Topic modeling neuronal non-paramétrique pour l'extraction d'insight client : application à l'industrie du pneumatique

Miguel Palencia-Olivar*,**, Stéphane Bonnevay*,**, Alexandre Aussem***, Bruno Canitia*

* Lizeo IT, 42 quai Rambaud, 69002 Lyon, France
miguel.palencia-olivar, stephane.bonnevay, bruno.canitia@lizeo-group.com
** Lab. ERIC, Université de Lyon, 5 Av. Pierre Mendès France, 69500 Bron, France
*** LIRIS, Université de Lyon, 25 Av. Pierre de Coubertin, 69100 Villeurbanne, France
alexandre.aussem@univ-lyon1.fr

Résumé. À l'ère des médias sociaux, les clients sont devenus des faiseurs d'opinion : toute personne intéressée par un produit peut rechercher des avis sur les plateformes d'internet. Le web-scraping reste souvent la seule voie d'accès, et malgré le recours à l'ETL, l'hétérogénéité des données rend la tâche d'extraction d'insight ardue et induit le besoin d'outils ad-hoc. Pour contourner ce problème, nous appliquons l'Embedded Dirichlet Process et l'Embedded Hierarchical Dirichlet Process dans un cadre industriel autour du cas du pneumatique. Ces topic models non-paramétriques apprennent les thèmes et leur nombre, des plongements de thèmes et des plongements de mots, de telle sorte à désambiguïser les mots et à offrir davantage de niveaux analytiques. Ils peuvent également servir à affiner des processus ETL. L'EDP et l'EHDP atteignent des niveaux de vraisemblance similaires voire plus élevés que ceux des techniques de l'état de l'art testées, sans ré-exécutions pour trouver le nombre de thèmes.

1 Introduction

Les topic models constituent un ensemble de techniques de référence en matière de text mining non supervisé. Leurs applications sont nombreuses, allant du résumé de rapports cliniques à l'extraction de tendances dans la littérature scientifique (Boyd-Graber et al. (2017)). Depuis l'introduction de l'Allocation de Dirichlet Latente (Blei et al. (2003)), le domaine a connu des développements dans de nombreuses directions. Les dernières en date consistent à mêler réseaux de neurones profonds et programmation probabiliste afin de tirer parti des propriétés et avancées des deux domaines (Miao et al. (2016); Srivastava et Sutton (2017)), en particulier en termes de plongements (Dieng et al. (2020)). L'inclusion de plongements semble prometteuse, puisqu'elle revient à ajouter des éléments contextuels. Ceux-ci semblent rendre plus aisée la capture de thèmes tout en permettant de visualiser les liens entre mots et entre thématiques. Ces capacités sont particulièrement intéressantes dans le cadre de l'analyse de réseaux sociaux. Toutefois, les applications de topic models neuronaux sur de tels jeux de données portent souvent sur des corpus provenant de sources uniques accessibles par API ; ce

faisant, il est aisé de réaliser des prétraitements permettant de standardiser un jeu de données *ante*-analyse. Ces situations ne sont pas représentatives des analyses portant sur des données issues d'activités de web scraping, qui portent sur plusieurs sources et nécessitent souvent plusieurs passes de nettoyage afin que les analyses en aval ne soient pas bruitées par des langages de balisage, par exemple. Il s'agit ici de notre cas d'espèce. Nous proposons ici d'appliquer l'Embedded Dirichet Process (EDP) et l'Embedded Hierarchical Dirichlet Process (EHDP) (Palencia-Olivar et al. (2021)) à un corpus issu d'une activité industrielle en rapport avec des avis clients portant sur le pneumatique. Le présent article est un résumé de l'article publié dans la conférence IJCNN 2022 (Palencia-Olivar et al. (2022)). Ces modèles initalement testés sur des jeux de données benchmark ont eu des performances satisfaisantes dans notre cadre et surpassent d'autres modèles de l'état de l'art. De plus, et parce qu'ils sont non-paramétriques, ceux-ci permettent de déterminer automatiquement le nombre de thèmes au sein d'un corpus sans coûteuse ré-exécution. Cet article est organisé comme suit : nous présentons l'EDP et l'EHDP en Section 2, puis présentons jeu de données, expérimentations et discussions en Section 3. Enfin, nous présentons nos conclusions et de possibles axes de travail en Section 4.

