HAL
open science

Predicting Medical Roles in Online Health Fora

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

To cite this version:

Amine Abdaoui, Jérôme Azé, Sandra Bringay, Natalia Grabar, Pascal Poncelet. Predicting Medical Roles in Online Health Fora. SLSP: Statistical Language and Speech Processing, Oct 2014, Grenoble, France. pp.247-258, 10.1007/978-3-319-11397-5_19 . lirmm-01130749

HAL Id: lirmm-01130749

https://hal-lirmm.ccsd.cnrs.fr/lirmm-01130749v1

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.

Predicting Medical Roles in Online Health Fora

Amine ABDAOUI[a], Jerome 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 Lille 1, France

Abstract. Online health fora are increasingly visited by patients to get help and information related to their health. However, these fora are not limited to patients: a significant number of health professionals actively participate in many discussions. As experts their posted information are very important since, they are able to well explain the problems, the symptoms, correct false affirmations and give useful advices, etc. For someone interested in trusty medical information, obtaining only these kinds of posts can be very useful and informative. Unfortunately, extracting such knowledge needs to navigate over the fora in order to evaluate the information. Navigation and selection are time consuming, tedious, difficult and error-prone activities when done manually. It is thus important to propose a new method for automatically categorize information proposed both by non-experts as well as by professionals in online health fora. In this paper, we propose to use a supervised approach to evaluate what are the most representative components of a post considering vocabularies, uncertainty markers, emotions, misspellings and interrogative forms to perform efficiently this categorization. Experiments have been conducted on two real fora and shown that our approach is efficient for extracting posts done by professionals.

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

1 Introduction

The Text Mining and Natural Language Processing communities have extensively investigated the huge amount of data on online health fora for different purposes, such as: classifying lay requests to an internal medical expert [1], assisting moderators on online health fora [2], identifying sentiments and emotions [3], identifying the targets of the emotions [4], etc. Indeed, online health fora are increasingly visited by both sick and healthy users to get help and information related to their health [2]. However, these fora are not limited to non-health professional users. More and more frequently, significant number of medical experts is involved in online discussions. For example, many websites, also called “Ask the doctor” services, allow non-health expert users to interact with medical experts [1].

For users searching medical information in online health fora, it may be interesting to automatically distinguish between posts made by health professionals and those made by lay men. For instance, users may be more interested by posts made by health professionals, who should give more precise and trustier answers. Some medical web-

sites hire health experts and indicate explicitly their health role. The main purpose of this study is to use such websites to build classification models that can be used to predict roles of users (medical experts vs patients) on websites while this information is not explicitly indicated. Indeed, according to personal indications we obtained, medical experts confirmed that they usually post messages to help non health professional users online, although their medical expert role may not be indicated.

Health professional and non-health professional posts present some differences that are related, for instance, to the used vocabulary, to the practice of subjectivity markers (emotion and uncertainty) and to the nature and the quality of the produced text (question forms and misspellings). We assume that health professionals may use a different vocabulary by comparison with non-health professionals. Then, lay men may show their emotions more easily than health experts, for example to express their sadness due to their illness: *the pain was so bad*, etc., while health professionals may use more uncertainty words, for example to make an uncertain diagnosis: *you may have an arteritis*, etc. Finally, non-health professionals may ask more questions and make more misspellings. In our work, we propose to consider these differences to evaluate what are the most representative components of a forum post to perform efficiently medical role categorization in online health fora.

Several studies have been proposed for user profiling [5] as well as the studies proposed for the identification of user roles on social media [6], [7], etc. but less works are concerned with the identification of medical roles. Among the works done to automatically categorize the discourse of doctors and the discourse of patients, Chauveau-Thoumelin and Grabar [8] have proposed to use subjectivity markers (emotions and uncertainty markers) in a supervised approach. The discourse of doctors was obtained from scientific papers and clinical reports, while the discourse of patients was obtained from fora posts. The results obtained by the Random Forests algorithm [9] showed high F-scores (from 0.91 to 0.95 for bi-class classification and from 0.88 to 0.90 for tri-class classification). A medical consultations transcriptions corpus has been used by Tanguy et al. [10]. Using linguistic and statistical techniques, the authors have highlighted some characteristics (for example the length of the discourse, the used vocabulary, the gender, the proportion of questions, etc.) that can be interesting for improving our categorization task.

