offrir un point de vue alternatifs à celui des essais cliniques où la qualité de vie est mesurée sur la base de questionnaires pré-formulés; 4) Aider les experts cliniques à comprendre les besoins et les préoccupations des patients tout le long de leur prise en charge.

# 2 Méthodologie utilisée

Dans ce travail, nous utilisons la LDA [3] pour découvrir les thèmes évoqués dans les forums de santé et illustrer comment ces thèmes peuvent être utilisés dans l'étude de la qualité de vie. Le modèle d'apprentissage non supervisé LDA est généralement utilisé pour la détection des thèmes dans un corpus de textes. C'est un modèle probabiliste avec une définition hiérarchique de ses composants. Il est plus constructif, en ce sens qu'on pourrait générer de nouveaux documents à partir d'un modèle donné. Il est relativement simple et robuste, avec une représentation en "sac-de-mots", il ne prend en considération ni l'ordre d'apparition des termes, ni la structure du texte lors des traitements.

Dans un corpus textuel, nous avons un vocabulaire ayant une taille $n_W$, il s'agit d'une collection de termes prétraités qui occurrent dans le corpus. Nous avons également un nombre $n_D$ de documents et un nombre de thèmes $k$.

Le modèle de génération du corpus de LDA est le suivant [8] :

- Pour chaque thème $t \in \{1 \dots k\}$, tirer aléatoirement les paramètres des lois discrètes probabilisant les occurrences des mots du vocabulaire selon une loi de Dirichlet $\beta_t = (\beta_{t1}, \dots, \beta_{tn_w})$
- Pour chaque document $d \in \{1, \dots, n_D\}$ :
  - Tirer aléatoirement la distribution des thèmes dans $d$ selon $\alpha_d = (\alpha_{d1}, \dots, \alpha_{dn_T}) \sim Dir(\lambda_\alpha, \dots, \lambda_\alpha)$. Chaque $\alpha_{dt}$ indique donc la proportion des occurrences du document $d$ qui sont associées au thème $t$.
  - Pour chaque position $i$ dans $d$, $i \in \{1, \dots l_d\}$
    - Tirer aléatoirement un thème selon une loi discrète $T_{di} \sim Disc(\alpha_d)$.
    - Tirer aléatoirement un mot conditionnellement au thème selon : $W_{di} \sim Disc(\beta_{T_{di}})$

L'information principale que nous pouvons apprendre en ajustant un tel modèle sur un corpus de données textuelles est la détection des thèmes et la distribution des thèmes sur les documents contenus dans le corpus. Le nombre élevé de paramètres inconnus dans ce modèle rend l'inférence difficile, mais les techniques de Bayes comme l'échantillonnage de Gibbs se sont avérées fiables. Basées sur la distribution a priori des poids des termes dans les thèmes et des thèmes dans les documents, ces techniques d'inférence confrontent le modèle à des données et estiment les distributions a posteriori. Plus important encore, la structure des thèmes la plus probable et les probabilités d'occurrence des thèmes dans chaque document sont proposées.

Pour le modèle LDA, il faut spécifier deux distributions a priori de type Dirichlet, chacune caractérisée par un paramètre de concentration $\nu > 0$; un pour les $k$ vecteurs de poids $\omega_{ti}$ des termes dans les thèmes avec paramètre $\nu_{theme}$ et un autre pour les $D$ vecteurs de poids $\omega_{dt}$ de thèmes dans les documents avec des paramètres $\nu_{doc}$.

## 2.1 Choix des paramètres $\alpha$ et $\delta$

Outre le paramètre $k$, deux autres variables souvent notées $\alpha$ et $\delta$ ont une influence sur la répartition des probabilités pour chaque thème des messages. Ce sont des paramètres de concentration pour les distributions a priori des thèmes sur un message ($\alpha$) et d'un mot sur un thème ($\delta$).