2 Les modèles

L'EDP et l'EHDP sont deux topic models neuronaux non-paramétriques par rapport au nombre de thèmes. Ils sont basés sur des autoencodeurs variationnels (VAE) et font usage de Processus de Dirichlet (DP) avec une construction en bris de bâton (stick-breaking). Les modèles présentent la particularité d'extraire les thèmes, leur nombre, ainsi que des plongements de thèmes et des plongements de mots sans supervision aucune. Nous en présentons le fonctionnement ci-après.

2.1 L'Embedded Dirichlet Process

Soit $\{\mathbf{w}_1, \ldots, \mathbf{w}_d\}$ un corpus de D documents, où $\mathbf{w}_d$ est un ensemble de N_d mots. Chaque document admet une représentation sac de mots $\mathbf{w}_d$. Le processus génératif de l'EDP pour un document est le suivant :

1. Échantillonner $G^{(d)}\left(\theta; \pi^{(d)}, \Theta\right) = \sum_{k=1}^{\infty} \pi_k^{(d)} \delta_{\theta_k}(\boldsymbol{\theta})$, avec $\pi^{(d)} \sim Beta\,(1, \beta)$
2. Pour chaque mot w_n de $\mathbf{w}_d$:
   (a) Échantillonner un thème $\hat{\theta}_n \sim G^{(d)}\left(\theta; \pi^{(d)}, \Theta\right)$
   (b) Échantillonner un mot $w_n \sim Multinomiale\left(\rho^T \phi\right)$

L'EDP décompose le paramètre de la multinomiale afférente aux mots en un produit entre une matrice (transposée) de plongements de mots ρ et une autre de plongements de thèmes ϕ. Cette décomposition forme un modèle log-linéaire ; ce faisant, nous pouvons visualiser les plongements dans le même espace - dans un esprit similaire à celui de l'ACP - et donc comparer les positions relatives des thèmes et des mots. La distribution jointe du modèle est la suivante :

$$Pr\left(\mathbf{w}_{1:N}, \pi, \hat{\theta}_{1:N} \mid \beta, \Theta, \xi\right) = Pr\,(\pi \mid \beta) \times \Pi_{i=1}^{N} Pr\left(w_i \mid \hat{\theta}_{\mathbf{i}}, \xi\right) Pr\left(\hat{\theta}_{\mathbf{i}} \mid \pi, \Theta\right) \quad (1)$$

avec $Pr(\pi \mid \beta) = Beta(1, \beta)$, $Pr(\theta \mid \pi, \Theta) = G(\theta; \pi, \Theta)$, $Pr(w \mid \theta, \xi) = \sigma(\theta\xi)$, avec $\xi = \sigma\left(\rho^T \phi\right)$, $\sigma(\cdot)$ étant la fonction softmax, G étant l'*a priori* sur les thèmes (ou processus de Dirichlet), π étant les bris de bâton (stick-breaks), θ étant les proportions de thèmes. Enfin, nous cherchons à optimiser l'objectif suivant :

$$\begin{aligned}\mathcal{L}(\mathbf{w}_{1:N} \mid \Theta, \psi, \xi) = \mathbb{E}_{q_\psi(\nu|\mathbf{w}_{1:N})}[\log Pr(\mathbf{w}_{1:N} \mid \pi, \Theta, \xi)] \\ -KL(q_\psi(\nu \mid \mathbf{w}_{1:N}) \parallel Pr(\nu \mid \beta))\end{aligned} \tag{2}$$

où $q_\psi(\cdot)$ est une famille de distributions variationnelles, ψ l'ensemble des paramètres d'un réseau de neurones génératif, v représente les poids de la construction stick-breaking du DP, et a et b sont les paramètres variationnels de la distribution Beta appris par encodage. Nous entraînons ce modèle par *inférence variationnelle amortie* et utilisons l'optimiseur Adam pour l'ensemble des paramètres.

