Classification of Online Health Discussions with Text and Health Feature Sets

Mi Zhang and Christopher C. Yang
College of Computing and Informatics, Drexel University
mi.zhang@drexel.edu, chris.yang@drexel.edu

Abstract

Nowadays, many health groups and forums are established on the Internet, where health consumers discuss health issues and interact with each other. Although there is a large amount of user generated content about healthcare on different social media sites, few studies have applied data mining or artificial intelligence techniques for knowledge discovery on a large scale of data in this particular emerging area. In online health forums, it is difficult for users to find relevant topics or peers due to the large amount of information. Traditional recommendation systems may not work well for health online forums, because health consumers have different intentions of participation or may be interest in different types of supports even if the content matches their interest. To help solving this problem, we apply Naïve Bayes methods in this study to classify posts and comments on QuitStop forum, which is an online community for smoking cessation intervention. Classifiers are built on different text features and health features of user quit status. Two different classification tasks are investigated: (1) classification of user intentions, and (2) classification of types of social support exchanged in interactions. We developed classifiers for posts and comments separately, and conducted experiments to compare classifiers with different text and health feature sets. It is found that using thread title or post content can achieve the highest classification accuracy on both posts and comments for user intention classification with text features. On the other hand, using the content of post or comment itself performs the best for the classification of social support types. In particular for the post, integrating health features of the post author can boost the text classifications of user intention and support type. However, user health features cannot help in improving text classifiers for the comments.

Introduction

With the development of Web 2.0, the concept of Health 2.0 emerges with a variety of features, including social networking, participation, apomediation, collaboration and openness (Eysenbach 2008; Belt, Engelen et al. 2010). Many online communities and social networking platforms are developed for people to discuss health issues and interact with each other. Fox et al. reported that “the social life of health information is robust”, with the fact that 52% online health inquires involved interaction with others (Fox and Jones 2009).

In Web 2.0 era, user generated content in various social media sites provides a rich resource for knowledge discovery. Data mining techniques are applied to extract knowledge from the unstructured data (Rajman and Besançon 1998). In medical and healthcare areas, data mining is applied to formal biomedical records in many studies (Cohen and Hersh 2005; Saeys, Inza et al. 2007). Although a lot of online communities, including forums and discussion groups, are built for health discussions and user interactions, few studies focus on this emerging field for knowledge extraction and discovery. In this study, we extract user discussion content from a smoking cessation forum, QuitStop, and apply classification technology to classify messages according to user intentions and social support exchange types in interactions.

QuitStop is a forum on QuitNet website, which is one of the most popular websites for smoking cessation (http://www.quitnet.com/qnhomepage.aspx), where different intervention services are provided (An, Schillo et al. 2008). QuitNet has developed 11 Web forums. QuitStop (http://forums.quitnet.com/aspBanjo/Message_List.asp?Conference_ID=10&Forum_ID=8&r=100777) is the most popular one among them, on which users can discuss the tobacco quitting process, ask questions, and give or receive social support.

QuitStop arouses a large number of discussions every day. Usually, a thread can only stay on the first page for a very short time. So it is difficult for users to look for relevant topics to discuss, or identify proper peers to communicate with. It would be helpful if we could recommend interesting topics or predict potential users for QuitStop forum. Traditionally, collaborative filtering and content-based approaches are used for recommendation. However, collaborative filtering usually suffers from cold start problem. For health forums like QuitStop, the content text is usually short, so recommendation directly on text content may not perform well. In healthcare social media, health consumers may have different intentions of participation or may be interested in different types of social support. In addition, the health status of health consumers may also indicate their specific interests. Thus in this study, we try to classify threads on QuitStop forum in terms of user intention and social support types. Based on classification result, better recommendation services could be provided on QuitStop forum or other health social media.

