MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)

doi:10.3233/978-1-61499-564-7-137

E-Patient Reputation in Health Forums

Amine Abdaoui[a], Jérôme Azé[a], Sandra Bringay[a] and Pascal Poncelet[a]

[a] *LIRMM UM2 CNRS, UMR 5506, 161 Rue Ada, 34095 Montpellier, France*

Abstract

Online health forums are increasingly used by patients to get information and help related to their health. However, information reliability in these forums is unfortunately not always guaranteed. Obviously, consequences of self-diagnosis may be severe on the patient's health if measures are taken without consulting a doctor. Many works on trust issues related to social media have been proposed, but most of them mainly focus only on the structure part of the social network (number of posts, number of likes, etc.). In the case of online health forums, a lot of trust and distrust is expressed inside the posted messages and cannot be inferred by only considering the structure. In this study, we rather suggest inferring the user's trustworthiness from the replies he receives in the forum. The proposed method is divided into three main steps: First, the recipient(s) of each post must be identified. Next, the trust or distrust expressed in these posts is evaluated. Finally, the user's reputation is computed by aggregating all the posts he received. Conducted experiments using a manually annotated corpus are encouraging.

Keywords:

Trust, Reputation, Social media, Online health forums.

Introduction

Internet is allowing patients to play a more active role in their own health care. Today, more than 46% of patients in 12 different countries use internet for self-diagnosis [1]. The use of Web-derived health information is rapidly increasing and has been termed as the *e-patient revolution* [2]. They have a strong desire to learn, to understand their own symptoms and to have access to medical knowledge. Although accessible and easy to use, the reliability of information on Internet represents a major risk on e-patients health. The consequences of an erroneous self-diagnosis are difficult to estimate if measures are taken without consulting a doctor. Indeed, only 21% of e-patients ask their physician confirmation of information obtained from Internet [1]. According to a recent study conducted by the Health On the Net foundation, 90% of e-patients use search engines to initiate their requests. Most of the returned links propose health forums which are used by more than 50% of e-patients. Thus, these forums are becoming the first source of medical information on the Internet. They are areas of exchange where patients, on condition of anonymity, freely relate their personal experiences and give their views and advices.

It is difficult to prevent e-patients to consult irrelevant or unreliable information on health forums; however it is possible to design tools to highlight the trustworthiness of information as well as the users in it. Many online health forums give a rank to each user, which is usually based only on the number of messages posted since his registration. Actually such ranking does not really give a good estimation of users' trustworthiness. A discussion with moderators of a French forum confirmed this intuition, since they know trusted users who post few messages and untrusted users who posted a lot of messages. In a previous work, we proposed a method to automatically distinguish posts made by health experts from those made by laymen [3]. In this new work, we are interested on the reputation of online health forum users independently from their medical roles. Many definitions of trust and reputation can be found in the literature according to each context [4]–[7]. Here we define the trust that a user *A* gives to a user *B* as: "*the belief of A in the accuracy of the information posted by B*", and the reputation of a user *A* as "*the aggregation of the values of trust given to user A*". Most studies of trust in social networks usually focus on the structure of the website (ratings, number of posts, number of likes, number of quotes, distance between posts, etc.) [8]. However, explicit liking and quoting functionalities are rarely used in the case of online health forums. For example, in the French health forum *CancerDuSein.org* only 2% of posts have explicit quoting. Besides, most users in this forum prefer posting a new reply where they express their agreement or recognition rather than simply pressing the like button.

In this study we suggest to infer the user's reputation from the replies he receives in the forum. Therefore, our method is divided into three main steps. First, links between each post and the person(s) to whom it is addressed, also called the recipient(s), are identified. Some works have already addressed this task [9], [10]. Three types of relationships have been extracted: structural relationships, name relationships and text quotation relationships. In this work, we consider nine kinds of different relationships. After that, posts are evaluated and classified into one of the following classes: trust, distrust and neutral, according to the use of agreement, disagreement and thanking expressions. Finally, the user's reputation is computed by aggregating the trust and distrust expressed in the posts he received. It may computed in the whole forum or in specific topics by aggregating only replies posted in that specific topic. Users' reputations can be used either by moderators in order to investigate users having very bad reputations, or by forum readers in order to have an idea about the authors trustworthiness.