2.2 L'Embedded Hierarchical Dirichlet Process

L'*a priori* de l'EDP implique la distribution Beta, dont le paramètre β est de poids : plus il est important, plus nous obtiendrons de thèmes et inversement. Afin de contrôler la croissance de leur nombre, il est possible de l'apprendre à partir des données. Pour ce faire, nous utilisons la distribution Gamma en tant qu'*hyper-a priori*. Le processus génératif pour un document devient le suivant :

1. Échantillonner $\beta \sim Gamma(\delta_1, \delta_2)$
2. Échantillonner $G^{(d)}\left(\theta; \pi^{(d)}, \Theta\right) = \sum_{k=1}^{\infty} \pi_k^{(d)} \delta_{\theta_k}(\boldsymbol{\theta})$, avec $\pi^{(d)} \sim Beta(1, \beta)$
3. Pour chaque mot w_n de $\mathbf{w}_d$:
   (a) Échantillonner un thème $\hat{\theta}_n \sim G^{(d)}\left(\theta; \pi^{(d)}, \Theta\right)$
   (b) Échantillonner un mot $w_n \sim Multinomiale\left(\rho^T \phi\right)$

L'objectif d'optimisation devient alors :

$$\begin{aligned}\mathcal{L}(\mathbf{w}_{1:N} \mid \Theta, \psi, \xi) = \mathbb{E}_{q_\psi(\nu|\mathbf{w}_{1:N})}[\log Pr(\mathbf{w}_{1:N} \mid \pi, \Theta, \xi)] \\ +\mathbb{E}_{q_\psi(\nu|\mathbf{w}_{1:N})q(\beta|\gamma_1,\gamma_2)}[\log Pr(\nu \mid \beta)] \\ -\mathbb{E}_{q_\psi(\nu|\mathbf{w}_{1:N})}[\log q_\psi(\nu \mid \mathbf{w}_{1:N})] \\ -KL(q(\beta \mid \gamma_1, \gamma_2) \parallel Pr(\beta \mid \delta_1, \delta_2))\end{aligned} \tag{3}$$

où γ_1 et γ_2 sont des paramètres variationnels, tandis que δ_1 et δ_2 sont des paramètres *a priori*. À l'instar de l'EDP, ce modèle est entraîné par inférence variationnelle amortie et l'optimiseur Adam est utilisé pour l'ensemble des paramètres.

3 Étude empirique

Notre jeu de données est issu d'une activité industrielle consistant à collecter des avis clients sur des pneumatiques. Nous en proposons une description avant de présenter nos expérimentations et les conclusions en découlant.