In intervention programs of smoking cessation, social support plays an important role in supporting patients to establish positive attitudes and confidence. Social support is "an exchange of resources between two individuals perceived by the provider or the recipient to be intended to enhance the well-being of the recipient" (Shumaker and Brownell 1984). In our previous study (Zhang, Yang et al. 2013), we used qualitative analysis to classify messages (including posts and comment) of QuitStop forum into five categories, which are offering social support, requesting social support, receiving social support, other activities of smoking cessation, and irrelevant content. These categories represent user intentions to publish corresponding posts or comments. Also, it is found that informational support and nurturant support are two major types of social support exchanged in threads of QuitStop forum. Identifying user intentions and types of social support is helpful in understanding user interactions in online health communities. The result is also helpful for developing recommendation systems that match users or content to enhance the user interactions and experience. Unfortunately, these classifications are usually done manually with multiple human annotators, which is very time consuming. In this study, we use classification techniques to classify messages for user intentions and social support types automatically. There are three potential benefits of the automatic classification of messages:

(1) Compared to manual classification of messages in health groups, automatic classification could be applied to a large scale of data, which can help us to develop a comprehensive framework to understand user discussions and interactions.
(2) Navigation could be provided in health forums based on automatic classification of threads, which supports users to find interesting information and discussion topics.
(3) With classification of user intentions and social support types, we could better match message topics with user intentions, and recommend proper topics for users to improve recommendation and user prediction.

Classification of online messages is usually based on text features. For a thread on QuitStop forum, text could be extracted from the title, post or comments. In this study, we choose different text features to classify posts and comments. In addition, we extract the health status from user profiles of QuitNet website. The health information (user quit status and user quit stage) is used as health feature sets to boost text classification. Experiments are conducted to evaluate the effectiveness of classification with different text and health feature sets.

Related Works

There is an increasing number of health forums and groups being developed on the Internet. It has drawn attention of healthcare and informatics researchers to analyze the discussion content of these online communities. These studies developed different classification categories for the discussion topics. For example, Bender et al. (Bender, Jimenez-Marroquin et al. 2011) extracted three discussion themes from 620 breast cancer groups on Facebook, including fundraising, awareness, product or service promotion related to fundraising or awareness, or patient/ caregiver support. For diabetes groups on Facebook, Greene et. al(Greene, Choudhry et al. 2011) summarized five themes, which are advertisements, providing information, requesting information, support and irrelevant. Five themes of discussions on QuitStop forum were extracted, including offering social support, requesting social support, receiving social support, other activities and irrelevant content (Zhang, Yang et al. 2013). These themes reflect the intention of message authors.

There are different types of social support identified from online healthcare communities, including forums or groups of stress control, weight loss, AIDS, and alcoholism (Cutrona and Suhr 1992; Mo and Coulson. 2008; Chuang and Yang 2010; Chuang and Yang 2011; Hwang, Ottenbacher et al. 2011). Informational support and nurturant support are two main types of social support exchanged in online health communities (Eichhorn 2008; Chuang and Yang 2010; Chuang and Yang 2011). Informational support offers information to assist patients in resolving health problems. Nurturant support comforts and consoles patients, without direct efforts to solve the problems (Zhang, Yang et al. 2013).

In the current studies, the content analysis and classification of online health discussions are usually conducted by qualitative analysis with manual coding using a relatively small dataset. As a result, it is difficult to conduct this analysis on a large volume of data in real time. In this study, we apply data mining techniques to classify the online discussions automatically; thus enabling continuous monitoring of the trends in online healthcare group and timely healthcare support to online users.

