Analyse des messages des patients et des médecins dans les fora de santé

Amine Abdaoui1, Jérôme Azé1, Sandra Bringay1, Pascal Poncelet1 et Natalia Grabar2

[1] LIRMM UM2 CNRS, UMR 5506, 161 Rue Ada, 34095 Montpellier.
[2] STL UMR 8163 CNRS, Université Lille 3 et Lille 1

Résumé : Les messages des fora de santé diffèrent selon la nature socio-professionnelle de leurs auteurs. A titre d'exemple les patients, qui représentent la principale catégorie sur les fora de santé en ligne, expriment souvent leurs émotions (par exemple leur tristesse causée par la maladie), alors que les médecins expriment souvent leur incertitude (généralement pour établir un diagnostic incertain). Dans un but de recherche d'information, il peut être intéressant de distinguer automatiquement les messages postés par les patients et ceux postés par les médecins. Dans ce travail, nous proposons une approche de fouille de texte qui considère les marqueurs de subjectivité (émotions, incertitude et fautes d'orthographe) associés aux classiques n-grammes afin de distinguer les messages postés par les patients et ceux postés par les médecins.

Mots-clés : fouille de texte, classification supervisée, analyse des fora de santé.

1 Introduction

Les fora de santé en ligne sont de plus en plus consultés par les patients à la recherche de soutien et d'informations se rapportant à leur santé (Huh et al., 2013). Cependant, les auteurs dans ces fora ne se limitent pas aux patients, car de nombreux professionnels de santé participent aussi aux discussions. Pourtant, ces deux catégories d'utilisateurs présentent des différences qui concernent, entre autres, les marqueurs de subjectivité tels que : les émotions (les patients expriment leurs émotions plus facilement que les professionnels de santé), l'incertitude (les professionnels de santé ont tendance à utiliser les mots d'incertitude plus que les patients) et enfin la qualité du texte (les professionnels de santé font moins de fautes d'orthographes que les patients), etc. Dans un but de recherche d'information, il peut être intéressant d'utiliser ces spécificités pour distinguer automatiquement les messages produits par les patients et ceux produits par les professionnels de santé, par exemple un utilisateur recherchant de l'information médicale aura plus de confiance aux messages postés par les professionnels de santé.

Les marqueurs de subjectivité ont déjà été utilisés (Chauveau et Grabar, 2014) pour distinguer les discours des médecins (articles scientifiques et rapports cliniques) et des patients (messages de forums de santé). Dans ce travail, nous proposons une approche qui se base sur les n-grammes mais qui prend aussi en considération la subjectivité via trois marqueurs qui sont : les marqueurs d'émotions (émoticônes, caractères répétés, mots d'émotions, etc.), incertitude (possible, peut-être, etc.) et les fautes d'orthographes.

La suite de cet article est organisée de la manière suivante : la section 2 présente notre corpus d'étude. La section 3 décrit la méthode proposée pour distinguer les messages créés par les professionnels de santé (les médecins) et ceux créés par les non-professionnels de santé (les patients). Les sections 4 et 5 présentent et analysent les résultats obtenus. Enfin, la section 6 conclut et donne nos perspectives futures.

2 Corpus

Notre corpus a été obtenu en deux étapes.

2.1 Récupération des données

16,609 messages ont été récupérés à partir du forum de santé français *Allodocteurs*[1]. Ce forum couvre une large gamme de sujets tels que le cancer, la nutrition, les médicaments potentiellement dangereux, la maternité, etc. Il contient deux catégories d'utilisateurs : les professionnels de santé et les non-professionnels. En général, la première catégorie d'utilisateurs répond à des questions posées par la deuxième catégorie. Les professionnels de santé peuvent être des médecins généralistes ou spécialistes mais ils peuvent aussi être des étudiants en médecine. Bien que leur nombre soit restreint (environs 16 utilisateurs), leur participation aux échanges sur le forum est considérable : ils ont postés 3,050 messages sur les 16,609 collectés.

2.2 Nettoyage et préparation

Afin d'améliorer la qualité de notre corpus, les messages contenant des citations ont été éliminés. En effet, certains professionnels de santé citent les questions avant d'y répondre, ce qui peut introduire des textes rédigés par les patients dans les messages des professionnels de santé. Les messages très courts (tel que « Merci », « oui », etc.) ont été supprimés également. Deux jeux de données distincts et équilibrés ont été générés: un jeu d'apprentissage contenant 4,000 messages (2,000 par catégorie) et un jeu de test contenant 450 messages (225 par catégorie).

