Health Informatics Journal

Original Article

Expertise in French health forums

Health Informatics Journal
2019, Vol. 25(1) 17–26

DOI: 10.1177/1460458216682356
journals.sagepub.com/home/jhi
SAGE

Amine Abdaoui and Jérôme Azé
Université Montpellier (UM), France

Sandra Bringay
Université Montpellier (UM), France; Université Paul-Valéry Montpellier 3 (UM3), France

Natalia Grabar
Université Lille 1 and Université Lille 3, France

Pascal Poncelet
Université Montpellier (UM), France

Abstract

More and more health websites hire medical experts (physicians, medical students, experienced volunteers, etc.) and indicate explicitly their medical role in order to notify that they provide high-quality answers. However, medical experts may participate in forum discussions even when their role is not officially indicated. Detecting posts written by medical experts facilitates the quick access to posts that have more chances of being correct and informative. The main objective of this work is to learn classification models that can be used to detect posts written by medical experts in any health forum discussions. Two French health forums have been used to discover the best features and methods for this text categorization task. The obtained results confirm that models learned on appropriate websites may be used efficiently on other websites (more than 98% of F1-measure has been obtained using a Random Forest classifier). A study of misclassified posts highlights the participation of medical experts in forum discussions even if their role is not explicitly indicated.

Keywords

author-profiling, health forums, medical expertise, text categorization

Corresponding author:
Amine Abdaoui, Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), Université Montpellier (UM), UMR 55, 860 Saint Priest Street, 34095 Montpellier, France.
Email: abdaoui@lirmm.fr

Introduction

Health forums are increasingly visited by both sick and healthy users when they want to get help and information related to their health.[1] According to a study conducted by the Health On the Net (HON; www.healthonnet.org) foundation, 50 percent of e-patients use online health forums to acquire medical information. However, these forums are not limited to patients. More and more frequently, a significant number of medical experts are involved in online discussions.[2] Indeed, some medical websites hire health experts (physicians, medical students, volunteers, etc.) and indicate explicitly their role. Others visit health forums unofficially and answer the patient's questions without a special indication about their expertise. Being experts, they are able to clearly explain the problems, the symptoms, to correct false affirmations, and to give precise and trustworthy answers. Furthermore, patients may acquire expertise through their own experience with a particular disease. After recovery, many of them go back to online forums in order to share their experience and help other patients. The aim of this study is to distinguish between posts written by medical experts (health practitioner or experienced patients) and by non-expert users.

Identifying expert posts may have many useful applications. For example, highlighting these posts facilitates the identification of best answers that are more likely to be trustworthy and informative. Furthermore, expert posts detection can help forum administrators to find new potential moderators who have enough expertise to answer the forum questions and moderate the discussions. Finally, this information allows studying the expertise evolution of the forum users over time. The main objective of our study is to use posts from websites, in which the medical roles are indicated, in order to build efficient classification models that can predict the potential expertise in other health forums. We intend to tackle the question through the analysis of the posts content. The proposed method uses supervised machine-learning algorithms in order to perform text categorization. Similar methods have been developed for the author-profiling tasks PAN[3–5] in order to identify the age, gender, and personality traits of a text author. In fact, companies are increasingly interested in discovering these characteristics about users who liked or disliked their products based on web blog posts.[5,6] Similarly, health organizations can extract valuable knowledge from expert and non-expert posts written on health forums.[7] They may study and use this knowledge in order to improve their practice (treatments, medications, etc.).