Automatic classification is an important topic in areas of data mining and artificial intelligence. Different algorithms and models are developed for classification, including rule-based classification, Naïve Bayes, Bayesian Belief Networks, Support Vector Machine (SVM), Artificial Neural Networks and so on. Classification methods are also applied to question/answer detection and knowledge extraction in online discussion forums. To identify questions and answers from a forum, previous studies developed categories and applied classification methods based on different features. For example, Kim et al. (Kim, Chern et al. 2006) classified threads in a student discussion board into categories of question, announcement, answer/suggestion, elaborate, correct/object, acknowledge/support. They considered speech act patterns in posts and proposed methods to detect conversation focus. However, they only focused on a small dataset, and the classification was implemented manually. Antonelli and Sapino (Antonelli and Sapino 2005) adopted a rule-based classifier to identify the relations of different postings. The categories included announcement, question, answer, cue, relation and entry point. Hong and Davison (Hong and Davison 2009) intended to solve problems of identifying question-related threads and their potential answers with classification techniques based on different features. Kim et al. (Kim, Wang et al. 2010) developed 12 categories to classify posts in a web forum to different types of questions and answers. They selected different feature sets, including structural features, post context features and semantic features. They also compared different algorithms of maximum entropy, SVM-HMM and CRF. Besides detecting questions and answers in web forums, some studies used classification method to evaluate the quality of post content. Huang et al. (Huang, Zhou et al. 2007) identified high-quality threads in forums with a SVM classification algorithm. They selected structural features as well as content features for classification. Similarly, Weimer and Gurevych (Weimer and Gurevych 2007) applied SVM classification based on features as surface, lexical, syntactic, forum specific and similarity to assess the quality of user generated discourse automatically. To analyze the completeness, solvedness, spam and problem types of threads in a Linux user forum, Baldwin el al. (Baldwin, Martinez et al. 2007) extracted text features from different positions of threads, and used different classification algorithms for thread classification.

Although classification and other data mining techniques have been widely used in the field of bioinformatics, most studies applied them to biologic data that are well-defined and structured, like attributes of cells, genes, proteins, etc. (Saeys, Inza et al. 2007). Text classification is usually applied to academic records, such as documents in Medline/PubMed (A survey of current work in biomedical text mining). For social media analysis of healthcare, text mining and social network analysis are used to propagate infectious diseases with hospital records, predict pandemic increase with Twitter data, model hospital structure network, or analyze health social network for some websites (Wegrzyn-Wolska, Bougueroua et al. 2011).

In online health forums and groups, there are increasing number of posts generated by users. However, it is difficult to detect the topics or themes from the unstructured data. Some studies tried to analyze the content of health posts on the web. Text mining is implemented to analyze posts of H1N1 (Kim, Pinkerton et al. 2012) and Sexually Transmitted Diseases (Oh and Park 2013) on Yahoo! Answers, as well as cancer blog posts (Kim 2009). However, these studies extracted concepts and terms based on standard medical vocabularies, like Medical Subject Headings (MeSH) Resource. As a result, only medical concepts can be identified from the social media. Some important user interactions, like social support exchange, cannot be indicated from the vocabulary-based text mining.

In this study, we classify posts and comments with text features and health features. For text features, we do not restrict the terms to medical vocabularies. Two different schemes of categories are designed to reflect author intentions and types of social support, respectively. The health features are built on user quit statuses and quit stages. To our knowledge, there are few studies considering user health status for text mining of unstructured social media data.

Problem Description

Data set

We collect data from QuitStop forum, the most popular forum on QuitNet website. Every registered user of QuitNet website can participate in discussions of threads on QuitStop. Each thread is composed of one initiating post and zero or more comments replying to the posts. We randomly collected 375 threads on QuitStop in periods of 05/01/2011 - 05/31/2011 and 07/01/2013 – 07/31/2013, which include 375 posts and 1365 comments. The messages are manually classified as gold standard. Experiments are conducted for post classification and comment classification, respectively. For each of classifications, 80% of all the records were extracted as training data and others are used as test data. The training data and test data are randomly generated five times, and the results of their average performances in the experiments are reported.

Task Description

As mentioned in above sections, two tasks are proposed in this study: classification of user intentions (Task I), and classification of social support types (Task II). The two tasks are conducted on posts and comments separately. Our goal is to classify the messages using all information that could be accessed from the context. Posts and comments are separated in classification because the intentions and social supports of posts are usually quite different from those comments that reply to them, but the intentions and social supports of the comments in the same thread can be relatively more similar.