3 Méthodes

La méthode proposée se compose de trois étapes (l'annotation, le prétraitement et la classification).

3.1 Annotation

Un protocole d'annotation a été appliqué dans le but d'avoir trois types de marqueurs :

3.1.1 Les mots médicaux

Les maladies, les traitements et les procédures médicales ont été détectés en utilisant les ressources suivantes :

- Le système des terminologies biomédicales UMLS[2]
- La base de médicaments Thériaque[3]
- La terminologie médicale SNOMED international[4]
- La base UCD couvrant les médicaments qui disposent d'une autorisation de mise sur le marché et sont commercialisés en France.

Une liste de tous les mots médicaux détectés a été extraite pour une utilisation ultérieure.

---

[1] www.allodocteurs.fr/forum-rubrique.asp [collecté: 19-11-2013]

[2] www.nlm.nih.gov/research/umls [dernier accès: 23-01-2014]

[3] www.theriaque.org [dernier accès: 23-01-2014]

[4] www.ihtsdo.org/snomed-ct [dernier accès: 23-01-2014]

3.1.2 Les émotions

Un lexique d'émotion pour le français (Augustyn et al., 2008), contenant environs 1,200 mots, a été utilisé pour annoter les adjectifs, les verbes et les noms porteurs d'émotions. En plus de ce lexique, certaines structures non-lexicales ont été identifiées comme marqueurs d'émotions : signes de ponctuation et lettres répétées, émoticônes, argots et majuscules.

3.1.3 Les marqueurs d'incertitudes

Une liste de 101 marqueurs d'incertitude a été utilisée pour annoter les verbes, noms, adjectifs et adverbes susceptibles d'être porteurs d'incertitude.

Le tableau 1 ci-dessous indique le nombre de mots de chaque type de marqueurs de subjectivité, dans les messages des professionnels de santé et des non-professionnels de santé ainsi que le nombre de messages où ces types de marqueurs apparaissent au moins une fois.

TABLE 1 – *Fréquence de chaque type de marqueurs de subjectivité dans les messages des patients et des médecins ainsi que le nombre de messages où ils apparaissent au moins une fois*

Types de marqueurs	Messages postés par les patients		Messages postés par les médecins	
	Nombre de mots	Nombre de messages	Nombre de mots	Nombre de messages
Mots médicaux	7,942	1,643	8,042	1,725
Émotions	1,589	815	525	385
Incertitude	3,423	1,342	4,988	1,675

Le tableau 1 montre que les mots médicaux sont répartis de manière plus ou moins équitable entre les messages des patients et les messages des médecins. Pour cette raison, nous ne les avons pas considérés lors de l'étape de classification. Par ailleurs, les émotions et les marqueurs d'incertitude semblent être des descripteurs intéressants pour la classification. En effet, les patients utilisent beaucoup de marqueurs d'émotion alors que les médecins utilisent beaucoup de marqueurs d'incertitude.

3.2 Prétraitements

Comme mentionné dans la littérature (Balahur, 2008), les messages des fora ont plusieurs particularités linguistiques qui peuvent influencer les performances de la classification. Pour cette raison nous avons appliqué les prétraitements suivants :

- **Argots :** Certaines expressions sont fréquemment utilisées sur les réseaux sociaux ('mdr', etc.). Ces expressions ont été remplacées par le texte correspondant ('mdr' devient 'mort de rire').
- **Tags d'utilisateurs :** Les tags des utilisateurs ont été identifiés et remplacés par le mot 'Tag', par exemple '@Laurie…' devient 'Tag Laurie…'.
- **Liens hypertextes et adresses mail :** Tous les liens hypertextes ont été remplacés par le mot 'lien' et toutes les adresses mails ont été remplacées par le mot 'mail'.
- **Pseudonymes :** une liste de tous les pseudonymes des utilisateurs du forum a été extraite. Cette liste a été utilisée pour remplacer tous les pseudonymes des médecins présents dans les messages par le mot 'Fmédecin' et tous les pseudonymes des patients par le mot 'Fpatient'.
- **Mise en minuscule et correction d'orthographe :** Tous les mots du corpus ont été mis en minuscules et corrigés avec le correcteur orthographique « Aspell »[5]. Le dictionnaire par défaut de « Aspell » a été étendu avec les mots médicaux extraits

[5] www.aspell.net [dernier accès : 20-01-2014]

dans la partie annotation. Le nombre de mots jugés erronés par « Aspell » a été calculé pour chaque message et utilisé comme attribut pour la classification.