The rest of the paper is organized as follows. Section 2 presents the corpus and describes the used methods. Section 3 presents the obtained results and Section 4 discusses them. Finally, Section 5 concludes and gives our main perspectives.

Materials and Methods

First, the corpus of study is described. Then, the three main steps of our method are presented: identifying the recipient(s) of each post, inferring the trust expressed by each post and computing the reputation of each user.

Corpus of study

CancerDuSein.org is a French health forum specialized in breast cancer. 1,050 threads have been collected which holds 16,961 messages posted by 675 users. It represents all the data that have been posted between October 2011 and November 2013. Some threads have more than 500 posts, which make the use of semi-automatic systems a challenging task. This forum gathers women with breast cancer or their families, who want to exchange their experiences, advices and emotional support. However, the consequences of acting on incorrect advices can be severe. The forum gives a rank to each user based on the number of posts since his registration as described in Table 1. However, an active member on the forum is not necessarily a trusted member and similarly a new member is not necessarily an untrusted one.

Table 1 – Number of posts for each rank

Rank	Number of posts
New member	[0, 20[
Regular member	[20, 40[
Accustomed member	[40, 80[
Active member	>80

Step 1: Finding the recipient(s)

The first step of our method consists in finding the recipient(s) of each post in the forum. In order to construct a network of replies, a rule based heuristic has been developed. Nine rules have been designed and applied chronologically as described below. If a message does not match the first rule, the heuristic will check the second one and so on. The first post in each thread is not considered since it does not answer anybody.

Explicit quoting: *CancerDuSein.org* allows users to explicitly quote another user's post. However, only 2% of posts on the Website contain explicit quoting. These quotes have been detected using the HTML tag <*quote*> and by comparing the content of these tags and the pseudonym of the quoted user with the messages posted before in the same thread. This process allowed us to detect 312 quotes automatically. The rest of quotes (37) have been related manually because the quoted text has been modified or truncated by the user.

Second posts: Messages posted at the second place in each thread have been considered as replying to the first one.

Names and pseudonyms: If a message contains the pseudonym or the name of a user who previously posted a message in the same thread, then this user is considered as the recipient of the message. The pseudonyms have been extracted automatically while the names have been extracted from the signatures and validated manually. The following preprocessings have been applied in order to detect names and pseudonyms inside the text:

- Remove all non-alphabetic characters except spaces (***John Woe 34*** becomes *John Woe*).
- Replace all accented characters by the corresponding non-accented ones (*Jérôme* becomes *Jerome*).
- Lowercasing (*Sandra* becomes *sandra*).

Grouped posts: If a message contains a group marker ("hello everyone", "Hi girls", "Thank you all", etc.) then all the users who previously posted in the same thread are considered as recipients for this post.

Second person pronouns: In French, singular second person pronouns and plural second person pronouns are different. If a singular second person pronoun is used then the recipient is the author of the previous post.

Activator posts: When the activator of the thread (the user who opened the the thread by posting the first message) posts a new message in the same thread, we consider that all the users who posted after his last message are recipients of his new message.

Questions: If the message contains a question then the message is addressed to all the users who previously posted in the same thread.

Answers: If there is a question posted before in the thread, the recipient is the user who posted this question.

Default: If none of the rules mentioned before is satisfied, we consider that the recipient of the message is the activator.

Table 2 presents the number of posts that match each rule.