Previous research summarized five themes of messages on QuitStop forum (Zhang, Yang et al. 2013) that reflect author intention to publish corresponding messages. Based on these themes, we developed five categories for Task I. The definitions and manually-coded examples are shown in Table 1.

Informational support and nurturant support are the main types of social support found in previous studies. Informational support is specific information about the disease, or treatment or coping (Cutrona and Suhr 1992; Chuang and Yang 2011), which includes advice, referral, fact, perceptual knowledge, personal experiences and feedback/opinion. Nurturant support is expressing caring or concern, as well as expressing the importance of relationship (Cutrona and Suhr 1992; Chuang and Yang 2011), including subcategories of esteem, network and emotional support. Task II in this study classifies messages to informational support or nurturant support.

Table 1. Intention Categories for Task I

Category	Example
Offering Social Support	Great job on your quit so far.
Requesting Social Support	I do sleep a lot and my mind receptors are ‘flickering’ to make me foggy ... and I do have those mood swings.
Receiving Social Support	Thanks everyone! I really appreciate the support and will be checking in often.
Other Activities of smoking cessation	To those of us that have been blessed to know PetroJMan (Bruce) please take a moment to send him a big hug as he has always done for us.
Irrelevant Content	Iced mochas are my favorite form of caffeine

Feature Description

Text Feature Sets

In a discussion thread, text could be extracted from its title, post and comments. To look for proper text for post classification and comment classification effectively, we select text at different thread positions, and conduct experiments to classify posts and comments separately. The text feature sets for post and comment classifications are listed in Table 2. For either post or comment classification, all information in corresponding thread is considered as text features. For post classification, besides title and the post content, we also use text of all comments in the thread as features, because comments provide context information for the corresponding post. For comment classification, besides the text content of the comment itself, we also consider the thread title and post content in the thread as the context information to build text feature sets.

Table 2. Text Feature Sets

Message Type	Text Feature Sets	Explanation
Post	Title	Title of corresponding thread
	Post	Post of corresponding thread
	Comments	All comments in corresponding thread
Comment	Title	Title of corresponding thread
	Post	Post of corresponding thread
	Comment	The comment of itself

For each text feature set, we extract text from corresponding positions of the threads, processing the raw text by discarding non-alphabetic content, removing general stop words, stemming and lemmatizing with WordNet database. The generated terms are transferred to term-count vectors. Each text feature set is composed of features of bag-of-words.

Health Feature Sets

Besides text features, we also construct health feature sets for classification. It takes a substantial long time for people to quit smoking because many still have a craving of smoking even if they have not smoked for a year or more. Quit status is an import characteristic of smoking quitters, and it reflects the outcome of smoking cessation intervention. Some users of QuitStop forum describe their quit date in their profile pages, from which we could calculate their quit statuses. For the author of each message in our datasets, we represent his/her quit status by the number of days that he/she has been abstinent from the self-reported day he/she stops smoking on the profile page to the day he/she posts the message.

According to quit statuses, users of QuitStop forum could be divided into five quit stages (Zhang, Yang et al. 2012). Users with the quit statuses of 0 to 3 months are at Stage 1 – early action stage; users with quit statuses of 3 to 6 months are at Stage 2 – late action stage; Users at Stage 3, early maintenance stage, are those who have been quitted for 6 months to 2 years; Those with quit statuses of 2 years to 5 years are at Stage 4 – late maintenance stage; and those who have been abstinent for more than 5 years are at Stage 5, which means that they have completed smoking cessation.

Quit status and quit stage of users are used as health features in this study to boost the text classification of posts and comments. We develop different health feature sets based on quit status and quit stage as shown in Table 3.

For items of quit status listed in Table 3, each item is regarded as a single feature with a continuous value. For items of quit stage, each item is regard as a feature set composed by features of five quit stages.

