DIPLÔME NATIONAL DE DOCTORAT

(Arrêté du 25 mai 2016)

Date de la soutenance : **15 mars 2023**

Nom de famille et prénom de l'auteur : **Monsieur PALENCIA Miguel**

Titre de la thèse : *« Une approche par modélisation thématique pour capturer l'insight client dans les médias sociaux »*

Résumé

L'ère des médias sociaux a ouvert de nouvelles perspectives aux entreprises. Cette richesse florissante d'informations se situe en dehors des canaux et des cadres traditionnels de la recherche marketing classique, y compris celui du Marketing Mix Modeling (MMM). Les données textuelles, en particulier, posent de nombreux défis que les praticiens de l'analyse de données doivent relever. Les médias sociaux constituent des sources de documents massives, hétérogènes et bruitées. Les processus industriels d'acquisition de données comprennent une certaine quantité d'Extract-Transform-Load (ETL), cependant, la variabilité du bruit dans les données et l'hétérogénéité induite par les différentes sources créent le besoin d'outils *ad hoc*. En d'autres termes, l'extraction d'insight client dans des contextes bruités et totalement non supervisés est une tâche ardue.

Nous nous intéressons ici à l'extraction de thématiques entièrement non supervisée dans des contextes Big Data bruités. Nous présentons trois approches construites suivant le framework de l'autoencodeur variationnel : l'Embedded Dirichlet Process (EDP), l'Embedded Hierarchical Dirichlet Process (EHDP) et le Dynamic Embedded Dirichlet Process (D-EDP). Ces approches non paramétriques concernant les thèmes présentent la particularité de déterminer des embeddings de mots et des embeddings de sujets. Ces embeddings ne nécessitent pas d'apprentissage par transfert, mais le transfert de connaissances reste possible. Nous testons ces approches sur des jeux de données de référence et des jeux de données liés à l'industrie automobile, issus d'un cas d'utilisation réel. Nous montrons que nos modèles atteignent des performances égales ou supérieures à celles de l'état de l'art et que le domaine du topic modeling bénéficierait de meilleures mesures d'évaluation.

Enfin, nous tirons parti du cadre Autoencoding Variational Bayes (AEVB) et du Deep Learning (DL) pour concevoir une boîte à outils adaptée à la pratique industrielle. Cette boîte à outils permet un entraînement et un développement rapides et évolutifs de nouveaux modèles, comblant ainsi le fossé entre la modélisation statistique et le développement de logiciels et permettant de travailler à la fois avec les méthodes de gestion de projet itératives et les mises à jour de connaissances métier.

Mots-clés : Statistique bayésienne, Topic Modeling, Traitement automatique du langage naturel, Machine Learning, Deep Learning, Business Analytics