Tous ces prétraitements ont été faits automatiquement mais supervisés par un humain.

3.2 Sélection d'attributs et Classification

Une étape de sélection d'attributs a été appliquée pour les n-grammes pour en choisir les plus discriminants. Cette étape a fait ressortir des uni-grammes et les bi-grammes qui sont à la fois très utilisés par une catégorie et très peu utilisés par l'autre. Parmi les uni-grammes et les bi-grammes décrivant la classe des patients nous avons trouvé: merci, je, j'ai, me, je suis, ma, etc. Parmi les uni-grammes et les bi-grammes décrivant la classe des médecins nous avons trouvé: cordialement, vous, pouvez-vous, avez-vous, etc. Ceci s'explique par le fait que le sujet de l'échange est le patient et sa maladie donc ce dernier s'exprime avec je, me, ma, etc. et le médecin par vous, votre, etc.

Les descripteurs listés dans le tableau 2 ont été utilisés pour distinguer les messages des patients et des médecins. Chaque descripteur peut être représenté par un ou plusieurs attributs. Pour chaque attribut, nous calculons une fréquence normalisée par le nombre de mots du message.

TABLE 2 – *Les descripteurs utilisés dans notre modèle de classification et leur nombre d'attributs*

Descripteurs	Nombre d'attributs
Uni-grammes (U)	1,005
Uni-grammes et bi-grammes (U+B)	2,024
Emotions (Emo)	1
Marqueur d'incertitude (Inc)	1
Fautes d'orthographe (Ort)	1

4 Résultats

Les résultats fournis par Weka (Hall et al., 2009) en termes de F-mesure sont présentés dans le tableau 3 ci-dessous pour quatre algorithmes de classification, ces résultats sont obtenus en utilisant le jeu d'apprentissage (4,000 messages) pour apprendre à chaque fois les modèles de classification et le jeu de test (450 messages). Nous avons également effectué une validation croisée à 10 échantillons sur l'ensemble des messages du corpus. Les résultats obtenus étant très similaires, nous ne présentons ici que ceux de la première expérience.

TABLE 3 – *Les résultats en termes de F-mesures en apprenant les modèles sur le jeu d'apprentissage et en testant sur le jeu de test du forum Allodocteurs*

Descripteurs	SVM SMO	RandomForest	NaivesBayes	JRip
U	**0.940**	0.902	0.870	0.936
U+B	**0.931**	0.909	0.869	0.902
Emo	0.653	0.577	0.522	**0.653**
Inc	0.704	0.673	0.683	**0.705**
Ort	0.633	0.634	0.596	**0.669**
Emo+Inc+Ort	0.724	0.72	0.691	**0.762**
U+Emo+Inc+Ort	**0.947**	0.916	0.872	0.882
U+B+Emo+Inc+Ort	**0.942**	0.915	0.869	0.900

Afin de tester si les modèles de classification appris sur un forum fonctionnent aussi bien sur d'autres fora, nous avons récupéré 12,000 messages à partir du site français MaSanteNet[6] dont le principe est similaire à Allodocteurs : les patients posent des questions et les médecins y répondent. Les messages récupérés sont aussi équilibrés entre les deux classes (6,000 questions et 6,000 réponses). Les mêmes étapes de nettoyage, annotations et prétraitements ont été appliquées sur ce nouveau corpus. Nous avons ensuite utilisé les messages du premier forum (4,450) pour construire les modèles de classification et les messages du nouveau forum pour les tester (12,000). Les F-mesures obtenues sont très élevées (tableau 4 ci-dessous).