3.1 Jeu de données et pré-traitement

Nous travaillons avec un sous-ensemble d'une base de données contenant des documents recueillis sur 1073 sites, dont 442 sont en anglais. Nos modèles n'étant pas prévus pour traiter les corpus multilingues, nous écartons les documents qui ne sont pas en anglais. Ce filtrage est principalement basé sur la langue du site de provenance du document ; il n'y a donc aucune garantie qu'il n'existe pas de document dans une langue autre que l'anglais. Puisque nous ne nous intéressons qu'à l'anglais, toute autre langue est ici considérée comme élément de bruitage. De même, notre base de données est partiellement annotée selon des éclairages d'experts, sous forme de chaînes de caractères spéciales. Ces annotations ne sont pas notre centre d'intérêt [1], mais n'ont pas pu être retirées car ce retrait aurait impliqué de modifier et de rejouer tout le processus d'ETL [2] interne. Nous considérons donc ces chaînes comme étant des éléments de bruitage. Malgré l'existence d'ETL en amont, il n'existe aucune garantie à propos de l'absence d'autres types de bruit dans les données. Enfin, il n'était possible ni de retirer la ponctuation et les chiffres, ni de passer le texte en minuscule. Concernant la ponctuation, les références produit (ex : 205/50ZR15) et dimensions de pneu (ex : 7.2/32") en incluent. En revanche, celles-ci sont présentes sans notation standard, de telle sorte qu'il n'est pas possible de créer d'expression régulière permettant de toutes les préserver. Concernant cette fois le passage en minuscule, il s'agissait de pouvoir distinguer les éléments. Ainsi, de la même manière que le latex se distingue de LaTeX, un climat continental se distingue de la marque Continental. Nos sources de données sont diverses ; cela vaut également pour nos documents. Ces derniers comprennent des avis client provenant de sites de e-commerce, de revues d'articles, de posts de blog, de fils de discussion ainsi que de tweets. Pour notre pré-traitement, les éléments issus du comptage des mots sont les seuls objectivement observables sans faire appel à un modèle. Nous nous sommes focalisés sur la longueur des documents pour trois raisons : 1/ il y a un lien direct entre longueur d'un document et sparsité d'un document ; 2/ un long document devrait être moins sparse [3] qu'un court, mais avoir peu de mots qui reviennent souvent du fait de la loi de Zipf [4] ; 3/ nous voulons représenter tous les documents à travers tous les thèmes et non seulement les courts ou les longs. Afin d'éliminer les documents de longueur marginale, nous avons mené une étude sur un échantillon de 96910 documents de notre base. Notre unité statistique est le mot, au sens d'une chaîne de caractères entourée d'espaces. Les collocations sont ainsi considérées de manière identique à tous les autres mots. Il a résulté de notre étude sur la longueur des documents que celle-ci suit une loi de Poisson de paramètre $\lambda = 107.9$ ($p < 0.05$ selon une adaptation aux données discrètes du test de Kolmogorov-Smirnov, cf Conover (1972)), ce qui représente 98.75% des données que nous traitons. Sur la base de notre inférence, nous avons retiré tous les documents de longueur supérieure à 450 mots. De même, nous ne pouvions pas retenir la totalité de l'échantillon d'étude pour les expérimentations du fait de capacités de calcul limitées. Toujours grâce à notre étude préalable, nous avons segmenté notre base en 22 classes d'amplitude 20. Pour réaliser cette segmentation, nous avons fait un histogramme dont le nombre de classes (binning) a été calculé grâce à la formule de Doane du fait du caractère non-Gaussien des données de comptage. La finalité de la manoeuvre

1. Les experts se focalisent sur des caractéristiques techniques, tandis que les consommateurs se focalisent sur leur expérience du produit.
2. Ce dernier inclut des phases de parsing et remplace des chaînes de caractères selon une ontologie interne, entre autres. Il n'y a en revanche aucune notion de détection d'entité nommée.
3. La gestion de la sparsité est un point faible notoire des VAE.
4. Cette hypothèse a été testée statistiquement dans l'article originel.

était de pouvoir créer un sous-échantillon de 40000 documents par échantillonnage stratifié ; pour cela, il nous fallait respecter la densité originelle tout en maintenant un nombre raisonnable de classes. Ce sous-échantillon de 40000 documents a par la suite servi à former un vocabulaire de 26731 mots et à former nos jeux d'entraînement (80%), de validation et de test (10% chacun).

