Analysis of Forum Posts Written by Patients and Health Professionals

Amine ABDAOUI[a], Jérôme AZE[a], Sandra BRINGAY[a], Natalia GRABAR[b] and Pascal PONCELET[a]
[a]*LIRMM UM2 CNRS, UMR 5506, 161 Rue Ada, 34095 Montpellier, France*
[b]*STL UMR 8163 CNRS, Université Lille 3 et Lille 1, France*

Keywords. Text categorization, text mining, online health fora.

Introduction

For information retrieval purposes, it may be interesting to categorize posts produced by patients and those produced by health professionals on online health fora.

In this work, we evaluated a supervised approach based on n-grams (vocabulary), emotion markers, uncertainty markers and misspellings to distinguish the two categories of posts on the French health forum AlloDocteurs.fr[1].

Methods

First, external resources were used to annotate medical words, emotion markers and uncertainty markers on two cleaned corpus which where balanced between the two roles: a train corpus of 4000 posts and a test corpus of 450 posts.

Then, a linguistic pre-processing step was applied to replace slangs, users' tags, emails, hypertext links, etc. and to correct and count misspellings.

Finally, a feature selection step was applied to select the most discriminant n-grams.

Results

Different combinations of features (uni-grams, bi-grams, emotions, uncertainty markers and misspellings) have been tested with four data mining algorithms (SVM SMO, RandomForest, NaïveBayes and JRip) to distinguish the two categories of posts.

Discussion

The results obtained by uni-grams and bi-grams have shown very high F-scores (up to 0.94), while the results obtained by considering only emotions, uncertainty markers and misspellings have shown low F-scores (from 0.52 to 0.76) comparing to chance (0.50).

Finally, conducted experiments by grouping all the features have slightly improved the classification performance (F-scores up to 0.95).

The main perspective is to test the learned models on new fora.

1 www.allodocteurs.fr/forum-rubrique.asp [data collected on: 19-11-2013]