Une méthode hybride pour la prédiction du profil des auteurs

Seifeddine Mechti, Maher Jaoua, Lamia Hadrich Belguith

mechtiseif@gmail.com
maher.jaoua@fsegs.rnu.tn
l.belguith@fsegs.rnu.tn
Faculté des Sciences Economiques et de Gestion (FSEGS),
Laboratoire MIRACL, B.P 1088, 3018, Sfax, Tunisie

Résumé. Dans cet article, nous nous intéressons à la détection du profil des auteurs (âge, genre) à travers leurs discussions. La méthode proposée s'appuie sur la classification automatique qui utilise certaines données extraites d'une manière statistique à partir de corpus source. Nous présentons une méthode hybride qui combine l'analyse de surface dans les textes avec une méthode d'apprentissage automatique. A fin d'obtenir une meilleure gestion de ces données, nous nous sommes basés sur l'utilisation des arbres de décision. Notre méthode a donné des résultats intéressants pour la détection du genre.

1 Introduction

De nos jours, les réseaux sociaux ont connu une croissance importante. Sur Twitter ou sur Facebook la plus part des utilisateurs renseignent seulement 20% de leurs profils. La détection du profil peut être utilisée dans plusieurs domaines, par exemple du point de vue marketing, les entreprises peuvent être intéressées à déterminer quels types de personnes préfèrent leurs produits. Dans la littérature, beaucoup de travaux ont focalisé sur la classification d'une conversation ou d'un texte donné et plus précisément la détection de l'âge de l'auteur, de son genre, sa personnalité, sa langue native, etc. Argamon et al. (2009); Schler et al. (2006); Koppel et al. (2003); Pennebaker (2011).

Les travaux réalisés par Koppel et al. (2003) ont montré qu'au niveau du genre il y a des différences linguistiques entre les hommes et les femmes. En effet, les hommes qui préfèrent catégoriser les choses, utilisent plus de déterminants (le/la, cette/ce, un/une, etc.) et de quantificateurs (deux, plus, peu, etc.). Les femmes, s'intéressent aux relations et plus que les hommes recourent aux pronoms personnels (je, tu, moi, etc.).
La suite de ce papier est organisée comme suit, dans la section 2 nous présentons notre méthode d'apprentissage en focalisant sur le choix des classes et l'algorithme employé. La dernière section présente notre étude expérimentale.

2 Méthode proposée

La méthode proposée s'appuie sur la classifcation de discussions en fonction du genre et de l'âge en se basant sur les probabilités d'apparitions des mots. La dimension genre est repré-