The rest of this paper is organized as follows: Section 2 presents two fora that have been used for evaluating the most significant components of a post. Section 3 introduces our categorization method. Section 4 presents the results obtained and a discussion is proposed in Section 5. Finally, Section 6 concludes and proposes future work.

2 Studied corpora

Two French corpora from two different fora have been collected and cleaned as described below.

2.1 Data collection

Posts from two French websites have been collected.

AlloDocteurs. *AlloDocteur* is a French health forum with more than 16,000 posts[1] covering a large number of topics related to health issues like potentially dangerous medicines, alcoholism, diseases, pregnancy, and sexuality. The forum contains two categories of users: health professional users and non-health professional users. The health professional category may include professional physicians or medical students. Even if their number is limited (16 health professional users are indicated to participate in the forum discussions), their participation in the forum exchanges is important. Indeed, they posted more than 3,000 posts among the 16,000 collected.

MaSanteNet. *MaSanteNet* is an online 'ask the doctor service' subject to charges, that allows users to ask 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. All the questions published on the website have answers. More than 12,000 posts[2] have been collected from this website equitably divided between patient questions and doctor answers.

2.2 Data cleaning

Once the two corpora collected, a cleaning step has been applied in order to improve their quality. First, all posts containing quotes have been filtered out. Indeed, some health professionals repeat the questions before answering them, which may introduce patient statements into health professional posts. Furthermore, all pieces of texts such as author signatures and date of the last modification have been deleted. Finally, posts with less than 10 words have been considered as irrelevant and therefore removed.

2.3 Data preparation

After this cleaning step, we obtained two datasets with more or less balanced data from health professional posts and non-health professional posts.

Table 1. The number of words and the number of posts in the two datasets

	AlloDocteurs		**MaSanteNet**	
	Health professionals	Non health professionals	Health professionals	Non health professionals
Number of words	147,419	222,463	233,565	452,453
Number of posts	2,193	2,179	5,876	6,136
Mean words/posts	67	102	40	74

[1] www.allodocteurs.fr/forum-rubrique.asp [collected on: 19-11-2013]
[2] www.masantenet.com/questions.php [collected on: 18-02-2014]

Table 1 shows that the first corpus has fewer posts than the second (about 4,400 posts from AlloDocteurs and about 12,000 posts from MaSanteNet) but more words per post (on average 85 words per post of AlloDocteurs and 57 words per post of MaSanteNet). It also shows that in both datasets non-health professional posts are longer than health professional posts.

3 Methods

The proposed and implemented method consists in three main steps: annotation, pre-processing and classification.

3.1 Annotation

The Ogmios platform [11] was used to perform the following annotations:

Medical concepts. Terms belonging to three semantic types (diseases, treatments and procedures) have been detected as medical concepts using the following medical terminologies and classifications:

- The Systematized Nomenclature of Human and Veterinary Medicine[3]
- The Thériaque database[4]
- The Unified Medical Language System[5]
- The list of authorized medication that can be marketed in France.

Two lists of all medical terms detected in each corpus have been extracted for a later use.

Emotions. A French emotion lexicon [12], containing about 1,200 words, was used to annotate adjectives, verbs and nouns conveying emotions (joy, sadness, anger, fear, surprise, etc.). In addition to this lexicon, some non-lexical expressions of emotions, such as repeated letters, repeated punctuation signs, smileys, slang and capital letters, have been detected and annotated with specifically designed regular expressions.

Uncertainty. A set of 101 uncertainty words, built in previous study [7], has been used to annotate verbs, nouns, adjectives and even adverbs conveying uncertainty meaning in our corpus.

3.2 Pre-processing

As observed by Balahur [13], fora posts have several linguistic peculiarities that may influence the classification performance. For this reason the following pre-processing steps have been applied:

---

3 www.ihtsdo.org/snomed-ct [last access: 06-05-2014]
4 www.theriaque.org [last access: 06-05-2014]
5 www.nlm.nih.gov/research/umls [last access: 06-05-2014]

Slang replacement. Some expressions are frequently used in Social Media ("*lol*"). They have been replaced by the corresponding standard text ("*lot of laugh*").

Replacement of user tags. All user tags have been identified in our corpora and replaced by the word "Tag" (for example *"@Laurie…"* becomes "*Tag Laurie …*").

Hyperlinks and email addresses. All the hypertext links have been replaced by the word "*link*" and all the email addresses have been replaced by the word "*mail*".

