VIPE : un outil interactif de classification multilabel de messages courts

Frank Meyer *, Sylvie Tricot *
Pascale Kuntz**, Wissam Siblini*,**

*Orange Labs - 2 av. Pierre Marzin - 22 300 Lannion, France
prenom.nom@orange.com,
**Laboratoire d'Informatique de Nantes Atlantique - Site Polytech 44300 Nantes, France
prenom.nom@univ-nantes.fr

Résumé. Nous présentons un outil interactif de classification multilabel développé au sein du groupe Orange et utilisé pour l'analyse d'opinions. Basé sur un algorithme de factorisation rapide de matrice, il permet à un utilisateur d'importer des textes courts (tweets, mails, enquêtes, ...), de définir des labels d'intérêts (« client globalement satisfait », « évoque la rapidité du débit »,...) et de proposer pour chaque texte des recommandations de labels et pour chaque label des recommandations de textes.

1 Introduction

L'analyse d'opinions est un enjeu majeur pour les entreprises qui visent à améliorer en permanence leur relation client. Aux enquêtes par sondage s'ajoutent pour l'analyse les informations extraites sur les médias sociaux. Ces informations contribuent à déterminer le degré d'engouement suscité par les offres d'entreprises, à identifier les différents points de vue et les points de convergence entre les clients, et à recueillir de l'information « fraîche » (Gauzente et al. (2012)). Cependant, l'acquisition des informations utiles est une tâche difficile car les sources complémentaires dont elles sont extraites sont hétérogènes et contiennent des données volumineuses, bruitées et non structurées. Les problèmes associés à l'analyse de ces données rendent le traitement automatique délicat en pratique et l'implication de l'utilisateur est cruciale (Keim et al. (2013)).

L'intégration de l'humain dans la boucle d'apprentissage connaît en effet un essor croissant et des systèmes de classification interactifs ont été développés pour des applications variées : e.g. classification d'images (cueFlick), sélection de fichiers (Smart Selection), classification de gestes (Wekinator), classification de documents (iCluster), tri d'alarmes (CueT). Dans ce cadre, l'utilisateur annote, via une interface adaptée, un nombre limité d'exemples et, à partir de ces quelques exemples, un algorithme d'apprentissage tente de capturer l'expertise pour apprendre un premier modèle prédictif. En fonction de sa satisfaction, l'utilisateur peut arrêter l'apprentissage ou continuer à entraîner le modèle. Les retours expérimentaux menés sur des petits échantillons d'utilisateurs semblent très prometteurs. Cependant, la plupart des systèmes existants se limitent à une classification monolabel où un seul label peut être affecté à la fois à un exemple ; ce qui est peu expressif d'autant plus que les données sont très souvent de