Many features can be exploited in order to perform author-profiling from text posts.[8] Here, we focus on those that can be efficient for medical expertise categorization. Tapi-Nzali et al.[9] mentioned that medical experts and patients use different vocabularies. Patients write more about symptoms and about themselves: I have a headache, etc., while experts should write more about treatments and about the non-experts: you should pass a mammography test, etc. Therefore, a bag of words configuration is considered. Rangel and Rosso[10] studied the impact of emotions and sentiments in author-profiling (age and gender). They proposed an emotion graph to model the way people use the language and the emotions when writing. They obtained, respectively, the first and the second best results for age and gender on the Spanish partition of PAN 2013 corpus. Grabar et al.[11] compared documents written by medical doctors and researchers (clinical reports and scientific literature) with the patient discourse (discussions from health forums). They observed differences in the use of descriptors like uncertainty markers, non-lexical (smileys, repeated punctuations, etc.) and lexical emotional markers, and medical terms related to disorders, medications, and procedures. In this work, these features are considered along with further annotations and preprocessing in order to evaluate the most representative components of a forum post that allow to perform efficiently medical expertise categorization.

The rest of the article is organized as follows. Section "Materials and methods" introduces the studied corpora and details of the proposed method. Section "Experiments" presents the obtained

results, and section "Discussions" discusses them. Finally, section "Conclusion and prospects" concludes and gives our main prospects.

Materials and methods

This section discusses the used corpora and the proposed methods, which are based on supervised machine-learning. Indeed, these methods are known to perform well when trained on appropriate annotated datasets. In our case, many online forums indicate explicitly the medical expertise of their users, which provides good and inexpensive annotated datasets.

Corpora

Two French corpora have been collected from two health forums as described below.

AlloDocteurs.fr is a French health forum covering a large number of topics related to health such as alcoholism, pregnancy, and sexuality. A total of 16,000 messages posted from June 2009 to November 2013 have been collected. The forum contains both expert and non-expert users. Medical experts include professional physicians and medical students. Even if their number is limited (16 medical experts over more than 6000 registered users), their participation in the forum exchanges is important. Indeed, they posted more than 3000 posts among the 16,000.

MaSanteNet.com is an online ask-the-doctor service that allows users to submit one or more questions to two doctors. The range of topics covered is also large. Users can ask questions on more than 20 different topics such as nutrition, dermatology, and pregnancy. More than 12,000 messages posted from January 2011 to March 2014 have been collected from this website. All the questions published on the website have answers. Therefore, the collected posts are equitably divided between patients' questions and doctors' answers.

Cleaning

Once the two corpora are collected, a cleaning step is applied to improve their quality. First, quotes present inside some posts are filtered out. Indeed, some medical experts quote the questions before answering them, which may introduce non-expert statements into posts of health professionals. All additional pieces of text, such as author signatures and date of the last modification, are removed. Finally, posts with less than 10 characters (blank posts or very short posts such as "yes") are considered as irrelevant and also removed since they do not convey enough information. Figure 1 presents the number of posts and words in the obtained datasets. On the one hand, it appears that the first corpus has fewer posts than the second one: approximately 4400 posts for AlloDocteurs and approximately 12,000 posts for MaSanteNet. On the other hand, it appears that in both datasets, posts written by non-experts are longer than those written by medical experts.

Preprocessing

Texts from social media have several linguistic peculiarities that may influence the classification performance.[12] Therefore, the following preprocessing steps are applied:

Slang. Some abbreviations are frequently used in social media. They are replaced by the corresponding standard text (e.g. "lol" is replaced by "lot of laugh").