Health pseudonyms. The health professional pseudonyms, previously extracted from each website, are used to replace these pseudonyms in posts by the word "*fdoctor*". Similarly, pseudonyms of non-health professionals have been extracted and used for their replacement by the word "*fpatient*".

Lowercasing and spelling correction. All words have been lowercased and processed with the spell checker Aspell[6]. The default Aspell French dictionary was expanded with medical words extracted from our corpora during the annotation step. The number of misspellings has been computed for each post and used as attribute for the classification.

3.3 Classification

Supervised classifications to categorize health professional and non-health professional posts have been done as follows.

Descriptors used. In order to detect the most discriminative features for our classification task, the number of occurrences of medical concepts, emotions, uncertainty markers, misspellings and question marks have been calculated in both health and non-health professional posts for the two websites processed

Table 2. The number of occurrences of each feature group in both health and non-health professional posts for the two websites

	AlloDocteurs		**MaSanteNet**	
	Health professionals	Non-health professionals	Health professionals	Non-health professionals
Medical concepts	8,924	8,888	21,690	22,921
Emotions (EM)	554	2,137	865	2,962
Uncertainty markers (UM)	5,561	3,871	8,449	7,356
Misspellings (MI)	3,828	12,921	11,529	22,137
Question marks (QM)	560	2,594	509	16,991

From Table 2, we can notice that medical words are used massively by both health and non-health professionals and that there is no significant difference between

[6] www.aspell.net [last access: 06-05-2014]

the two categories of users. Nevertheless, the other descriptors indicate that there is difference between these two kinds of users. Non-health professionals express their emotions more frequently than health professionals. Uncertainty markers are slightly more frequent in health professional posts. And as expected, there are also more misspellings and question marks in non-health professional posts.

According to these observations, emotions, uncertainty markers, misspellings and question marks have been chosen as descriptors in our classification task. For each feature, 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.

In addition to the four features presented before, word ngrams have been considered. The following process has been applied to each corpus: First, all unigrams (words) and bigrams (two words sequences) that appear at least two times are extracted. Then, the number of occurrences of each considered ngram is computed for every post. This number is also normalized by the corresponding post length (number of words) and weighted by its *tf-idf* score (term frequency * inverse document frequency) [14]. Finally, ngrams obtained from the first corpus have been also used on the second corpus and those obtained from the second have been used on the first, which allowed us to test models learned on posts provided by one corpus with the posts from the other corpus. All these treatments were performed with the *"StringToWordVector"* filter from the *Weka* platform [15].

Feature selection. A feature selection step has been applied to select the most discriminant features: those that frequently appear in one category of posts but not in the other one. Therefore, the selected features should characterize one category of users as compared to the other category. Another filter algorithm from the *Weka* platform, named *"InfoGainAttributeEval",* has been used to perform the selection. The gain of each attribute on the classification task has been computed and features that have negative gain (i.e. those that don't improve the classification) have been removed. Table 3 indicates most discriminant ngrams (those that had the best gain scores) for each category in the two datasets[7].

Table 3. Most discriminant unigrams and bigrams for the two categories of users in AlloDocteurs and in MaSanteNet

	AlloDocteurs		**MaSanteNet**	
	Non-health professionals	Health professionals	Non-health professionals	Health professionals
Unigrams (U)	*I, am, thanks, me, have, my*	*Cordially, hello, you, can*	*Am, I, thanks, hello, my*	*Fpatient, must, good, cordially, have*
Bigrams (B)	*I am, I have, thanks, my, that I*	*Your doctor, cordially, fdoctor, can you*	*I am, I have, thanks, that I, is it*	*Cordially, you must, you have, fpatient you*

[7] For readability reasons, ngrams have been translated from French to English.

Table 3 shows that each group uses a specific vocabulary, which is almost the same in both websites. Non-health professionals use the first person singular pronouns (*I, my*) while health professionals use the second person pronouns (*you, your*), which makes sense because the subject of the talk is often the patient and his illness. Besides, non-health professionals show their acknowledgment (*thanks*) while health professionals prefer using a more formal discourse (*cordially*).

Evaluation. Four classification algorithms implemented in *Weka* have been used to test our approach: SVM SMO [16], Naive Bayes [17], Random Forest [9], JRip [18]. For each algorithm, Weighted F-scores are computed with different combinations of features. F-score measures the accuracy of a class; it combines both precision and recall. Usually, it is computed as the harmonic mean of the precision and the recall of the class. Weighted F-score is the mean of all class F-scores weighted by the proportion of elements in each class.