3.2 Expérimentations et discussion

Nous avons testé plusieurs modèles selon deux grilles comparatives : modèle neuronal contre modèle non-neuronal et modèle paramétrique contre modèle non-paramétrique. Nous utilisons l'inférence variationnelle et des tailles de batch de 1000 documents pour l'ensemble des modèles testés. Les modèles concurrents sont l'iTM-VAE-Prod et l'iTM-VAE-G tels que décrits dans Ning et al. (2020), une version utilisant une distribution variationnelle Beta de l'iTM-VAE-Prod que l'on nommera SB-VAE implicit, l'ETM (Dieng et al. (2020)), la ProdLDA (Srivastava et Sutton (2017)), le HDP (Wang et al. (2011)) et la LDA (Blei et al. (2003)). Pour les modèles neuronaux, les encodeurs sont des perceptrons multicouche comprenant 2 couches de 100 neurones chacun, tandis qu'Adam admet un pas d'aprentissage de 0.002. Les modèles admettant des plongements (l'ETM, l'EDP et l'EHDP) utilisent des plongements de dimension 300. Nous avons réalisé des tests en initialisant les plongements de mots au hasard et avec des Skip-grams (Mikolov et al. (2013)) ; nous suffixons les noms des modèles avec -R dans le premier cas et -T dans le second. Tous les paramètres ont été fixés par validation croisée, jusqu'à convergence et dans une limite de 150 itérations dans le cas des modèles neuronaux. Suivant l'état de l'art du thème modeling, notre critère de référence pour l'ajustement statistique est la log-vraisemblance (LL). Concernant l'ajustement qualitatif, nous définissons la qualité (TQ) comme le produit de deux autres indicateurs : la diversité des thèmes (TD) et leur cohérence (TC), c'est-à-dire $TQ = TC \times TQ$. La TD est la proportion de mots uniques parmi les 10 premiers mots de chaque thème, tandis que la TC se définit comme suit :

$$TC = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{45}\sum_{i=1}^{10}\sum_{j=i+1}^{10} f\left(w_i^{(k)}, w_j^{(k)}\right) \text{avec} f\left(w_i^{(k)}, w_j^{(k)}\right) = \frac{\log\frac{Pr(w_i,w_j)}{Pr(w_i)Pr(w_j)}}{-\log Pr\left(w_i, w_j\right)}. \quad (4)$$

Cette formule correspond à l'information mutuelle normalisée (NPMI) et rend compte de la co-occurrence des termes. La TC prend ses valeurs entre -1 et 1 ; pour deux termes, -1 est l'absence systématique de co-occurrence, 0 une indépendance et 1 une co-occurrence systématique.

Interprétation des indicateurs Dans l'ensemble, les résultats obtenus en phase de test (Tab. 1) montrent que les modèles neuronaux non-paramétriques obtiennent les meilleurs résultats, tant en termes de vraisemblance que de qualité. L'iTM-VAE-Prod est une exception. Ce modèle utilise une distribution-substitut de la distribution Beta dite de Kumaraswamy. Dans le même contexte, son équivalent SB-VAE implicit qui utilise pour sa part la distribution Beta a obtenu parmi les meilleurs résultats, tandis que l'ETM et la ProdLDA ont certes pu être entraînés, mais n'arrivaient pas à généraliser. Les modèles ayant bénéficié de pré-entraînement (transfer learning) ne semblent pas avoir fait mieux que s'ils n'avaient pas été initialisés au hasard. L'iTM-VAE-G généralise aussi bien que ses concurrents neuronaux non-paramétriques,

Type	Modèle	# thèmes	LL	TC	TD	TQ
Neuronal non paramétrique	*EHDP-R*	9	-664	-0.05	0.97	-0.05
	EHDP-T	10	-655	-0.05	**1.0**	-0.05
	EDP-R	9	-625	0.05	**1.0**	0.05
	EDP-T	9	**-622**	-0.02	**1.0**	-0.02
	iTM-VAE	NaN	NaN	NaN	NaN	NaN
	SB-VAE implicit	12	-631	0.23	0.73	**0.17**
	iTM-VAE-G	10	-680	0.17	0.11	0.02
Non neuronal non paramétrique	*HDP*	50	-2.7×10^6	0.21	0.02	0
Neuronal paramétrique	*ProdLDA*	10	NaN	-0.65	0.8	-0.52
	ETM-R	10	NaN	-0.63	1.0	-0.63
	ETM-T	10	NaN	-0.65	0.93	-0.60
Non neuronal paramétrique	*LDA*	10	-2.06×10^5	0.06	0.48	0.03