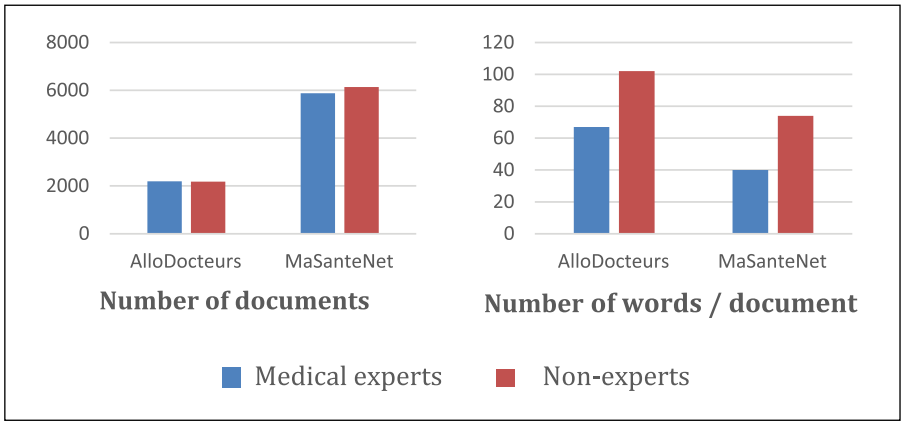

Figure 1. Number of documents and the average number of words per document in each corpus.

User tags. User tags are identified in our corpora and replaced by the word "tag" (e.g. "@Diana …" becomes "tag …").

Hyperlinks and emails. Hypertext links are replaced by the word "link" and email addresses are replaced by the word "mail."

Pseudonyms. The medical expert pseudonyms, previously extracted from each website, are used to replace all their apparitions inside the posts by the word "fdoctor." Similarly, pseudonyms of non-experts are extracted and used for their replacement by the word "fpatient."

Lowercasing and spelling correction. All words are lowercased and processed with the spell checker Aspell (www.aspell.net, accessed 26 February 2016). The default Aspell French dictionary is expanded with all the pseudonyms and all the medical words extracted from our corpora. The medical terms are obtained after an annotation step as described below.

Annotations

In order to categorize the discourse of medical experts and the discourse of non-experts, the descriptors proposed in Grabar et al.[11] have been annotated using the Ogmios platform.[13] This annotation step allows us to include them easily as features in the classification step.

Medical concepts. Terms belonging to three semantic types (diseases, treatments, and procedures) are detected using the following medical resources: the Systematized Nomenclature of Human and Veterinary Medicine (www.ihtsdo.org/snomed-ct, accessed 26 February 2016), the Thériaque database (www.theriaque.org, accessed 26 February 2016), the Unified Medical Language System (www.nlm.nih.gov/research/umls, accessed 26 February 2016), and the list of authorized medication that can be marketed in France.

Emotions. A French emotion lexicon made by the authors[14] is used to annotate adjectives, verbs, and nouns conveying seven types of emotions (joy, trust, sadness, anger, fear, disgust, and surprise). The lexicon contains about 14,000 emotional terms. In addition, some non-lexical expressions of

