Analysis of Forum Posts Written by Patients and Health Professionals
Amine Abdaoui, Jérôme Azé, Sandra Bringay, Natalia Grabar, Pascal Poncelet

To cite this version:

HAL Id: lirmm-01130727
https://hal-lirmm.ccsd.cnrs.fr/lirmm-01130727
Submitted on 12 Mar 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Analysis of Forum Posts Written by Patients and Health Professionals

Amine Abdaoui, Jérôme Azé, Sandra Bringay, Natalia Grabar and Pascal Poncelet

a LIRMM UM2 CNRS, UMR 5506, 161 Rue Ada, 34095 Montpellier, France
b STL UMR 8163 CNRS, Université Lille 3 et Lille 1, France

Context: Online health fora are increasingly visited by both patients and health professionals. For online fora visitors, posts written by health professionals may be more interesting since the professionals are able to well explain the problems, the symptoms, correct false affirmations and give useful advices, etc.

Objective: To automatically distinguish posts written by health professionals from those written by patients.

Intuition: Use a supervised approach and test the following features with different classification models:
- Vocabulary
- Emotions
- Uncertainty
- Question marks
- Misspellings

Methods:
4000 posts divided equally between patients and health professionals

Cleaning (quotes, short posts, etc.)
Annotation (emotions, uncertainty, questions and misspellings)

preprocessing (feature selection, slang replacement, spelling correction, etc.)

Classification (Weka)

Results: 10-folds cross validation (f-measures)

<table>
<thead>
<tr>
<th>Features</th>
<th>Number of features</th>
<th>SVM SMO</th>
<th>Naive Bayes</th>
<th>Random Forest</th>
<th>JRip</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>1,120</td>
<td>0.938</td>
<td>0.869</td>
<td>0.901</td>
<td>0.892</td>
</tr>
<tr>
<td>U+B</td>
<td>2,160</td>
<td>0.921</td>
<td>0.865</td>
<td>0.902</td>
<td>0.889</td>
</tr>
<tr>
<td>EM</td>
<td>1</td>
<td>0.565</td>
<td>0.529</td>
<td>0.564</td>
<td>0.609</td>
</tr>
<tr>
<td>UM</td>
<td>1</td>
<td>0.682</td>
<td>0.660</td>
<td>0.657</td>
<td>0.689</td>
</tr>
<tr>
<td>MI</td>
<td>1</td>
<td>0.636</td>
<td>0.601</td>
<td>0.641</td>
<td>0.653</td>
</tr>
<tr>
<td>QM</td>
<td>1</td>
<td>0.560</td>
<td>0.516</td>
<td>0.613</td>
<td>0.653</td>
</tr>
<tr>
<td>EM+UM+MI+QM</td>
<td>4</td>
<td>0.751</td>
<td>0.66</td>
<td>0.725</td>
<td>0.751</td>
</tr>
<tr>
<td>U+EM+UM+MI+QM</td>
<td>1,124</td>
<td>0.940</td>
<td>0.872</td>
<td>0.901</td>
<td>0.900</td>
</tr>
<tr>
<td>U+B+EM+UM+MI+QM</td>
<td>2,164</td>
<td>0.927</td>
<td>0.866</td>
<td>0.906</td>
<td>0.897</td>
</tr>
</tbody>
</table>

Acronyms:
- U: Unigrams
- B: Bigrams
- EM: Emotion Markers
- UM: Uncertainty Markers
- MI: Misspellings
- QM: Question Marks