Dans ce qui suit, nous expliquons notre choix de $\alpha$ sur la base de son influence sur la distribution de probabilités des thèmes pour les messages et la distribution des termes pour les thèmes. Lorsque $\alpha = 1$, la distribution a priori pour le vecteur de sujet probabilités correspond à une distribution uniforme dans le simplexe avec $k$ sommets. Lorsque $\alpha$ augmente, cette distribution se concentre de plus en plus fortement vers le centre du simplexe, de sorte que la plupart des probabilités sont près de $1/k$. Lorsque $\alpha$ diminue, elle

se concentre de plus en plus fortement vers les sommets, ce qui conduit à des probabilités éloignées de $1/k$. Avec $\alpha$ fixé, les probabilités se concentrent de plus en plus autour de $1/k$ lorsque $k$ augmente. Dans [6], les valeurs $\alpha = \alpha_0/k$ avec la constante $\alpha_0 = 50$ sont préconisées, où la division par $k$ maintient constante une certaine mesure de la complexité sur le modèle. Une analyse exploratoire a montré que $\alpha_0 = 50$ conduit à des vecteurs de probabilité "très plats" dans notre cas, ce qui rend difficile l'attribution d'un petit nombre de thèmes à l'indexation pour chaque message. D'autre part, une valeur très petite de $\alpha_0$ conduit à des thèmes difficilement interprétables. Ceci étant dû à une distribution de probabilités plate des termes dans les thèmes et une apparition des termes similaires dans plusieurs thèmes. Certaines études [5] ont montré que que le choix automatique des paramètres à travers un critère de sélection d'un modèle rend l'interprétation difficile des thèmes. Nous allons donc utiliser plusieurs méthodes pour choisir le meilleur $k$ sur nos données.

## 2.2 Choix du paramètre $k$

Nos expérimentations ont été faites sur les corpus décrits dans la section 3.1. Un accent a été particulièrement mis sur le choix du nombre de thèmes $k$. Pour cela, nous cherchons à trouver le modèle pour lequel on obtient le paramètre $k$ optimal. Étant donné qu'il est impossible de comparer les modèles LDA par des techniques classiques telles que l'AIC ou le BIC, car difficile de calculer la vraisemblance du modèle, nous utilisons des méthodes permettant de trouver le modèle ayant le $k$ optimal. Les trois méthodes que nous avons utilisées sont les suivantes pour trouver le nombre de thèmes $k$ optimal :

- **Méthode basée sur la moyenne harmonique [6]**. Le but ici est de calculer la probabilité $p(w|k)$. Vu qu'il est très difficile, voire impossible d'avoir cette valeur, on peut approximer $p(w|k)$ en prenant la moyenne harmonique d'un ensemble de valeurs de $p(w|z,k)$. Le modèle que nous retiendrons en faisant varier k sera celui qui aura la valeur la plus élevée.

- **Méthode basée sur la densité [4] et méthode basée sur la divergence de Kullback-Leibler [1]**. Ces deux méthodes suivent le même principe, mais diffèrent juste sur certains points. Le principe revient à calculer des similarités (ou des distances) entre toutes les paires de thèmes pour différents modèles obtenus en faisant varier le nombre de thèmes. LDA peut être vu comme une méthode de "Factorisation par matrices non négatives" $M$ pouvant être décomposée en une matrice $M1$ de thèmes-mots et une matrice $M2$ de documents-thèmes. La mesure de divergence utilisée est calculée par la formule 1 pour chaque valeur attribuée à $k$. Le $k$ optimal est celui qui a la plus faible divergence.

$$Div(M1, M2) = KL(C_{M1}||C_{M2}) + kl(C_{M2}||C_{M1)} \quad (1)$$

où $C_{M1}$ est la décomposition en valeurs singulières de la matrice M1 et $C_{M2}$ est la distribution obtenue par normalisation du vecteur $L * M2$. $L$ est le vecteur de taille $1 * D$ contenant la longueur de chaque document dans le corpus.