4 Results

Four experiments have been tested: (1) 10-fold cross validation [19] on AlloDocteurs, (2) 10-fold cross validation on MaSanteNet, (3) AlloDocteurs as train set and MaSanteNet as test set and finally (4) MaSanteNet as train set and AlloDocteurs as test set.

4.1 10-fold cross validation on AlloDocteurs

Table 4. Weighted F-scores obtained with 10-fold cross validation on AlloDocteurs.

Feature group	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

4.2 10-fold cross validation on MaSanteNet

Table 5. Weighted F-scores obtained with 10-fold cross validation on MaSanteNet.

Feature group	Number of features	SVM SMO	Naive Bayes	Random Forest	JRip
U	3,096	**1.000**	0.935	0.999	**1.000**
U+B	4,567	**1.000**	0.949	**1.000**	**1.000**
EM	1	0.503	0.495	0.542	0.558
UM	1	0.678	0.653	0.690	0.680
MI	1	0.438	0.648	0.739	0.686
QM	1	0.748	0.715	0.773	0.773
EM+UM+MI+QM	4	0.761	0.741	0.858	0.851
U+EM+UM+MI+QM	3,100	**1.000**	0.942	0.999	**1.000**
U+B+EM+UM+MI+QM	4,571	**1.000**	0.953	**1.000**	0.999

4.3 AlloDocteurs as train set and MaSanteNet as test set

Table 6. Weighted F-scores obtained by considering AlloDocteurs as train set and MaSanteNet as test set

Feature group	Number of features	SVM SMO	Naive Bayes	Random Forest	JRip
U	1,120	0.948	0.862	0.938	0.960
U+B	2,160	0.940	0.914	0.938	**0.970**
EM	1	0.558	0.460	0.504	0.558
UM	1	0.679	0.665	0.654	0.681
MI	1	0.436	0.453	0.44	0.371
QM	1	0.773	0.608	0.720	0.773
EM+UM+MI+QM	4	0.677	0.605	0.679	0.705
U+EM+UM+MI+QM	1,124	0.930	0.866	0.946	0.975
U+B+EM+UM+MI+QM	2,164	0.954	0.915	**0.970**	0.961

4.4 MaSanteNet as train set and AlloDocteurs as test set

Table 7. Weighted F-scores obtained by considering MaSanteNet as train set and AlloDocteurs as test set

Feature group	Number of features	SVM SMO	Naive Bayes	Random Forest	JRip
U	3,096	0.559	0.816	0.615	0.334
U+B	4,567	0.421	**0.841**	0.599	0.335
EM	1	0.610	0.555	0.579	0.610
UM	1	0.681	0.656	0.669	0.681
MI	1	0.313	0.438	0.490	0.440
QM	1	0.653	0.584	0.650	0.653
EM+UM+MI+QM	4	0.685	0.645	0.641	0.595
U+EM+UM+MI+ QM	3,100	0.582	0.821	0.555	0.334
U+B+EM+UM+ MI+QM	4,571	0.434	**0.841**	0.560	0.335

5 Discussion

Globally, the cross validations on both websites processed shows good results. First, the use of ngrams shows high F-scores (between 0.865 and 0.938 obtained on *Allo-Docteurs* and between 0.935 and 1 obtained on *MaSanteNet*) comparing to the use of emotions, uncertainty markers, misspellings and question marks which shows low and medium F-scores (between 0.516 and 0.751 obtained on *AlloDocteurs* and between 0.438 and 0.858 obtained on *MaSanteNet*). The combination of ngrams with the rest of the features increases slightly the classification performances (between 0.866 and 0.94 obtained on *AlloDocteurs* and between 0.942 and 1 obtained on *MaSanteNet*). This increase is so small (between 0.001 and 0.008) that it tends to be statistically insignificant.

The models learned on *AlloDocteurs* and tested on *MaSanteNet* shows similar results. Ngrams show high F-scores (between 0.862 and 0.97) while emotions, uncertainty markers, misspellings and question marks show low F-scores (between 0.371 and 0.773). The combination of all features doesn't improve the classification performances or improves them very little (F-scores obtained by considering all the features are between 0.866 and 0.97). These results tend to confirm the hypothesis according to which the models learned on one website can be efficiently used on other websites.

