1-1-2010

INFORMATIONAL SUPPORT OR EMOTIONAL SUPPORT: PRELIMINARY STUDY OF AN AUTOMATED APPROACH TO ANALYZE ONLINE SUPPORT COMMUNITY CONTENTS

Kuang-Yuan Huang
University at Albany, kh799292@albany.edu

Priya Nambisan
University at Albany, pnambisan@uamail.albany.edu

Özlem Uzuner
University at Albany, ouzuner@uamail.albany.edu

INFORMATIONAL SUPPORT OR EMOTIONAL SUPPORT: PRELIMINARY STUDY OF AN AUTOMATED APPROACH TO ANALYZE ONLINE SUPPORT COMMUNITY CONTENTS

Research-in-Progress

Kuang-Yuan Huang
College of Computing and Information
University at Albany, SUNY
7A Harriman Campus, Suite 220
1400 Washington Avenue
Albany, NY 12222
kh799292@albany.edu

Priya Nambisan
School of Public Health
University at Albany, SUNY
1 University Place
Rensselaer, NY 12144
pnambisan@uamail.albany.edu

Özlem Uzuner
College of Computing and Information
University at Albany, SUNY
7A Harriman Campus, Suite 220
1400 Washington Avenue
Albany, NY 12222
ouzuner@uamail.albany.edu

Abstract

Recognizing the need for analyzing large amounts of data in the study of online support communities, an automated content analysis method is introduced in this article. By adopting machine learning techniques and tools, this method requires minimal manual intervention while capable of analyzing large amounts of data automatically. Through this method, contents of messages from online support communities spanning over years are categorized as either informational support or emotional support. A case study on the analysis of online breast cancer and prostate cancer message boards is presented to demonstrate that the proposed method generates results comparable to results concluded from traditional manual qualitative content analysis methods.

Keywords: Research method, content analysis, online support community, machine learning

Introduction

Online support communities are formed by people with similar life situations (e.g., pregnancy) or illnesses (e.g., cancer) to discuss their feelings and thoughts and to search for support anonymously at any time and from any place (Pfeil 2009). The number of such online communities seems to be growing quite rapidly (Fox & Fallows, 2003) and this has led several researchers to study different aspects of online communities. For example, there are several studies focusing on perceived social support (Uden-Kraan et al., 2008), types of support sought and provided in online support communities (Coulson, 2005), relationships between the members of such communities (Pfeil and Zaphiris 2007), and outcomes of participants of online support communities (Lieberman & Goldstein, 2005). According to Pfeil (2009, p. 122), "it is important to investigate online support communities in order to make an informed judgment about their benefits and problems."

Among these studies on online support communities, knowing which types of social support are sought and provided is one of the main focuses. By understanding the type of social support sought/provided, researchers and practitioners could get an insight into the dynamics of user behavior in various online support communities specifically for patients with different diseases, demographic groups, or for family members. For example, studies found gender of the members influenced the type of support sought/provided in online cancer support communities (Gooden and Winefield 2007; Klemm et al. 1999; Seale et al. 2006); Coulson (2005) analyzed support communicated through online support networks for individuals with irritable bowel syndrome; Braithwaite et al. (1999) studied the online support community for patients with disabilities. Many of these studies used qualitative methods such as grounded theory (Corbin and Strauss 2007) or deductive thematic analysis (Boyatzis 1998) to examine online message contents.

Qualitative methods are helpful in analyzing and conceptualizing support sought/provided in online message contents compared to quantitative analysis. Qualitative content analysis, however, is normally a tedious and time-consuming task, and thus limits the amount of online messages being analyzed. The lack of enough data spanning a long period of time would engender the analysis results less representative of the whole community, as pointed out by Coulson (2005) in his effort to study online communities. In addition, due to the huge amount of effort and time required, it is often difficult to collect and compare message contents of multiple online support groups. Alternative methods that shorten the time and minimize the effort required to analyze the discourse of online support communities thus would be useful. Comparative keyword analysis is one of these alternatives (Seale et al. 2006). In their study, message contents of online breast and prostate cancer support communities are processed through WordSmith Tools (Scott 2004) to generate a list of keywords based on their relative frequency of occurrence in the two support communities. Their method, however, still requires considerable human intervention to analyze identified keywords. The need of relieving human efforts while analyzing large amounts of data thus calls for an automated content analysis method.