Table 3. Health Feature Sets

Message	Health Feature Sets	Abbreviation
Post	Quit status of the post author	PA Status
	Quit stage of the post author	PA Stage
	Mean value of the quit statuses of all comment authors in corresponding thread	Mean of CA Status
	Standard Deviation of the quit statuses of all comment authors in corresponding thread	SD of CA Status
	Quit stages of all comment authors in corresponding thread	CA Stage
Comment	Quit status of the post author in corresponding thread	PA Status
	Quit stage of the post author in corresponding thread	PA Stage
	Quit status of the comment author	CA Status
	Quit stage of the comment author	CA Stage

Approaches of Classification

In this study, methods of Naïve Bayes are used for classifications. Statistical classifiers are constructed with different features based on the training data. For each of the tasks, there are k classes denoted as $c_1, c_2, \ldots, c_k$. A message could be represented as a set of observed features $X = (x_1, x_2, \ldots, x_n)$. The probability that this message X belonging to class c_i could be calculated as:

$$P(c_i|X) = \frac{P(X|c_i)P(c_i)}{P(X)} \qquad (1)$$

Assuming that features are independent in a same dataset, the probability could be simplified to:

$$P(c_i|X) == \prod_{j=1}^{n} P(x_j|c_i) \qquad (2)$$

The likelihood $P(x_j|c_i)$ is estimated with training data. For text feature sets and the feature sets of quit stages, x_j is a categorical value in multinomial distribution. In this case, the likelihood is calculated as:

$$P(x_j|c_i) = \frac{N_{ji}}{N_j} \qquad (3)$$

where N_{ji} is the number of messages with feature x_j in class c_i in the training data, and N_j is the number of all messages with feature x_j in the training data.

For the feature of quit status, x_j a continuous value, which is assumed in Gaussian distribution. The likelihood is calculated as:

$$P(x_j|c_i) = \frac{1}{\sqrt{2\pi\sigma_i{}^2}} \exp\left(-\frac{(x_j-\mu_i)^2}{2\sigma_i{}^2}\right) \qquad (4)$$

where μ_i and σ_i are the mean and the standard deviation of the feature value for all messages in the training set.

For each message X in the test data, $P(c_i|X)$ is calculated according to formula (2) for all the classes. The message is assigned to the class with the highest probability.

For each of the tasks, the experiments are implemented on posts and comments, respectively. First, classification models are built on different text feature sets as shown in Table 2, and the results are compared. Then, we use each of the health feature sets in Table 3 to boost text classifications. Specifically, to combine a text feature set and a health feature set, we built two different classifiers based on each of the feature sets separately. For a class c_i and a message X in the test set, let $P_{text}(c_i|X)$ be the probability returned by the text classifier, and $P_{health}(c_i|X)$ be the probability returned by the classifier with a health feature set, we compute the logarithms of probabilities, and combine them by:

$$L(c_i|X) = Log(P_{text}(c_i|X)) + w * Log(P_{health}(c_i|X)) \quad (5)$$

where w is a weight to combine the two classifiers, and $Log(V)$ is the logarithm value of V. For each of the k classes, $L(c_i|X)$ is calculated by formula (5), and the message X is assigned to the class with the highest value of $L(c_i|X)$.

Accuracy is used to evaluate the performances of different classification models. For the test data, accuracy is computed by the ratio of the number of messages with correct prediction and the total number of all messages classified.

Results and Discussions

Task I - Classification of User Intentions

In this task, we classify posts and comments from the perspective of author intention. We first use different text feature sets to classify posts and comments separately, and then apply health feature sets to boost text classifications. For the post classification, the accuracy of using title, post and comments are 0.547, 0.477 and 0.464, respectively. So, using title as the text feature can achieve the best classification result of post intentions. For comment classification, the accuracy of using title, post and comment are 0.810, 0.830 and 0.682, respectively. It indicates that text in a comment itself may not be the best reflection of the author's intention. The thread title and the post content can better indicate the author intention of a comment. The reason is that the comments in a thread are responses to the corresponding post. Usually, the author intention to make a comment is determined by the content of corresponding post. Different comments in a same thread may use various terms, but the user intentions are usually the same. In addition, some comments only consist of a few words (e.g. “Keep up your good work”), which make the classification task more challenging.

