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

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

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

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

1 Introduction

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