Furthermore, the knowledge of the different types of support sought/provided in online support communities also will allow more studies in this area. For example, it will allow us to further our research in social network analysis on social structures of online support communities. In addition, by understanding the different types of support sought/provided, studies can focus on the trend of support sought/provided across time periods, whether at the individual, group or community level. To achieve these research goals, first a method that provides timely analysis on large amounts of online message contents to understand various aspects of online social support is required. Online support communities are such a fascinating research area with many existing and undiscovered research topics. Alternate research methods would help study the full potential of different research topics in this area.

In order to facilitate studies on online support communities, a new method to automatically classify message contents of online support communities into types of support is proposed in this article. Specifically, by applying machine learning techniques and tools, this method analyzes and classifies message contents of online support communities automatically as either informational support or emotional support. With this analysis method at hand, various existing and potential research topics in studying online support communities, and even other studies that require analyzing large amounts of data can then be investigated more efficiently.

In the following sections, first the background of social support and online social support groups is provided. Concepts of automated text classification are discussed next. Then the method of automated social support classification using a machine learning approach is proposed and evaluated, followed by a case study. This article will conclude with the limitations and implications of this method and future works based upon it.

Background

Social Support and Online Social Support Communities

Social support can be defined as "information leading the subject to believe that he is cared for and loved, esteemed, and a member of a network of mutual obligations" (Cobb 1976, p. 300). More specifically, social support is "the exchange of verbal as well as nonverbal messages in order to communicate emotional and informational messages that reduce the retriever's stress" (Pfeil 2009, p. 124). Coulson (2005) indicated that support groups provide "mutual aid and self-help for individuals facing chronic disease, life threatening illness and dependency issues" (p. 580). Social support has been studied from various perspectives (Heaney and Israel 2002; Lakey and Cohen 2000) as to its benefits to support seekers, either through direct effect by increasing support seekers' physical, mental and social health and reducing their mortality rate (Berkman and Glass 2000; Thoits 1995) or through a buffering effect between stressful life events and patients' health by helping them cope with these events (Cohen and Wills 1985; Thoits 1995).

With the advent of the Internet and online discussion boards and their growing users in recent years, scholars also shift their focus to the study of online support communities. Differences between online support communities and its offline counterpart are pointed out (e.g., Klemm et al. 1998; Pfeil 2009). Examples of advantages of online support communities are: greater accessibility in terms of time and space; encouraging connection through weak ties, resulting in better access to diverse information and experts (Wellman et al., 1996); and offering anonymous communication. On the other hand, compared to offline support communities, online communities can have some disadvantages such as: information posted in online support communities may be incorrect or hostile to other users; lack of nonverbal cues of communication; excluding people with low literacy levels etc. There have been many studies in this area where researchers endeavor to understand the dynamics of online interaction (e.g., Klemm et al. 1999; Maloney-Krichmar and Preece 2005; White and Dorman 2000; Pfeil and Zaphiris 2009).

Among the studies of social support communities, online or offline, a constant interest is the type of support sought/provided within support communities. As indicated by Schaefer et al. (1981), "social support can have a number of independent components serving a variety of supportive functions" (p. 385). Knowledge about the different types of support sought/provided would provide us with an insight into the behavior and relationships among participants of diverse social support communities. A literature review in this area shows that there are various classifications of social support. Schaefer et al. (1981) classified social support into emotional, tangible, and informational support. House (1981) categorized social support as four types of behavior: emotional, instrumental, informational and appraisal. Cutrona and Russell (1990) identified five types of social support: informational, emotional, esteem, tangible assistance, and social network support, which are further grouped into two broad categories: action-facilitating support (including informational and tangible support) and nurturant support (including emotional, network, and esteem support) (Cutrona and Suhr 1992). Kleem et al. (1998) used support types of information giving/seeking, encouragement/support, personal experience, personal opinion, prayer, thanks, humor and miscellaneous in their study on online cancer support groups. These classifications are also further adopted as the theoretical basis of various research (e.g., Coulson 2005). Despite the different theoretical framework of support classification these studies use, informational and emotional support have been generally concluded as the most common types of social support (Pfeil 2009).

In the research on types of social support sought/provided in online cancer support communities, qualitative content analysis has been one of the most widely adopted methods of study. Through content analysis, online message contents are scrutinized manually and are classified into different categories, inductively (e.g., Gooden and Winefield 2007) or deductively (e.g., White and Dorman 2000). In order to fully understand the target online support communities, analysis of messages spanning a long period of time is expected. Analyzing message contents manually however, is a time consuming task and requires much time and effort. This limits the amount of data that can be analyzed. For example, by adopting grounded theory, Gooden and Winefield (2007) collected and analyzed messages that spread over only a one-month period from online support communities. In another study by Klemm et al. (1999), online messages spanning a forty-one day period were collected for analysis. The use of such limited amounts of data for analysis risks a less holistic view and may lead to a biased view of the intricacy of support givers/seekers' online behaviors. Some studies attempt to address this issue by collecting data evenly spaced over multiple periods of a year, considering possible differences of message contents in different months of a year. For instance, Klemm et al. (1998) collected data from four days in June and five days in January; in White and Dorman

(2000), data were collected from the first five days of March, June, September, and December. However, the collected data for analysis were still trivial compared to the total messages posted over years on many online support communities. Moreover, this same concern about content analysis on a small, non-representative dataset is also pointed out in other Computer Mediated Communication (CMC) research (e.g., De Wever et al. 2005).