For posts and comments, we use health feature sets listed in Table 3 to boost text classifications with different w. The result is shown in Figure 1 and Figure 2. For the post classification in Figure 1, with a low weight of the quit status or quit stage of the post author as the health feature set, the accuracies of text classifiers can all be improved, which indicates that the health characteristics of the post authors are associated with their intention to publish posts. However, classifiers with heath features of comment authors could not outperform any text classifiers. As shown in Figure 2, none of the health feature sets can help improving the text classifications if comments, which indicate that the user intention to publish comments is not associated with their health statuses.

Task II – Classification of Support Types

In this task, posts and comments are classified to different types of social support, including informational support and nurturant support. For the post classification, the accuracies of using title, post and comments are 0.759, 0.801and 0.788, respectively. Text in the post can best indicate the types of social support. For the comment classification, the accuracies of using title, post, and comment are 0.891, 0.880 and 0.917, respectively. The classifier using the comment text can achieve the best accuracy. So, to classify types of social support of both posts and comments, using content of the messages themselves can achieve the best result.

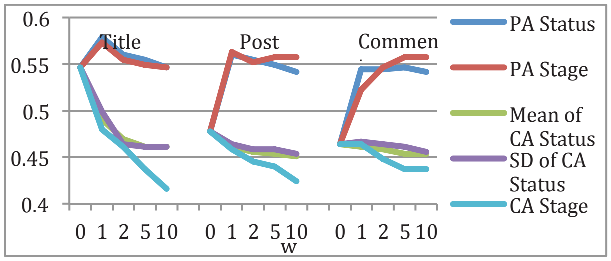

Figure 1 Task I with Text and Health Feature Sets for the Posts

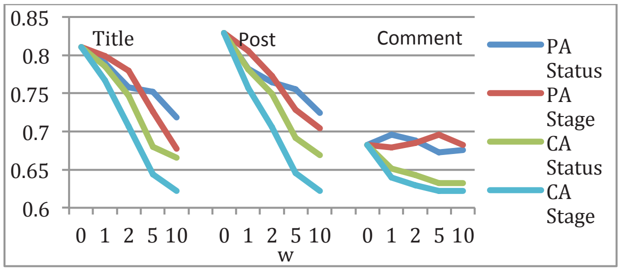

Figure 2 Task I with Text and Health Feature Sets for Comments

Health feature sets in Table 3 are used to boost text classifications, and the results are shown in Figure 3 and Figure 4. For the post classification as shown in Figure 3, it achieves the highest accuracy when integrating the health feature of post authors’ quit status with a large weight (w > 5) in any text classifier. In this case, the quit status of the post author plays an important role for classification that overwhelm text features. The quit status of the post author can effectively predict the types of social support in the posts. However, as shown in Figure 4, none of the health features sets can improve the text classification of comments. These results reflect that the quit status of post authors is useful in determining the social support types they involve but it is not the case for the quit status of the comment authors. It conforms to our observations that the users in the early quit status tend to involve in informational support but the users in the late quit status tend to involve in nurturant support when they initiate a post in a thread. However, this pattern is not observed when the users make a comment in a thread. The users of any quit status can involve in informational support or nurturant support when they write a comment in a thread.

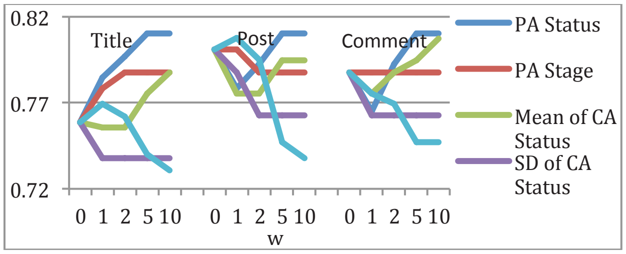

Figure 3 Task II with Text and Health Feature Sets for the Post Classification

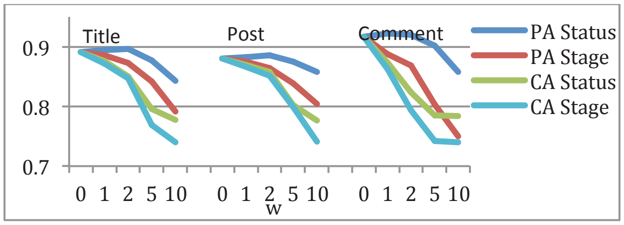

Figure 4 Task II with Text and Health Feature Sets for the Comment Classification