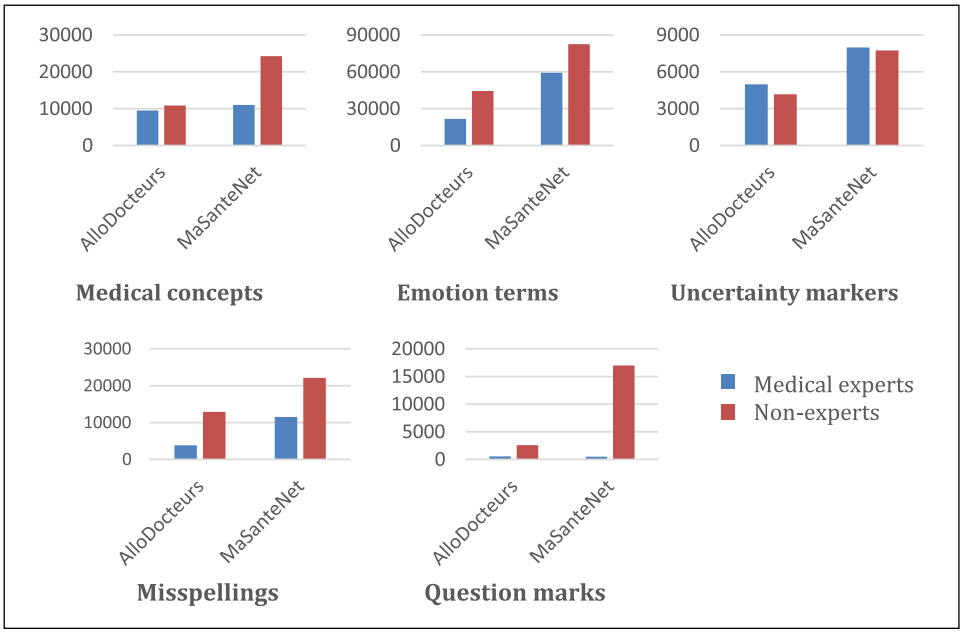

Figure 2. Number of medical concepts, emotion terms, uncertainty markers, misspellings, and question marks in each corpus.

emotions, such as repeated letters, repeated punctuation signs, smileys, slang, and capital letters, are detected and annotated with specifically designed regular expressions.

Uncertainty. A set of uncertainty words[11] is used to annotate verbs, nouns, adjectives, and even adverbs conveying uncertainty meaning (e.g. to seem, possible, and probably). Three levels of uncertainty are considered: weak, medium, and strong.

Classification

Features. In addition to the features based on the annotation step, the number of misspellings and question marks are included in the categorization. Figure 2 shows the number of medical concepts, emotions terms, uncertainty markers, misspellings, and question marks in each benchmark. It appears that non-experts use more medical concepts and emotion terms, ask much more questions, and do more spelling mistakes, while medical experts use slightly more uncertainty markers (usually to make an uncertain diagnosis). Therefore, 15 attributes representing these descriptors are included in our classification task (medical concepts: three attributes, emotion terms: seven attributes, uncertainty markers: three attributes, questions: one attribute, and misspellings: one attribute). For each attribute, we compute the number of occurrences normalized by the corresponding post length. The length of each post corresponds to the number of words it contains. We call these attributes as "Dictionary-Based Features."

Moreover, a bag of words representation is considered. Words that appear at least two times in the training sets are included. Each word is represented by his normalized number of occurrences (number of occurrences divided by the corresponding post length). In the next section, we evaluate all these features on the classification performances.

Table 1. Weighted F-scores obtained with 10-fold cross validation on AlloDocteurs.

Feature group	SVM SMO	J48	Random Forest	JRip
Bag of words	92	90.6	92.1	89.7
Dictionary-based markers	71.6	73	74	75
Bag of words + dictionary-based markers	92.7	90.7	92.7	90.3

SVM SMO: support vector machines–sequential minimal optimization.

Feature selection. Feature subset selection is applied to select the most discriminant features: those that frequently appear in only one category of posts. Therefore, the selected features should characterize one category of users. The information gain method is used as a filter to select attributes in each experiment.

Classifiers. The Weka data-mining platform[15] is used to learn the classification models. We tested the following models that have been reported in the literature as efficient for text categorization:[16] support vector machines–sequential minimal optimization (SVM SMO), decision trees (J48 and Random Forest), and rule-based models (JRip). Since feature selection does not remove redundant attributes, models that assume the independency of the features (such as Naïve Bayes) are not adapted. The Weka default configuration is used for each classification model.

Evaluation metric. Weighted F1-scores are used to evaluate the classification performances of different combinations of features and algorithms. F1-score is computed as the harmonic mean of the precision and the recall of a given class. Weighted F-score is the mean of all class F-scores weighted by the proportion of elements in each class. For a balanced dataset, chance will produce a weighted F-score of 0.5 that can be considered as a baseline for evaluating our results.

Experiments

In this section, the conducted experiments and the obtained results are described.

Cross validation

First, 10-fold cross validation has been performed on each dataset separately. K-fold cross validation is a validation technique that randomly partitions the dataset into k equal size subsets. A single subset is used for testing, while the remaining k−1 subsets are used as training set. This process is repeated k times so that each of the k subsets is used as a testing set exactly once. The features construction, selection, and classification models are learned on the training subset of each fold. Moreover, the same training and testing sets are used to learn and test our four classification models in each fold.

