



**HAL**  
open science

## Analysis of Forum Posts Written by Patients and Health Professionals

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

► **To cite this version:**

Amine Abdaoui, Jérôme Azé, Sandra Bringay, Natalia Grabar, Pascal Poncelet. Analysis of Forum Posts Written by Patients and Health Professionals. MIE: Medical Informatics Europe, Aug 2014, Istanbul, Turkey. 25th European Medical Informatics Conference, pp.1185-1185, 2014. lirmm-01130727

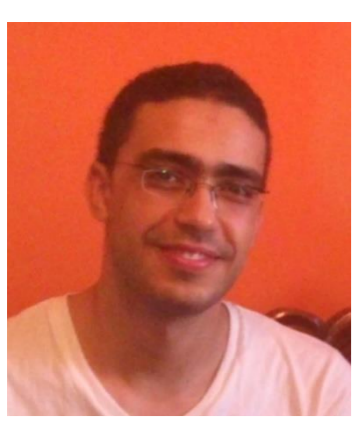
**HAL Id: lirmm-01130727**

**<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<sup>a</sup>, Jérôme Azé<sup>a</sup>, Sandra Bringay<sup>a</sup>, Natalia Grabar<sup>b</sup> and Pascal Poncelet<sup>a</sup>*

<sup>a</sup> LIRMM UM2 CNRS, UMR 5506, 161 Rue Ada, 34095 Montpellier, France

<sup>b</sup> 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.

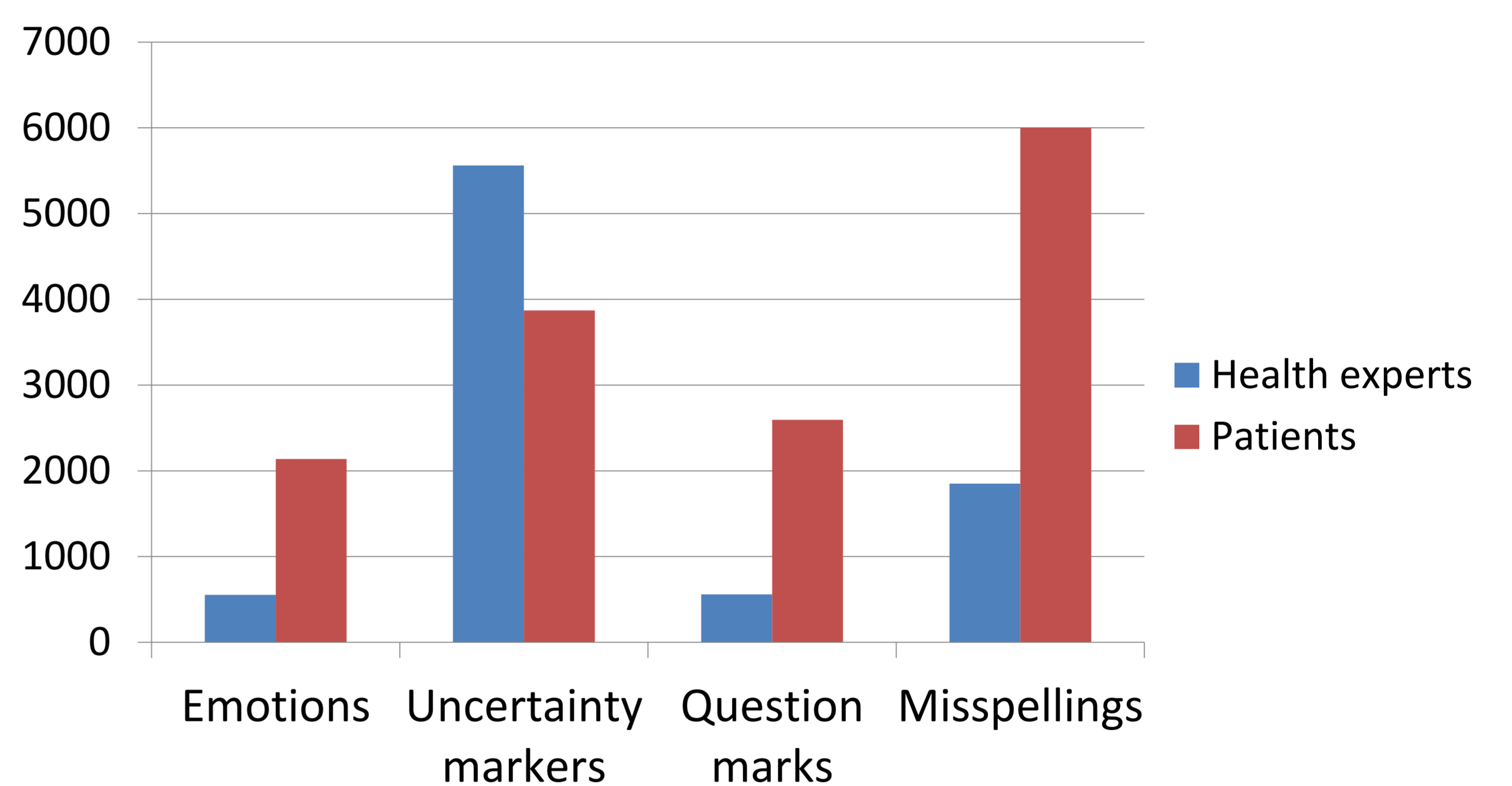
**Objective:** To automatically distinguish posts written by health professionals from those written by patients.

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

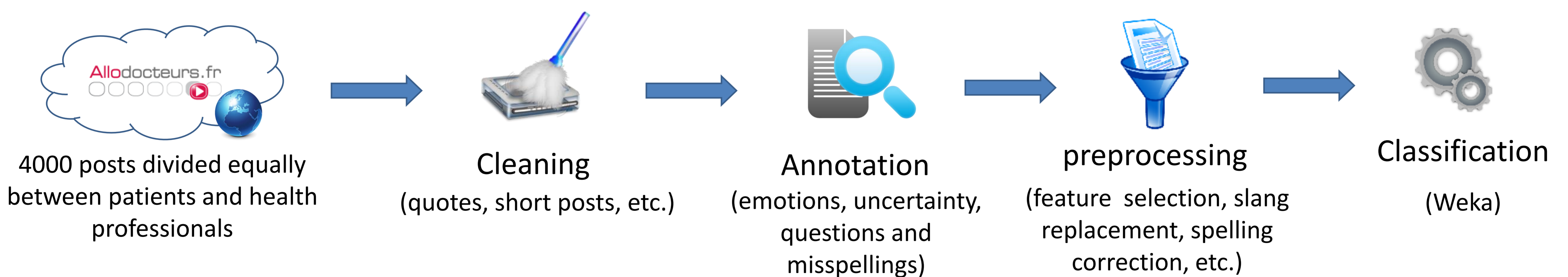
- Vocabulary
- Emotions
- Uncertainty
- Question marks
- Misspellings



The vocabulary used by each category



## Methods:



## 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
<b>U+EM+UM+MI+QM</b>	<b>1,124</b>	<b>0.940</b>	<b>0.872</b>	<b>0.901</b>	<b>0.900</b>
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