Conclusion

In this study, we apply Naïve Bayes method to classify posts and comments on QuitStop forum for author intentions and types of social support. Text features and health features are constructed for post and comment classification separately, and two different tasks are developed. There are three findings of our experiments:

(1) For intention classification with text features only, information provided by the initiated post is effective to classify both post and comment. The thread title and post content perform better than comment text for post and comment classification.

(2) For the classification of social support types with text features, using message content for either post or comment performs the best. Concretely, post content is effective for post classification, and comment content is effective for comment classification.

(3) Quit status and quit stage of the post author can help improve text classifications of posts. For intention classification, integrating quit status or quit stage with a low weight can achieve the best result; for support type classification, only using quit status of post authors can reach the highest accuracy.

(4) For comments, none of the health feature sets could improve text classifications of intention or social support type.

Nowadays, social media becomes an important platform for health discussions and interactions. Techniques of data mining and artificial intelligence could be applied on knowledge discovery and extraction from this data. In this study, we applied classification techniques to identify user intentions and social support types from a forum of smoking cessation intervention, which is an extension of previous works of manual coding. In the future, more techniques would be developed for user classifications and predictions for healthcare areas on social media.

References

An, L. C., and Schillo, B. A. eds. 2008. Utilization of Smoking Cessation Informational, Interactive, and Online Community researches as Predictors of Abstinence: Cohort Study. *J. Med. Internet Res* 10: e55.

Antonelli, F. and M. L. Sapino. 2005. A rule based approach to message board topics classification. In Proceedings of the 11th international conference on Advances in Multimedia Information Systems: 33-48. Sorrento, Italy: International Conference on Advances in Multimedia Information Systems.

Baldwin, T., D.; Martinez D.; and Penman R.B. 2007. Automatic Thread Classification for Linux User Forum Information Access. In Proceedings of the 12th Australasian Document Computing Symposium: 72-79. Melbourne, Australia: Australasian Document Computing Symposium.

Van De Belt, T. H; Engelen, L.J.L.P.G.; Berben, S.A.A.; and Schoonhoven, L. 2010. Definition of Health 2.0 and Medicine 2.0: A Systematic Review. *J. Med. Internet Res* 12(2): e18.

Bender, J. L.; Jimenez-Marroquin, M-C.; and Jadad, A.R. 2011. Seeking Support on Facebook: A Content Analysis of Breast Cancer Groups. *J Med Internet Res* 13(1): e16.

Chuang, K. and Yang, C.C. 2010. Social Support in Online Healthcare Social Networking. In Proceedings of iConference 2010. Urbana-Champaign, Illinois: iConference.

Chuang, K. and Yang, C.C. 2011. Helping You to Help Me: Exploring Supportive Interaction in Online Health Community. In Proceedings of ASIS&T 2010. Pittsburgh, PA: ASIS&T 2010.

Cohen, A. M. and Hersh, W. R. 2005. A survey of current work in biomedical text mining. *BRIEFINGS IN BIOINFORMATICS* 6(1): 57-71.

Cutrona, C. E. and Suhr J.A. 1992. Controllability of Stressful Events and Satisfaction With Spouse Support Behaviors. *Communication Research* 19(2): 154-174.

Eichhorn, K. C. 2008. Soliciting and Providing Social Support Over the Internet: An Investigation of Online Eating Disorder Support Groups. *Journal of Computer-Mediated Communication* 14(1): 67-78.

Eysenbach, G. 2008. Medicine 2.0: Social Networking, Collaboration, Participation, Apomediation, and Openness. *J. Med. Internet Res* 10(3): e22.

