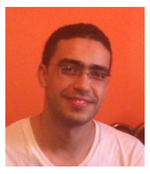


Analysis of Forum Posts Written by Patients and Health Professionals

Amine Abdaoui[a], Jérôme Azé[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

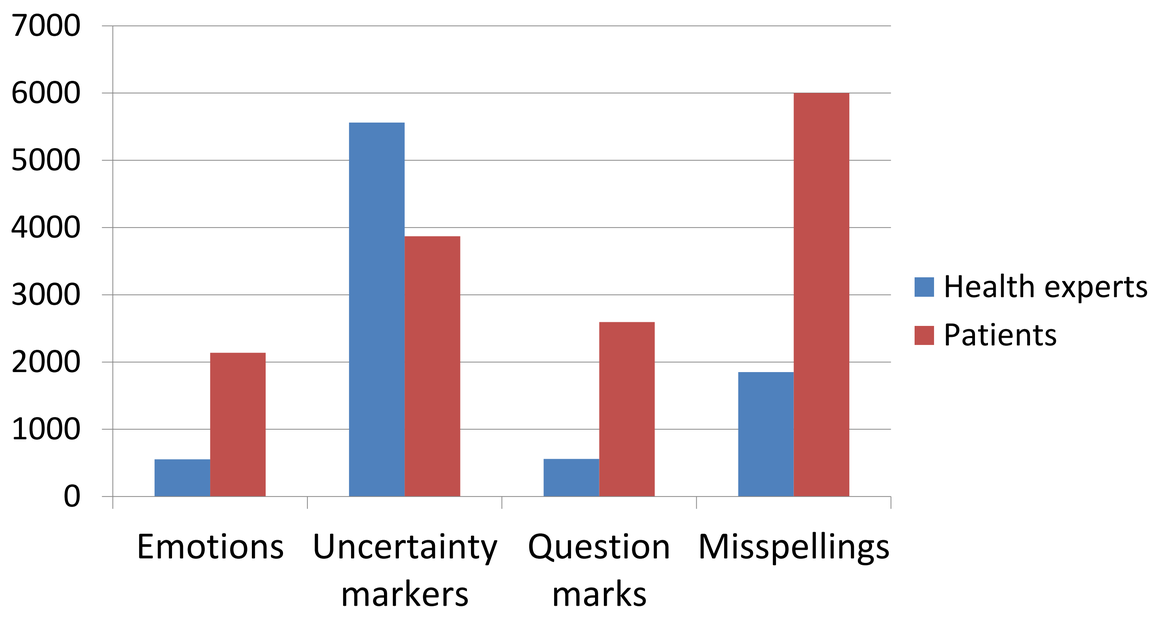
Context: Online health fora are increasingly visited by both patients and health professionals. For online fora visitors, posts written by health professionals may be more interesting since the professionals are able to well explain the problems, the symptoms, correct false affirmations and give useful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

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

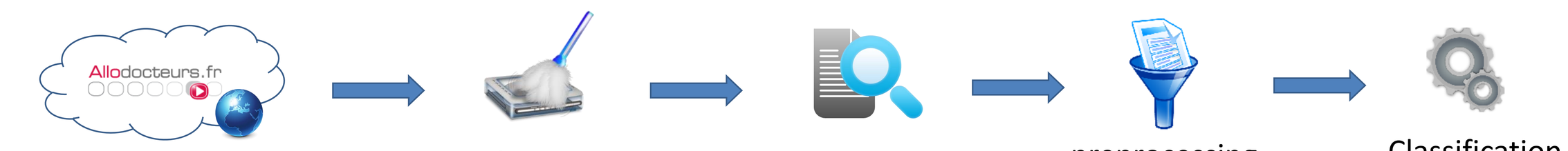
Intuition: Use a supervised approach and test the following features with different classification models:

- Vocabulary
- Emotions
- Uncertainty
- Question marks
- Misspellings

The vocabulary used by each category



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)

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

Acronyms:

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

Laboratoire Informatique Robotique Microélectronique Montpellier
UNIVERSITÉ MONTPELLIER 2 SCIENCES ET TECHNIQUES
cnrs
STL savoirs langage textes