TAB. 1 – *Résultats quantitatifs*

mais est de moindre qualité, tant en termes d'indicateurs que d'interprétabilité (cf les résultats qualitatifs). Enfin, les modèles ne faisant pas appel à des réseaux de neurones sont bien moins efficaces en généralisation que leurs pendants neuronaux et ont des indicateurs de qualité similaires. Pour autant, ils ne sont pas aussi interprétables.

Interprétation des thèmes et plongements Dans l'ensemble, nous sommes mitigés quant à la qualité des résultats obtenus par les différents modèles, en particulier lorsque nous confrontons les thèmes obtenus aux indicateurs. En effet, hormis les deux initialisations de l'EHDP, seules l'EDP-R, l'EDP-T et le SB-VAE implicit ont obtenu des résultats satisfaisants. Le SB-VAE implicit a extrait 2 à 3 thèmes de plus que les autres modèles, ces thèmes additionnels étant peu interprétables car contenant la même suite de mots sans cohérence apparente. Ceci explique une TD plus faible. De plus, le SB-VAE implicit n'inclut pas de plongements et ne permet donc pas d'aller plus loin. Les autres thèmes sont qualitativement proches de ceux de l'EHDP-R, qui sont eux-mêmes proches de ceux de l'EHDP-T dont nous produisons ici un extrait, les mots étant ordonnés par ordre d'importance pour chaque thème (Tab. 2). Notons que les thèmes ont été nommés sur la base de leur contenu sans aucune automatisation à cet égard mais sous contrôle d'un expert métier. Concernant l'évaluation qualitative des performances gagnées par pré-entraînement, nous avons extrait les plongements de mots produits par les modèles -R et les modèles -T puis examiné le voisinage des principaux mots des thèmes interprétables. Ces voisinages étant très proches, nous déduisons que nos modèles sont capables d'obtenir des plongements de mots à la qualité au moins équivalente à celle des Skip-grams dans notre cas d'espèce. Concernant cette fois-ci l'apport de la contextualisation, celle-ci nous a permis d'en savoir plus sur certains regroupements *a priori* ambigüs. Considérons le thème "Stopwords & online provider". En récupérant les documents présentant une forte caractéri-

Performance & maneuverability	"performance", "conditions,", . . ., "cornering", "dedicated", . . ., "braking"
Evaluation criteria	"5", "4", . . ., "winter_traction_note :", "handling_note :", "off_road_note :"
Tire references	"zz5's", FK45x", "assy", "ZZ5's", "Falkan"
Tire adhesiveness	"incline", "encountered", "icy", "snowy", "hills"
Durability	"lasted", "far.", "had.", "haven't", "replaced"
Tire adaptability	"fitment :", "road_types :", "road_conditions :", "Mixed", "4x2"
Confidence in the brand	"Continental", "spirited", "review", "Mazda", "snows"
Climatic conditions	"winter", "summer", "winter.", "season", "weather"
Stopwords & online provider	"expand. . .", "member :", . . ., "ZE914", "Blackcircles,", "£100"
Driving experience	"wore", . . ., "Civic, "Ecsta", . . ., "terrible", "noisy."

TAB. 2 – *Extrait des thèmes de l'EHDP-T*

sation par ce thème, nous nous sommes aperçus que Blackcircles est un e-commerçant britannique de pneumatiques. Ce faisant, il nous semble cohérent que ce terme soit entouré d'un modèle de pneu ("ZE914") et d'un prix ("£100"). Nous nous sommes également aperçus que les termes "expand..." et "member :" sont en fait des éléments résiduels post-ETL correspondant au modèle HTML du site. Il s'agit donc de bruitage. Grâce à la récupération de documents ainsi opérée, nous pouvons raffiner l'ETL afférent à cette source particulière.