The models learned on *MaSanteNet* and tested on *AlloDocteurs* gives the worst results. Ngrams show low and medium F-scores if we do not consider Naive Bayes (between 0.334 and 0.615), but high F-scores using Naïve Bayes (between 0.816 and 0.841). Similarly, the results obtained with emotions, uncertainty markers, misspellings and question marks show low F-scores if we do not consider Naive Bayes (between 0.371 and 0.685), low and medium using it (between 0.438 and 0.821). The

combination of all features doesn't improve the classification performances neither: the F-scores obtained by considering all the features are between 0.334 and 0.841.

The difference between the two last experiments can be explained by the fact that the first website is a forum, where 16 health professionals post messages in many threads. This makes the discourse of medical users more extensive and diversified, so that models learned on this website may cover the topics and medical discourse observed on the other website: these models have more chances to identify medical professional posts on other websites. On the other hand, the second website is an "Ask the doctor" service where only two medical experts answers the questions. Moreover, their answers are constrained and normalized, as they always answer in the same way. This makes the discourse of medical experts extremely specific to this website: for this reason it appears to be less adapted to learn language models that can be used on other data.

6 Conclusion and perspectives

In this paper, we presented a supervised method that allows categorizing posts made by health professionals and those made by non-health professionals. Several features have been tested to perform the categorization: ngrams, emotions, uncertainty markers, misspellings and question marks. The experiments indicate that ngrams are the most efficient. The results indicate that models leaned on appropriate websites may be used efficiently on other websites. Moreover, models learned on more general and varied websites (like fora) where many health professionals are involved provide better data for the learning step.

The results obtained are very encouraging but they can be improved. First, the filter used in the feature selection step computes the gain of each feature independently from the other features and doesn't treat the case of redundancy between the features, which may influence the results of some classification algorithms (such as: Naive Bayes) which assume that the features are independent. Furthermore, we used a small French emotion lexicon (containing about 1,200 words). A more comprehensive emotion lexicon [20] is now under construction; we are translating and expanding to synonyms the English emotion lexicon NRC [21] with the help of a professional translator. Up to now, the new emotion lexicon contains more than 20,000 emotion words and we expect it will become even more extensive. The spell checking can also be improved either by considering grammar rules or by a more stringent human supervision of the correction process which also implies that we may obtain a more correct number of misspellings.

The question of detecting trustier and more precise posts in online health fora may be addressed with different methods. Indeed, trust models tested on other social media may be applied either by looking at the structure of the threads (computing scores based on the number of quotes, the number of likes, the number of posts between each post and its replies, etc.) [22], [23] or by inferring these information from the text [24]. In addition to these models, we plan to include the emotional reaction of users to a specific post while computing the trust scores (for example posts arousing the anger of the users).

Finally, we are interested in other applications of Natural Language Processing and Text Mining on online health fora. Currently, we are working on a recommendation system that suggests appropriate topics where the user should post his message. We exploit the content of the posts (title and body), the gender and the age of users, etc. An additional descriptor may be related to the topics where the user has already posted the messages: we assume it may improve the automatic system because the previous preferences of the users may be indicative of his current interests.

7 Acknowledgement

This paper is based on studies supported by the "Maison des Sciences de l'Homme de Montpellier" (MSH-M) within the framework of the French project "Patient's mind"[8].

8 Bibliography

[1] W. Himmel, U. Reincke, and H. W. Michelmann, "Text Mining and Natural Language Processing Approaches for Automatic Categorization of Lay Requests to Web-Based Expert Forums," *J. Med. Internet Res.*, vol. 11, no. 3, pp. 1–1, Jul. 2009.

[2] J. Huh, M. Yetisgen-Yildiz, and W. Pratt, "Text classification for assisting moderators in online health communities," *J. Biomed. Inform.*, vol. 46, no. 6, pp. 998–1005, Dec. 2013.

[3] S. Melzi, A. Abdaoui, J. Azé, S. Bringay, P. Poncelet, and F. Galtier, "Patient's Rationale: Patient Knowledge Retrieval From Health Forums," in *eTELEMED 2014, The Sixth International Conference on eHealth, Telemedicine, and Social Medicine*, 2014, pp. 140–145.