To introduce a new method of studying online social support, in this study we will take Cutrona and Suhr's (1992) categories of action-facilitating support and nurturant support as our reference of classification of message contents in online support communities. In their definition, action-facilitating support is "intended to assist the stressed individual to solve or eliminate the problem that is causing his or her distress," and nurturant support "encompasses efforts to comfort or console, without direct efforts to solve the problem causing the stress" (p. 155). We made two minor changes to their use of the two types of support: 1. We intentionally disregard the tangible support in our analysis because as pointed out, in the online setting, tangible support of direct aid or services are very rare, as users of online support communities are generally dispersed geographically and can barely meet to provide tangible support (Pfeil 2009). 2. In this study we still refer action-facilitating support as informational support and nurturant support as emotional support since in a broader view, action-facilitating support provides problem-solving information and nurturant support brings emotional comfort and attachment.

Automated Text Classification

In the computer science discipline, the task of automated text classification is to categorize documents into a pre-defined set of classes automatically based on their content. Compared to the manual text categorization process, automated text classification has the advantage of effectiveness in terms of saving time and labor power, and higher portability by application to other task domains (Sebastiani 2002). Examples of automated text classification tasks include topical classification (e.g., to classify news articles as sports/politics/economic topics, Nigam et al. 2000); text filtering (e.g., spam mail detection, Sahami et al. 1998); text genre classification (e.g., factual/opinion text classification, Riloff and Wiebe 2003); and sentiment analysis (Pang and Lee 2008). In the task of classifying message contents into types of social support, an example is given here:

Suppose there are two sentences extracted from an online support community:

> *"I also take tamoxifen and I have had no problems on it."*

and

> *"I want to wish you good luck in all the coming procedures."*

As human classifiers we can categorize the first sentence into informational support and the second one as emotional support with minimal effort. However, to do it automatically through software tools, a mechanism used to characterize document contents is required such that decisions of classification can be made. This mechanism can be a set of human-generated rules specifying how to classify text. In this case, the class to which a text is categorized is dependent on how it is interpreted by the rules. These rules take the form of:

> ***if** [criterion] **then** [category]* (ex. **if** ["good luck" in message] **then** [emotional support])

A more popular and state-of-the-art mechanism is to do automated text classification through a machine learning approach (Sebastiani 2002), which is also the approach we use here. Through a machine learning approach, first a so-called "learner" software tool is fed with a small set of manually classified documents. By analyzing and comparing these manually classified documents via a set of pre-determined document characteristics (such as document length, specific words which occur in the document), the automated text classifier is "trained" and generated through the learner. The generated text classifier is then ready to classify a new set of unseen documents, which is also called "to predict" (the amount of unseen documents is normally much larger than those manually classified ones). In the studies of machine learning approach, what matters is the accuracy of the automated classification done by the classifier. The most important factors affecting the accuracy of classification are the above mentioned pre-determined document characteristics and the type of classifier to be trained to. Different document characteristics capture different aspects of content features, while different types of classifiers determine "how to" classify given document characteristics into classes. The choices of the document characteristics and classifier type will be described in more detail in the next section.

Automated Support Classification: A Machine Learning Approach

The proposed method is innovative in that, although the topic of automated content analysis has been extensively studied in the Computer Science research field (e.g., Pang and Lee 2008), the focus of these studies, however, is on various machine learning techniques. To the best of the authors' knowledge, this article is the first endeavor to apply this method from the Computer Science research field to the study of online support communities. Figure 1 illustrates this method. The basic approach of this method is to first separate downloaded data into two non-overlapping datasets, one small and one large. By adopting machine learning techniques and tools, message contents in the small dataset (which is also called the training set) are used to create the classifier capable of analyzing and classifying message contents automatically. The classifier is then applied to the large dataset (which is also called the testing set) to predict the categories of the online message contents, which is free from human intervention. The value of this proposed method lies in its capability of analyzing large amounts of data, resulting in classifications better representative of the online support community. In this section, the method used to automatically classify message contents of online support communities as informational support or emotional support is described, experimented and evaluated, and the focus will be put on the creation of the classifier. The prediction on the testing dataset will be demonstrated in detail in the next section. The unit of analysis in this study is the sentence. As a result, each sentence in a message will be processed and classified into one of the two pre-defined classes.