Tables 1 and 2 show that on both datasets, bag of words induce high weighted F1-measures. They obtain more than 0.90 on AlloDocteurs and perfect classification F1-measures[1] on MaSanteNet. However, the dictionary-based markers induce lower weighted F1-measures: between 0.70 and 0.75 on AlloDocteurs and between 0.55 and 0.60 on MaSanteNet. Regarding the classification models, SVM SMO and Random Forest obtained the highest F1-measures on MaSanteNet. Finally, the use of the dictionary-based features along with the bag of words configuration does not change the results (the obtained F1-measures are almost the same as those obtained only with bag of words). The presented results may indicate that our models are

Table 2. Weighted F-scores obtained with 10-fold cross validation on MaSanteNet.

Feature group	SVM SMO	J48	Random Forest	JRip
Bag of words	100	100	100	100
Dictionary-based markers	88.9	91.6	93.6	92
Bag of words + dictionary-based markers	100	100	100	100

SVM SMO: support vector machines–sequential minimal optimization.

Table 3. Weighted F-scores obtained with AlloDocteurs as training set and MaSanteNet as testing set.

Feature group	SVM SMO	J48	Random Forest	JRip
Bag of words	96.6	97.7	98	96.9
Dictionary-based markers	57	62.1	69.6	69.6
Bag of words + dictionary-based markers	96	97.3	98.2	96.6

SVM SMO: support vector machines–sequential minimal optimization.

Table 4. Weighted F-scores obtained with AlloDocteurs as training set and MaSanteNet as testing set.

Feature group	SVM SMO	J48	Random Forest	JRip
Bag of words	37.3	33.3	46.3	33.3
Dictionary-based markers	57.1	52.9	53.2	55.3
Bag of words + dictionary-based markers	37.5	33.3	43.7	33.3

SVM SMO: support vector machines–sequential minimal optimization.

dependent on the forum used for learning. Therefore, we evaluate the genericity of the models learned on each forum and test them on the other forum.

Training and testing on different datasets

In this study, we assume that models learned on specific forums can be used efficiently on other forums. In order to evaluate this claim, two more experiments are conducted. In each experiment, features and classification models are constructed and learned on one dataset and tested on the other dataset.

Table 3 shows that models learned on AlloDocteurs obtain significantly high F1-measures. The bag of words used alone or with the dictionary-based features induces more than 0.95 weighted F1-measures when tested on MaSanteNet. Once again, Random Forest obtains the highest F1-measure. The dictionary-based features induce F1-measures between 0.55 and 0.70. These results show that the models learned on AlloDocteurs remain highly efficient when applied on MaSanteNet. However, Table 4 shows that the classification models learned on MaSanteNet obtain low F1-measures. The weighted F1-measures of the bag of words features used alone or with the dictionary-based features drop significantly when tested on AlloDocteurs (between 0.35 and 0.55). The weighted F1-measures obtained by the dictionary-based features drop slightly when tested on AlloDocteurs (between 0.50 and 0.60). SVM SMO induces the highest F1-measure using these features. Finally, we can conclude that the bag of words models learned on MaSanteNet are extremely context dependent, which makes this forum inappropriate for training generic models.

Discussions

In this section, we discuss the obtained results and describe a manual error analysis step.

Results interpretation

Despite the high F1-measures obtained with cross validations on both datasets, the models learned on AlloDocteurs remain efficient when applied on MaSanteNet. However, those learned on MaSanteNet gave lower F1-measures when applied on AlloDocteurs. These results can be explained by the fact that the first website is a health forum, in which 16 medical experts participate in the forum discussions. They post messages in any thread where their expertise is needed, which make the discourse of the medical experts more extensive and diversified. Therefore, models learned on this forum may cover topics and medical discourse may also be found on MaSanteNet. However, MaSanteNet is a limited health forum (an ask-the-doctor service) in which only two medical experts answer the questions. There are no long discussions since each thread contains only one question and one answer. The answers are formed following the same pattern, which makes the discourse of the medical experts specific to this website. For this reason, MaSanteNet appears to be less suitable for learning classification models that can be used on other forums.