Table 2 – Number of posts that match each rule

Rule	Number of posts
Explicit quoting	349
Second posts	942
Names and pseudonyms	7,121
Grouped posts	740
Second person pronouns	2,406
Activator posts	1,239
Questions	298
Answers	1,790
Default	772
Total	**15,657**

Step 2: Inferring the trust

The second step consists in classifying each post according to the expressed trust. Posts containing agreement and thanking expressions have been considered as expressing trust to the answered person. Posts containing disagreement expressions have been considered as expressing distrust. The rest of posts have been considered as neutral.

Building the lists of expressions: The expressions of agreement, disagreement and thanking have been built manually based on terms extracted from the LAROUSSE thesaurus [11] and with the help of some specific Websites. All these expressions have been lemmatized in order to detect all the forms of words. Finally, the trust list contains 34 expressions of agreement and thanking while the distrust list contains 15 expressions of disagreement.

Negation: if one of the trust expressions is under the scope of a negation term, it is considered as a distrust expression and vice versa. The scope of a negation term may be two words after, two words before or two words after and two words before according to the nature of the negation term.

Computing the frequencies and classifying the posts: First, all posts have been lowercased, lemmatized using the TreeTagger tool [12] and corrected using the spell checker

Aspell (www.aspell.net). Next, posts are classified as expressing trust, distrut or neutral as follows:

- If a post contains more trust expressions than distrust ones, it is considered as expressing trust.
- If a post contains more distrust expressions than trust ones, it is considered as expressing distrust.
- Otherwise, it is considered as neutral.

Step 3: Computing the reputation

Once the posts addressed to each user are identified and the trust expressed in each post evaluated, we can compute the reputation of each user based on the trust and/or distrust expressed by the replies he received. We suggest computing the difference between the rate of replies expressing trust and the rate of replies expressing distrust. For a user "*u*" the reputation is computed as follows:

$$Reputation(u) = \begin{cases} \frac{NRT(u) - NRD(u)}{NR(u)}, if\ NR(u) \neq 0 \\ 0 \quad\quad\quad\quad\quad\quad , Otherwise \end{cases}$$

Where:

NRT(*u*) is the number of replies expressing trust to user "*u*"
NRD(*u*) is the number of replies expressing distrust to user "*u*"
NR(*u*) is the total number of replies addressed to user "*u*"

Reputation(*u*) belongs to the range [-1, 1]

Results

In order to evaluate the rule based heuristic that finds the recipient(s) of each post and the automatic inference of trust, 2,433 manual annotations have been done. The results obtained using these annotations and those obtained after computing the reputation are described below.

Step 1: Finding the recipient

Two datasets have been used to test our rule based heuristic. The rules have been designed according to the development set (the authors used this dataset to have an idea on the rules that need to be designed). After that, a test set has been used to test our heuristic on unseen threads.

Table 3 – Number of threads, posts and links found by the heuristic in each dataset

Datasets	Number of threads	Number of posts	Number of links
Development set	10	105	152
Test set	10	109	150

Prior-assessment: 15 non-expert annotators (they do not know the designed rules) annotated our two datasets. Each one annotated between 1 and 5 threads so that each thread had 3 annotators. The goal was to find the recipient(s) of each post without knowing the results of our heuristic. Classical measures of agreement are not well adapted, here we simply present the number of links (message, recipient) found by our three annotators at the same time, the number of links found by two out of the three annotators and the number of links found by only one annotator.

Table 4 – Links found by one, two and by the three annotators in each dataset

Datasets	Found by	Number of links	Percentage
Development set	3 annotators	102	53.4%
	2 annotators	54	28.3%
	1 annotator	35	18.3%
	Total	191	100%
Test set	3 annotators	103	60.3%
	2 annotators	44	25.7%
	1 annotator	24	14%
	Total	171	100%

Post-assessment: 3 expert annotators (the authors) annotated the links found by the heuristic in the two datasets. The goal was to validate or not the links found automatically with the possibility of adding a link which was not found by the heuristic. The agreement between the annotators was very good, (the obtained Fleiss' Kappa [13] is **0.89** for the development set and **0.74** for the test set).