# 3 Application aux données des forums

Les textes issus des forums de santé tels que `CancerDuSein.org` et `lesimpatientes.com` puis des réseaux sociaux tels que les " les groupes Facebook" ne donnent qu'une représentation relativement frustre des thèmes abordés par les utilisateurs. De plus, le nombre de "posts" sont fortement déséquilibrés dans toutes les catégories. Ici, nous cherchons une caractérisation plus fine des thèmes avec des informations sur l'attribution des termes aux thèmes et de l'attribution des thèmes aux "posts".

## 3.1 Données utilisées et prétraitements

Dans ce travail, nous utilisons différents médias sociaux (*Facebook, cancerdusein.org, lesimpatientes.com*). Nous avons 96 925 messages extraits du groupe Facebook, 130 000 messages extraits du forum "`lesimpatientes.com`" et 16 900 messages du forum "`CancerDuSein.org`".

Les messages des forums ont plusieurs particularités linguistiques qui peuvent influer sur l'exécution de plusieurs tâches [2]. Les utilisateurs de médias de santé sociale emploient de l'argot, un vocabulaire informel, font beaucoup de fautes d'orthographe et utilisent des abréviations. Pour cette raison, nous avons prétraité les données en supprimant les tags utilisateurs, hyperliens, adresses mails, pseudonymes et les expressions fréquemment utilisées dans les médias sociaux tels que "lol, mdr, ..."

## 3.2 Expérimentations

Le modèle a été appliqué sur les données présentées précédemment, plusieurs thèmes ont été sélectionnés et interprétés par un oncologue spécialiste du cancer du sein.

La liste des thèmes déterminés n'est pas exhaustive puisqu'elle est associée au choix du paramètre $k$ (ici $k = 30$). Les principales préoccupations des patientes atteintes d'un cancer du sein qui émergent des données textuelles étudiées ont été regroupées selon des thématiques qui concernent principalement cinq catégories : l'annonce de la maladie (*dépistage, diagnostic et annonce du cancer, recherche de renseignements médicaux*), les traitements et leurs effets indésirables (*chirurgie, réaction à l'annonce d'une chirurgie du sein, reconstruction mammaire, protocole de chimiothérapie/radiothérapie, alopécie (dû à la chimiothérapie), effet secondaire de la chimiothérapie, hormonothérapie et ses effets secondaires*), les relations avec le personnel médical (*interaction avec les infirmières, confiance dans le personnel médical*), le ressenti des patientes par rapport à leur maladie (*anxiété, état d'esprit, soins et image du corps pendant le cancer, soutien de l'entourage*),

et enfin, tout ce qui concerne des aspects plus "pratiques" de leur vie (*aspect financier, vie quotidienne, vie professionnelle au cours du cancer*).

# 4 Discussion

Nous avons remarqué de bonnes correspondances entre les thèmes détectés et les auto-questionnaires pour les patients atteints du cancer du sein, ce qui justifie la construction rationnelle de ces questionnaires. Par ailleurs, nous pouvons confirmer que les médias sociaux peuvent être une importante source d'informations pour les oncologues. De plus, le nombre de thèmes est très important dans la mesure où la taille des corpus à traiter diffère à chaque fois. Il est donc important de trouver une méthode semi-automatique qui nous permettra de faciliter le choix du paramètre $k$, car même si manuellement on aboutit parfois à de bons résultats, cela est très coûteux en temps.

# Bibliographie


[1] R. Arun, V. Suresh, C. V. Madhavan, and M. N. Murthy. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Advances in Knowledge Discovery and Data Mining*, pages 391–402. Springer, 2010.

[2] A. Balahur. Sentiment analysis in social media texts. In *4th workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 120–128. Citeseer, 2013.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[4] J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang. A density-based method for adaptive lda model selection. *Neurocomputing*, 72(7):1775–1781, 2009.

[5] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

[6] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[7] J. T. Hancock, C. Toma, and N. Ellison. The truth about lying in online dating profiles. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 449–452. ACM, 2007.

[8] L. Rigouste, O. Cappé, and F. Yvon. Quelques observations sur le modele lda. *Actes des IXe JADT*, pages 819–830, 2006.