TABLE 4 – *Les résultats en termes de F-mesures en apprenant les modèles sur Allodocteurs mais en les testant sur MaSanteNet*

Descripteurs	SVM SMO	RandomForest	NaivesBayes	JRip
U	0.801	0.948	0.908	**0.978**
U+B	0.908	**0.972**	0.912	0.953
Emo	0.452	0.526	0.461	**0.566**
Inc	0.568	**0.592**	0.582	0.589
Ort	0.573	0.563	0.559	**0.581**
Emo+Inc+Ort	0.572	0.581	0.597	**0.609**
U+Emo+Inc+Ort	0.743	0.944	0.913	**0.962**
U+B+Emo+Inc+Ort	0.908	**0.976**	0.914	0.973

5 Discussion

Globalement, les résultats obtenus sont très encourageants. Comme prévu, les n-grammes donnent des F-mesures très élevées, ce qui confirme l'hypothèse que les médecins et les patients utilisent un vocabulaire très différent. Les marqueurs de subjectivité (émotions, incertitude et fautes d'orthographe) donnent des F-mesures inférieures, ce qui signifie que les marqueurs de subjectivité ne suffisent pas à eux seuls pour distinguer efficacement les messages des patients et les messages des médecins mais qu'ils peuvent apporter un plus en combinaison avec d'autres marqueurs. Enfin l'utilisation des n-grammes avec les marqueurs de subjectivité donne les meilleurs résultats.

Il apparaît aussi que les modèles de classification appris sur le forum Allodocteurs fonctionnent aussi bien sur MaSanteNet (on obtient même des F-mesures qui frôlent les 98%). Ces résultats sont très encourageants pour la distinction des deux classes sur des sites où la catégorie des utilisateurs n'est pas visible.

En termes d'efficacité des algorithmes, SVM (Schölkopf et al., 1999) semble donner les meilleurs résultats lorsque les tests sont effectués sur le même forum mais en même temps les pires lorsque les tests sont effectués sur un autre forum. Dans ce dernier cas, ce sont RandomForest (Breiman, 2001) et JRip (Cohen et Singer, 1999) qui sont les plus efficaces.

6 Conclusion

Dans cet article nous avons présenté une méthode pour distinguer les messages des médecins et des patients dans les fora de santé de ligne. Notre méthode se base sur les n-grammes et les marqueurs de subjectivité (émotion, incertitude et fautes d'orthographe). Elle se compose de plusieurs étapes de nettoyage, annotation, prétraitements, sélection d'attributs et enfin classification. Les résultats obtenus sont très encourageants et prouvent que les modèles appris sur un forum où les médecins et les patients sont donnés explicitement peuvent être utilisés efficacement pour distinguer les messages des patients et des médecins sur d'autres forums.

[6] www.masantenet.com [collecté: 18/02/2014]

Ces résultats peuvent encore être améliorés. D'abord, nous avons utilisé un petit lexique d'émotions contenant 1,200 mots. Un autre lexique est en cours de construction contenant plus de 20,000 mots : nous sommes en train de traduire et d'étendre avec des synonymes le lexique d'émotion NRC (Mohammad et Turney, 2010) avec l'aide d'une traductrice professionnelle. La correction d'orthographe peut être aussi améliorée, déjà, en considérant les règles de grammaire, et ensuite, en respectant la casse car nous avons remarqué que la mise en minuscule, qui nous a simplifié la correction, diminue légèrement la qualité des résultats. Enfin, nous prévoyons d'utiliser les marqueurs de subjectivité pour d'autres buts tels que l'identification de thématiques. Nous pouvons par exemple supposer que certains sujets suscitent plus de subjectivité que d'autres. Pour cela nous avons constitué un autre corpus contenant d'autres forums qui traitent de plusieurs sujets relatifs à la santé.

Références

Augustyn M., Ben Hamou S., Bloquet G., Goossens V., Loiseau M. & Rynck F. (2008). Constitution de ressources pédagogiques numériques : le lexique des affects. Presses Universitaires de Grenoble, p. 407–414.

Balahur A. (2013). Sentiment analysis in social media texts. 4th workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, p. 120-128.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Cohen, W. W., & Singer, Y. (1999, July). A simple, fast, and effective rule learner. In Proceedings of the National Conference on Artificial Intelligence (pp. 335-342). John Wiley & Sons Ltd.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H. (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1.

Huh J., Yetisgen-Yildiz M. & Pratt W (2013). Text classification for assisting moderators in online health communities. Journal of Biomedical Informatics 46, p. 998-1005.

Mohammad S. & Turney P. D. (2010). Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, US, p. 26-34.

Thoumelin P. C. & Grabar N (2014). La subjectivité dans le discours médical: sur les traces de l'incertitude et des émotions. 14eme conférence sur l'Extraction et la Gestion des Connaissances.

Schölkopf, B., Burges, C. J., & Smola, A. J. (Eds.). (1999). *Advances in kernel methods: support vector learning*. MIT press.