Evaluation: Using these annotations, the quality of the developed heuristic has been evaluated. The links obtained automatically have been compared with those obtained from the annotations by considering only those that have been validated by more than two annotators (a majority vote). Table 5 presents the obtained precision, recall and F1-score.

Table 5 – Precision, recall and F1-score of the heuristic obtained on both dataset using prior and post assessments

		P	R	F
Prior assess-ment	Development set	0.70	0.68	0.69
	Test set	0.81	0.83	0.82
Post assess-ment	Development set	0.80	0.84	0.82
	Test set	0.83	1	0.91

Step 2: Inferring the trust

Two new datasets have been used to evaluate the automatic trust inference. Unlike the first step where both datasets had prior and post assessment, here prior-assessment has been done only for the first dataset and post-assessment has been done only for the second one.

Prior-assessment: 3 annotators annotated the trust expressed in 97 messages without knowing the results of the automatic system. The agreement between them was less than the recipient assessment but still acceptable (the obtained Fleiss' Kappa is **0.61**).

Post-assessment: The same 3 annotators annotated the trust expressed in 102 other messages. The results of the automatic system have been displayed, and annotators can chose the same value of trust or another one. The agreement between the annotators was also acceptable (the obtained Fleiss' Kappa is **0.69**).

Evaluation: The results obtained by comparing the classification made by the system with the annotations are presented Table 6. The annotations have been combined by using a majority vote.

Table 6 – Precision, recall and F1-score of the trust inference system using prior and post assessments

Datasets	Class	Number of posts	P	R	F
Prior assessment	Trust	28	0.67	0.93	0.78
	Distrust	4	0.50	0.25	0.33
	Neutral	65	0.96	0.83	0.89
	Global	**97**	**0.86**	**0.84**	**0.84**
Post assessment	Trust	31	0.77	0.87	0.82
	Distrust	5	0.23	0.60	0.33
	Neutral	64	0.92	0.75	0.83
	Global	**100**	**0.84**	**0.78**	**0.80**

Step 3: Computing the reputation

First, Table 7 presents the number of authors, the mean of the reputation and the standard deviation of the reputation for each rank.

Table 7 - The mean and the standard deviation of the reputation for each rank

Rank	Number of authors	Avg(R)	Std(R)
New member	561	0.29	0.29
Regular member	42	0.35	0.10
Accustomed member	26	0.35	0.09
Active member	41	0.38	0.07

Next, Figure 1 presents a scatterplot of the reputation and the number of posts while Figure 2 presents a scatterplot of the reputation and the number of replies.

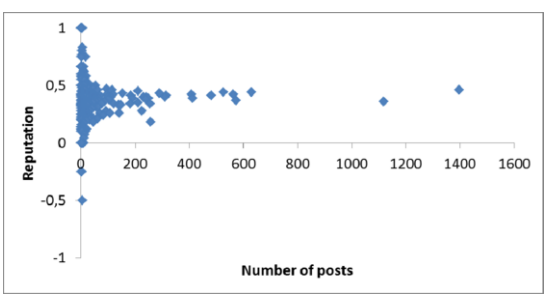

Figure 1 - The reputation and number of posts scatterplot

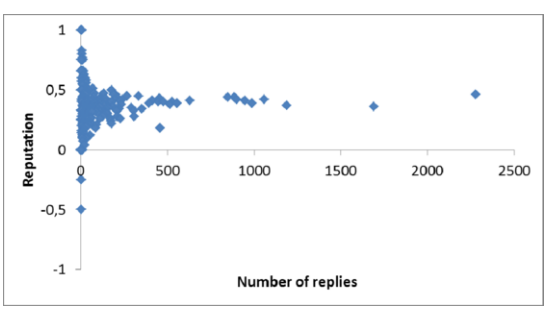

Figure 2- The reputation and the number of replies scatterplot

Discussion

In this section, we discuss the results obtained in each step.