[4] S. Bringay, E. Kergosien, P. Pompidor, and P. Poncelet, "Identifying the Targets of the Emotions Expressed in Health Forums," in *Computational Linguistics and Intelligent Text Processing*, A. Gelbukh, Ed. Springer Berlin Heidelberg, 2014, pp. 85–97.

[5] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches, "Overview of the author profiling task at PAN 2013," *Noteb. Pap. CLEF*, pp. 23–26, 2013.

[6] M. Bouguessa, B. Dumoulin, and S. Wang, "Identifying Authoritative Actors in Question-answering Forums: The Case of Yahoo! Answers," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2008, pp. 866–874.

[7] D. Fisher, M. Smith, and H. T. Welser, "You Are Who You Talk To: Detecting Roles in Usenet Newsgroups," in *Proceedings of the 39th Annual Hawaii International Conference on System Sciences, 2006. HICSS '06*, 2006, vol. 3, p. 59b–59b.

[8] P. C. Thoumelin and N. Grabar, "La subjectivité dans le discours médical : sur les traces de l' incertitude et des émotions," *Rev. Nouv. Technol. Inf.*, vol. Extraction et Gestion des Connaissances, RNTI-E-26, pp. 455–466, 2014.

[9] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[8] https://www.lirmm.fr/patient-mind/pmwiki/pmwiki.php?n=Site.Accueil

[10]L. Tanguy, C. Fabre, L.-M. Ho-Dac, and J. Rebeyrolle, “Caractérisation des échanges entre patients et médecins : approche outillée d’un corpus de consultations médicales,” *Corpus*, no. 10, pp. 137–154, Jun. 2012.

[11]T. Hamon and A. Nazarenko, “Le développement d’une plate-forme pour l’annotation spécialisée de documents Web : retour d’expérience,” *Trait. Autom. Lang.*, vol. 49, no. 2, pp. 127–154, 2008.

[12]M. Augustyn, S. B. Hamou, G. Bloquet, V. Goossens, M. Loiseau, and F. Rinck, “Lexique des affects : constitution de ressources pédagogiques numériques.,” in *Autour du langage et des langues : perspective pluridisciplinaire, Sélection d’articles du Colloque International des étudiants-chercheurs en didactique des langues et linguistique.*, 2008.

[13]A. Balahur, “Sentiment Analysis in Social Media Texts,” in *4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Atlanta, Georgia, 2013, pp. 120–128.

[14]G. Salton, “Developments in Automatic Text Retrieval,” *Science*, vol. 253, no. 5023, pp. 974–980, Aug. 1991.

[15]M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA Data Mining Software: An Update,” *SIGKDD Explor Newsl*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[16]J. C. Platt, “Advances in Kernel Methods,” B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA, USA: MIT Press, 1999, pp. 185–208.

[17]G. H. John and P. Langley, “Estimating Continuous Distributions in Bayesian Classifiers,” in *Eleventh Conference on Uncertainty in Artificial Intelligence*, San Mateo, 1995, pp. 338–345.

[18]W. W. Cohen, “Fast Effective Rule Induction,” in *Twelfth International Conference on Machine Learning*, 1995, pp. 115–123.

[19]“Cross-validation and selection of priors,” *Statistical Modeling, Causal Inference, and Social Science*. [Online]. Available: http://andrewgelman.com/2006/03/24/crossvalidation_2/. [Accessed: 07-May-2014].

[20]“Lexique des sentiments et des émotions français.” .

[21]S. M. Mohammad and P. D. Turney, “Emotions Evoked by Common Words and Phrases : Using Mechanical Turk to Create an Emotion Lexicon,” in *Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Stroudsburg, PA, USA, 2010, pp. 26–34.

[22]F. Skopik, H.-L. Truong, and S. Dustdar, “Trust and Reputation Mining in Professional Virtual Communities,” in *Web Engineering*, M. Gaedke, M. Grossniklaus, and O. Díaz, Eds. Springer Berlin Heidelberg, 2009, pp. 76–90.

[23]N. Wanas, M. El-Saban, H. Ashour, and W. Ammar, “Automatic Scoring of Online Discussion Posts,” in *Proceedings of the 2Nd ACM Workshop on Information Credibility on the Web*, New York, NY, USA, 2008, pp. 19–26.

[24]D. Feng, E. Shaw, J. Kim, and E. Hovy, “Learning to Detect Conversation Focus of Threaded Discussions,” in *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, Stroudsburg, PA, USA, 2006, pp. 208–215.