Using emotions, uncertainty markers, and medical concepts, Grabar et al.[11] obtained F1-measures between 0.91 and 0.95 when classifying forum posts produced by patients and clinical reports produced by medical experts. Our study shows the limits of using these markers in categorizing the patients' discourse and the medical experts' discourse when the text documents are of the same nature (forum posts). Our results suggest to use bag of words features, which are the most adapted to perform such categorization. This result confirms those obtained in the author-profiling challenge PAN,[5] where the best systems used content-based features (bag of words, words n-grams, TF-IDF n-grams, etc.).

Error analysis

An error analysis of the 10-fold cross validation applied on AlloDocteurs has been performed. In each fold, four classification algorithms have been trained on 90 percent of the data using all the features (bag of words and dictionary-based markers) and tested on the remaining 10 percent. If at least three algorithms agree to classify a post to the wrong category (with respect to the role given on the website), the post is to be studied manually. Therefore, this study included 164 posts among which 107 were written by patients but classified as medical experts and 57 which were written by medical experts but classified as patients.

On the one hand, the manual analysis of the 107 posts classified as medical experts allowed us to find new users having medical expertise but not indicated as such on the website. They may be either medical physicians (e.g. "… many similar cases come to see us in the hospital …") or only users who had the same experience before (e.g. "… the pain will disappear in few days, my mother had the same surgery …"). These users posted 79 messages among the 107, which confirms that medical experts may participate in the discussions even if their role is not explicitly indicated. In this case, only 47 posts have been considered as misclassified. On the other hand, the manual analysis of the 57 posts that has been written by medical experts and classified as patients showed that medical experts may have the same discourse as patients (e.g. they may ask questions). This observation highlights that even medical experts may lack expertise in a particular topic or need precision on the patient's condition.

Conclusion and prospects

In this article, we presented a supervised learning approach designed to distinguish posts written by medical experts and by patients in French online health forums. The performed experiments show very high F-scores with bag of words features. Moreover, they confirm that models learned on appropriate forums where many medical experts participate in various discussions can be applied on other websites with satisfactory results. Finally, analyzing the misclassified posts allowed us to find out that medical experts may write posts in online health forums even if their medical role is not indicated on the website. The study of the misclassified posts also shows that the expertise of a user may change according to the discussed topic.

As future work, a temporal dimension may be included to highlight the evolution of the author's expertise over time. Indeed, users may develop expertise especially in the case of chronic diseases. A study of a French forum on breast cancer (www.cancerdusein.org/forum, accessed 23 March 2016) shows that the discourse of patients changes according to the evolution of the disease and treatments.[17] Some of them start as information consumers and progressively acquire the status of information providers. Usually, for many of them, once they recover from the disease, they want to go back online to share their knowledge and experience with other patients. This observation also stands for technical and programming forums (www.stackoverflow.com/help/whats-reputation, accessed 23 March 2016), where a programmer begins as non-experienced and gradually acquire expertise. A temporal dimension may be easily included to this method in order to highlight those changes.

Another interesting research issue to detect expert users instead of expert posts consists of mining the answers received by a given user in the forum. We are working on another textual content-based method that analyzes the posts addressed to a user instead of the posts written by him.[18] Natural language processing methods may be used to detect the trust expressed in these answers. This trust may be inferred by searching agreement, disagreement, and thanking expressions. The first step is to identify the recipient(s) of each post. Then, the posts addressed to each user must be evaluated to detect trust expressions. Finally, the expertise of the user may be computed by measuring the number of positive and negative replies.