Step 1: Finding the recipient

As expected, results obtained by using a post-assessment have been better from those obtained using prior-assessment. The difference is larger between the recalls than between the precisions. The gain in precision is 0.10 for the development set and 0.02 for the test set whereas the gain in recall is 0.16 for the development set and 0.17 for the test set. This observation can be explained by two reasons. First, the nature of the prior assessment itself gives the non-expert annotators much more freedom to choose the links, which increases the chances of validating links that the heuristic will not find. Furthermore, the non-expert annotators do not know the heuristic rules at all, they may validate links that the heuristic is not designed to find. Their annotations may be useful to add or update some rules. But surprisingly, the results obtained from the test set have been better than those obtained from the development set. Luckily, the test set has less particular cases that we did not implement in our heuristic.

Step 2: Inferring the trust

Unlike the first step, results obtained by using a prior assessment have been slightly better than those obtained using a post assessment. This observation tends to reduce the effect of the knowing the system's results while annotating in the case of trust inference. This small difference may be due to the chosen posts and not to the way that annotations have been done. The results obtained on the trust class have been good, but the recall is higher than the precision using both prior and post assessment. It means that the system finds the majority of posts expressing trust but also gives some posts that are not expressing it as so. Therefore, even if the list of trust expressions has been built manually, it seems to be sufficient to find the majority of trust posts. The results obtained on the neutral class have also been good, but the precision is higher than the recall. It means that the majority of neutral posts have been correctly classified but some posts have not been found by the system (classified in other classes). Finally, the results obtained on the distrust class have been the worst but it is difficult to make conclusions since very few distrust posts have been annotated. In fact, users in this forum do not usually express a lot of disagreement since the first goal is to exchange emotional support.

Step 3: Computing the reputation

Since the forum has very few disagreement posts, the reputation of almost all users is positive. Only, 3 users had negative reputation (two points are superposed in Figure 1 and 2). They posted less than 7 posts and received less than 5 replies. Indeed, the more posts a user receives the more chances he has to receive trust posts and more importantly neutral posts. This is why the standard deviation of the reputation decreases with the increase of the number of posts. Because neutral posts have a mitigating effect to both good and bad reputations.

As presented Table 7, the majority of users have the first rank on the forum (New member). The users' reputations mean increases slightly from lower rank levels to higher ones while the standard deviation decreases. However, authors from the first rank level have reputations ranging from -0.5 to 1 which is very diverse. This observation confirms our hypothesis that the Website rank is not a good estimation of the users' trustworthiness.

Conclusion

In this paper, we presented a method to infer the user's reputation from the posts he receives in online health forums. Indeed, a lot of trust is expressed inside the posted text which can not be detected by structure based trust models. Our method may be used either by the users to have an idea on the trustworthiness of each user or by the moderators, for example to reward the users having the best reputations and to detect the ones having the worst reputation. The method is divided into three steps. First, the recipients of each post are identified using a rule based heuristic. Next, the trust expressed in each post is evaluated by searching the use of agreement, disagreement and thanking expressions. Finally, the reputation of each user is obtained by aggregating the trust or distrust expressed in all the posts addressed to him. The method has been tested on one French health forum specialized in breast cancer where users exchange a lot of emotional support but few disagreement. Therefore most of them have got positive reputations. Manual annotations have been done in order to evaluate the methods. The results obtained for the first and the second step are encouraging.

This paper presents a first implementation and test of a method that infers the trust from the posts addressed to each user. Many perspectives can be done in order to go further in this idea. First, the users' reputations are now computed in the whole forum, we may also compute the reputation in each topic since one user may be trusted differently in several topics (reputation is topic dependent), which can be done by considering only posts received for specific topics. Second, even if the annotated corpus is relevant, still now its size is quite small according to the size of the whole forum. Improving a more complete annotation could probably provide more accurate results. Therefore, we are planing to use crowdsourcing services in order to annotate the whole forum. The quality of annotations obtained using crowdsourcing services are usually questioned but many solutions may be used to overcome this issue. For example, we can put the posts that had a perfect agreement in this study between the posts of the new corpus to make sure that the future annotators will correctly do their work. A large annotated corpus will not only gives us a better estimation of the method's performances, but can also be used in order to learn models that will be able to better classify each post according to the expressed trust. Text mining and supervised classification techniques might give better results in this case.