Fox, S. and Jones S. 2009. The Social Life of Health Information. http://www.pewinternet.org/2009/06/11/the-social-life-of-health-information/

Greene, J. A.; Choudhry, N. K. ; Kilabuk E.; and Shrank W.H. 2011. Online Social Networking by Patients with Diabetes: A Qualitative Evaluation of Communication with Facebook. *J Gen Intern Med* 26(3): 287-292.

Hong, L. and Davison, B. D. 2009. A Classification-based Approach to Question Answering in Discussion Boards. In proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR'09). Boston, Massachusetts: SIGIR'09.

Huang, J.; Zhou, M. and Yang, D. 2007. Extracting Chatbot Knowledge from Online Discussion Forums. the 20th international joint conference on Artifical intelligence (IJCAI'07). Hyderabad, India: IJCAI'07.

Hwang, K. O.; Ottenbacher, A. J.; Green, A.P.; Cannon-Diehl, M.R.; Richardson, O.; Bernstam, E.V. and Thomas, E.J. 2011. Social Support in an Internet Weight Loss Community. *Int J Med Inform* 79(1): 5-13.

Kim, J.; Chern, G.; Feng, D.; Shaw, E. and Hovey, E. 2006. Mining and Assessing Discussions on the Web through Speech Act Analysis. In proceedingsof ISWC'06 Workshop on Web Content Mining with Human Language Technologies. Athens, GA, USA: ISWC'06.

Kim, S. 2009. Content analysis of cancer blog posts. *J Med Libr Assoc* 97(4): 260-266.

Kim, S.; Pinkerton, T. and Ganesh, N. 2012. Assessment of H1N1 questions and answers posted on the Web. *American Journal of Infection Control* 40(3): 211–217.

Kim, S. N.; Wang, L. and Baldwin T. 2010. Tagging and Linking Web Forum Posts. In proceedings of the Fourteenth Conference on Computational Natural Language Learning. Uppsala, Sweden.

Mo, P. K. H. and Coulson, N. S. 2008. Exploring the Communication of Social Support within Virtual Communities: A Content Analysis of Messages Posted to an Online HIV/AIDS Support Group. *CyberPsychology & Behavior* 11(2): 371-374.

Oh, S. and Park, M. S. 2013. Text Mining as a Method of Analyzing Health Questions in Social Q&A. in proceedings of ASIST 2013. Montreal, Quebec, Canada: ASIST 2013.

Rajman, M. and Besançon, R. 1998. Text Mining - Knowledge extraction from unstructured textual data. in proceedings of the 6th Conference of International Federation of Classification Societies (IFCS-98). Rome, Italy: IFCS-98.

Saeys, Y.; Inza, I. and Pedro, L. 2007. A review of feature selection techniques in bioinformatics. *bioinformatics* 23(19): 2507–2517.

Shumaker, S. A. and Brownell, A. 1984. Toward a Theory of Social Support: Closing Conceptual Gaps. *Journal of Social Issues* 40(4): 11-36.

Wegrzyn-Wolska, K. and Bougueroua, L. 2011. Social Media Analysis for e-Health and Medical Purposes. In proceedings of the 2011 International Conference on Computational Aspects of Social Networks (CASoN). Salamanca, Spain: 278-283.

Weimer, M. and Gurevych, L. 2007. Predicting the Perceived Quality of Web Forum Posts. In proceedings of Recent Advances in Natural Language Processing 2007 (RANLP '07). Borovets, Bulgaria.

Zhang, M; Yang, C.C. and Gong, X. 2013. Social Support and Exchange Patterns in an Online Smoking Cessation Intervention Program. In proceedings of IEEE International Conference on Healthcare Informatics (ICHI). Philadelphia, US: ICHI.

Zhang, M.; Yang, C.C. and Li, J. 2012. A comparative study of smoking cessation intervention programs on social media. In Proceedings of SBP2012, College Park, MD, USA: SBP 2012.