Approches Statistique et Linguistique Pour la Classification de Textes d'Opinion Portant sur les Films

Damien Poirier*, Françoise Fessant*, Cécile Bothorel*
Émilie Guimier de Neef*, Marc Boullé*

*France Telecom R&D, TECH / EASY
2 avenue Pierre Marzin, 22300 Lannion, FRANCE
prénom.nom@orange-ftgroup.com

Résumé. Les sites communautaires sont par nature des lieux consacrés à l'expression et au partage d'avis et d'opinions. *www.flixster.com* est un exemple de site participatif où se retrouvent chaque jour des dizaines de millions de fans dans le but de partager leurs impressions et sentiments sur les films. Une étude approfondie de cette richesse d'information permettrait une meilleure connaissance des utilisateurs, de leurs attentes, de leurs besoins. Pour y parvenir, une étape nécessaire est la classification automatique d'opinion. Dans ce papier nous décrivons trois approches permettant de classer des textes selon l'opinion qu'ils expriment. La première approche consiste à étiqueter les mots porteurs d'opinion à l'aide de techniques linguistiques, ces mots permettant par la suite de classer les textes. La deuxième approche est basée sur des techniques statistiques. La dernière approche est une approche hybride qui combine approche linguistique, pour prétraiter le corpus, et approche statistique, afin de classer les textes.

1 Introduction

Depuis l'émergence du Web 2.0 et des sites communautaires, une quantité croissante de textes non structurés prolifère sur la toile. Ces textes, généralement produits par les internautes, sont très souvent porteurs de sentiments et d'opinions sur des produits, des films, des musiques, etc. Ces données textuelles représentent potentiellement des sources d'information très riches permettant *a priori* de découvrir les attentes, désirs, besoins des utilisateurs ou encore de mesurer la popularité de certains produits, personnalités, décisions politiques, etc.

Le domaine de la fouille d'opinion peut-être divisé en trois sous-domaines (Pang et Lee, 2008) :

– l'identification des textes d'opinion, qui consiste à identifier dans une collection textuelle les textes porteurs d'opinion, ou encore à localiser les passages porteurs d'opinion dans un texte. Plus précisemment, on parle ici de classer les textes ou les parties de texte selon qu'ils sont objectifs ou subjectifs (Stoyanov et Cardie, 2008) ;
– le résumé d'opinion, qui consiste à rendre l'information rapidement et facilement accessible en mettant en avant les opinions exprimées et les cibles de ces opinions présentes dans un texte. Ce résumé peut être textuel (extraction des phrases ou expressions