Moreover, following the evolution of the user's reputation since his registration may also be very interesting. Indeed, we noticed that new users usually receive less trust replies than old ones. But, their reputation can increase over time. Especially in the case of chronic diseases, users remain on the forum for years. Using their experiences and the knowledge acquired, they will become experts and will receive more thanking and agreement replies. Finally, the posts expressing trust posted by users having good reputation may have more weight than posts expressing trust posted by users having a bad reputation. One way to include these propagation aspects is to use PageRank based trust models [14]. This algorithm ranks webpages according to their importance. The basic idea is to give more importance to web pages that are pointed by many other important pages. It has been wihdly applied in social media for example to find key users in terms of connectivity and communication activity [15].

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" (https://www.lirmm.fr/patient-mind/pmwiki/pmwiki.php?n=Site.Accueil).

References

[1] H. S. Wald, C. E. Dube, and D. C. Anthony, "Untangling the Web—The impact of Internet use on health care and the physician–patient relationship," *Patient Educ. Couns.*, vol. 68, no. 3, pp. 218–224, Nov. 2007.

[2] S. M. Akerkar and L. S. Bichile, "Doctor patient relationship: changing dynamics in the information age," *J. Postgrad. Med.*, vol. 50, no. 2, pp. 120–122, Jun. 2004.

[3] A. Abdaoui, J. Azé, S. Bringay, and P. Poncelet, "Predicting Medical Roles in Online Health Fora," presented at the 2nd International Conference on Statistical Language and Speech Processing, SLSP 2014, Grenoble, 2014.

[4] M. Deutsch, "Cooperation and trust: Some theoretical notes," in *Nebraska Symposium on Motivation, 1962*, Oxford, England: Univer. Nebraska Press, 1962, pp. 275–320.

[5] S. P. Marsh, "Formalising Trust as a Computational Concept," Ph.D. Thesis, University of Stirling, Department of Mathematics and Computer Science, 1994.

[6] R. Falcone and C. Castelfranchi, "Trust and Deception in Virtual Societies," C. Castelfranchi and Y.-H. Tan, Eds. Norwell, MA, USA: Kluwer Academic Publishers, 2001, pp. 55–90.

[7] J. Golbeck, "Trust and Nuanced Profile Similarity in Online Social Networks," *ACM Trans Web*, vol. 3, no. 4, pp. 12:1–12:33, Sep. 2009.

[8] F. Skopik, H.-L. Truong, and S. Dustdar, "Trust and Reputation Mining in Professional Virtual Communities," in *Proceedings of the 9th International Conference on Web Engineering*, Berlin, Heidelberg, 2009, pp. 76–90.

[9] A. Gruzd and C. Haythornthwaite, "Automated Discovery and Analysis of Social Networks from Threaded Discussions. Paper presented at," in *the International Network of Social Network Analysts, St. Pete Beach*, 2008.

[10] M. Forestier, J. Velcin, and D. Zighed, "Extracting Social Networks to Understand Interaction," in *2011 International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2011, pp. 213–219.

[11] D. Péchoin, *Thésaurus*. Paris: Larousse, 1999.

[12] H. Schmid, *Probabilistic Part-of-Speech Tagging Using Decision Trees*. 1994.

[13] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, 1971.

[14] J. Caverlee, L. Liu, and S. Webb, "Towards Robust Trust Establishment in Web-based Social Networks with Social trust," in *Proceedings of the 17th International Conference on World Wide Web*, New York, NY, USA, 2008, pp. 1163–1164.

[15] J. Heidemann, M. Klier, and F. Probst. Identifying Key Users in Online Social Networks: A PageRank Based Approach. *ICIS 2010 Proc.*, Jan. 2010.