Synthèse Il semble qu'il y ait hiatus entre ce que montrent les indicateurs et ce que l'on peut déduire de l'examen des thèmes et des plongements de mots. En effet, ce n'est pas parce qu'un modèle est efficace en généralisation qu'il s'interprète bien. De plus, la NPMI est proche de 0 dans plusieurs cas. Ceci indique plutôt une indépendance de co-occurrence des termes. Des éléments d'explication ont été avancés dans Palencia-Olivar et al. (2021) concernant les modèles à plongements. Ceux-ci permettent de tenir compte de relations transitives non-mesurées par la NPMI : si un mot A est proche d'un mot B lui-même proche d'un mot C, alors les mots A et C devraient également être proches. Cela laisse penser que la NPMI est insuffisante à mesurer la proximité entre les mots dans ces cas et que nous manquons d'indicateurs pour mesurer la proximité sémantique.

4 Conclusions et perspectives

Nous avons appliqué l'EDP et l'EHDP à une tâche d'extraction d'insights client relatifs au pneumatique. Ces avis proviennent d'un jeu de données bruité issu d'un cadre industriel. L'EDP et l'EHDP ont obtenu de meilleures performances que d'autres méthodes de l'état de l'art et permettent de raffiner des processus d'ETL. Nous avons également montré que l'utilisation d'une distribution-substitut donne de moins bons résultats que son équivalent. Enfin, nous avons constaté la contradiction entre métriques et qualité perçue des modèles. Le déve-

loppement de métriques efficaces reste un sujet ouvert dont le cadre dépasse largement celui du thème modeling.

Références

Blei, D. et al. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*.

Boyd-Graber, J. et al. (2017). Applications of Topic Models. *Foundations and Trends in Information Retrieval*.

Conover, W. J. (1972). A kolmogorov goodness-of-fit test for discontinuous distributions. *Journal of the American Statistical Association*.

Dieng, A. et al. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*.

Miao, Y. et al. (2016). Neural variational inference for text processing. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*.

Mikolov, T., K. Chen, G. Corrado, et J. Dean (2013). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations*.

Ning, X., Y. Zheng, Z. Jiang, Y. Wang, H. Yang, J. Huang, et P. Zhao (2020). Nonparametric topic modeling with neural inference. *Neurocomputing*.

Palencia-Olivar, M. et al. (2021). Neural Embedded Dirichlet Processes for Topic Modeling. In *Modeling Decisions for Artificial Intelligence*.

Palencia-Olivar, M. et al. (2022). Nonparametric neural topic modeling for customer insight extraction about the tire industry. In *Proceedings of the International Joint Conference on Neural Networks*.

Srivastava, A. et C. Sutton (2017). Autoencoding variational inference for topic models. In *International Conference on Learning Representations*.

Wang, C. et al. (2011). Online Variational Inference for the Hierarchical Dirichlet Process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*.

Summary

In the age of social media, customers have become opinion makers: anyone interested in a product can search for opinions on internet platforms. Web-scraping is often the only way to access them, and despite the use of ETL, the heterogeneity of the data makes the task of extracting insights arduous and leads to the need for ad-hoc tools. To circumvent this problem, we apply the Embedded Dirichlet Process and the Embedded Hierarchical Dirichlet Process in an industrial setting around the case of tires. These non-parametric topic models learn topics and their number, topic embeddings and word embeddings, in order to disambiguate words and to offer more analytical levels. They can also be used to refine ETL processes. EDP and EHDP achieve similar or even higher levels of likelihood than the state-of-the-art techniques tested, without re-runs to find the number of topics.