Acknowledgements

This work was based on studies supported by the "Maison des Sciences de l'Homme" (MSH-M) within the framework of the French project "Patients Mind" (http://www.msh-m.fr/la-recherche/programmes-clos/article/a-quoi-peuvent-bien-penser-les). It was also supported by the Algerian Ministry of Higher Education and Scientific Research (www.mesrs.dz) by funding a PhD grant.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Househ M, Borycki E and Kushniruk A. Empowering patients through social media: the benefits and challenges. *Health Informatics J* 2014; 20(1): 50–58.
2. Antheunis ML, Tates K and Nieboer TE. Patients' and health professionals' use of social media in health care: motives, barriers and expectations. *Patient Educ Couns* 2013; 92(3): 426–431.

3. Rangel F, Rosso P, Koppel M, et al. Overview of the author profiling task at PAN 2013. In: *Conference and labs of the evaluation forum (CLEF)*, Valencia, 23–26 September 2013.
4. Rangel F, Rosso P, Chugur I, et al. Overview of the 2nd author profiling task at PAN 2014. In: *Conference and labs of the evaluation forum (CLEF)*, Sheffield, 15–18 September 2014, pp. 898–927. CEUR Workshop Proceedings.
5. Rangel F, Rosso P, Potthast M, et al. Overview of the 3rd author profiling task at PAN 2015. In: *Conference and labs of the evaluation forum (CLEF)*, Toulouse, 8–11 September 2015.
6. Nguyen D, Gravel R, Trieschnigg D, et al. "How old do you think I am?" A study of language and age in Twitter. In: *Seventh international AAAI conference on weblogs and social media*. Boston, MA, 8–11 July 2013, pp. 439–448. Palo Alto, CA: AAAI Press.
7. Murdoch TB and Detsky AS. The inevitable application of big data to health care. *J Am Med Assoc: JAMA* 2013; 309(13): 1351–1352.
8. Weren ER, Kauer AU, Mizusaki L, et al. Examining multiple features for author profiling. *J Inf Data Manag* 2014; 5(3): 266.
9. Tapi Nzali MD, Bringay S, Lavergne C, et al. *Construction d'un vocabulaire patient/médecin dédié au cancer du sein à partir des médias sociaux*. In: *IC: Ingénierie des Connaissances*, Rennes, 1–3 June 2015, pp. 9–20. INRIA.
10. Rangel F and Rosso P. On the impact of emotions on author profiling. *Inform Process Manag* 2016; 52(1): 73–92.
11. Grabar N, Chauveau-Thoumelin P and Dumonet L. Medical discourse and subjectivity. In:Guillet F, Pinaud B, Venturini G, et al. (eds) *Advances in knowledge discovery and management (studies in computational intelligence)*. Gewerbestrasse: Springer International Publishing, 2016, pp. 33–54.
12. Farzindar A and Inkpen D. *Natural language processing for social media*. San Rafael, CA: Morgan & Claypool Publishers, 2015, p. 166.
13. Hamon T and Nazarenko A. Le développement d'une plate-forme pour l'annotation spécialisée de documents Web: retour d'expérience. *Trait Autom Lang* 2008; 49(2): 127–154.
14. Abdaoui A, Azé J, Bringay S, et al. FEEL: a French Expanded Emotion Lexicon. *Language Resources and Evaluation* 2016; 1–23.
15. Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update. *ACM SIGKDD Explor Newsl* 2009; 11(1): 10–18.
16. Sebastiani F. Machine learning in automated text categorization. *ACM Comput Surv: CSUR* 2002; 34(1): 1–47.
17. Opitz T, Azé J, Bringay S, et al. Breast cancer and quality of life: medical information extraction from health forums. In: *Medical Informatics Europe: MIE*, Istanbul, Turkey, 27–31 August 2014, pp. 1070–1074. IOS Press.
18. Abdaoui A, Azé J, Bringay S, et al. E-Patient reputation in Health Forums. In: *MEDINFO*, São Paulo, Brazil, 19–23 August 2015, pp. 137–141. IOS Press.