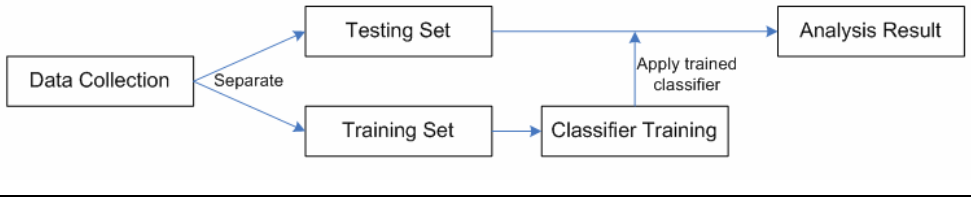

Figure 1. Method of automated support classification

Data Collection

The messages we used to do the classification task are acquired from a large online cancer support community with more than a hundred thousand registered members posting hundreds of messages every day. This community hosts various discussion boards for patients with different types of cancers and for their caregivers. For the purpose of the case study demonstrated in the next section, the focus here is on the breast cancer support discussion board. RSS (Really Simple Syndication) feed is used to download messages, resulting in two sets of messages. The small set includes 326 messages, containing 2174 sentences from the breast cancer support discussion board with date of post ranging from Oct. 2009 to Feb. 2010. These messages comprise the training dataset. The large set has 10000 messages, containing 55175 sentences. These are testing dataset messages.

Ethical Concern

Ethical issues of personal privacy and potential psychological harm should be considered before conducting qualitative research on online communities (Eysenbach 2001). Given the public nature of the target online support community, all the personal postings are publicly accessible without user registration and can be searched through google.com. In addition, the number of members in this online support community is expected to be more than 100,000, which far exceeds the number of 10 or 100 that may require privacy concerns pointed out by Eysenbach (2001). As a result, we regard the target online community as a public space and no informed consent is needed. Still, to ensure that there will be no ethical issues, we do not use direct quotes from the message content, nor do we disclose any information that is identifiable to members of the online support community in this article.

Procedure

As mentioned above, to do automated text classification using the machine learning approach, the pre-determined document characteristics (also called document "features") and choice of classifier type are needed. The type of classifier we chose for this study is support vector machine (SVM) (Joachims 1998). As a statistical learning

algorithm (Vapnik 1999), SVM has been shown to be effective doing automated text classification (Joachims 1998) and thus is suitable to our support classification task. In the experiment we adopt LIBSVM software library (Chang and Lin 2001) that implements the SVM algorithm to do the training and classifying task. While in this study these choices are made, the proposed automated method is not confined with a specific tool or classifier type.

In the current study, three types of document characteristics (features) are identified to represent the cancer online support community messages. The selection of the document features is crucial since these features determine how the trained classifier discerns the patterns of their occurrences in messages of different categories statistically and thus is able to differentiate between these messages. The three selected features are: 1. Bag-of-Word feature, a feature that includes all the words occurring in each sentence. By using this feature, it is hypothesized that informational support and emotional support can be differentiated by words that occur in sentences. For example, a sentence with words "love" and "hug" is statistically more likely to provide emotional support, while the words "website" and "physician" are more probable to occur in sentences expressing information support. Some words such as "is", "has", "that" are deemed as high occurrence frequency words in both support categories and do not help differentiate message contents of different support types. This type of words, which is also called "stopwords," are excluded from this feature. 2. Sentence length. It is supposed that, statistically, informational and emotional supports differ in their general sentence length. 3. Unified Medical Language System (UMLS)[1] semantic type. UMLS is an online meta-thesaurus of controlled vocabularies of medical terminologies. Each entry in UMLS has been assigned one of the total 134 semantic types such as "Disease or Syndrome", "Mental Process", or "Therapeutic or Preventive Procedure". In the support classification task, we assume that sentences providing different types of support tend to contain words belonging to different semantic types in UMLS and with different frequency. For instance, a word with semantic type "Disease or Syndrome" such as "hypertension" is more likely to occur in informational support sentences, and a word with a semantic type "Mental Process" such as "happy" is more likely to occur in emotional support sentences.

With these features and classifier type at hand, the machine learning approach can be described as a five-stage process:

Stage 1: Manually Classify Documents for Training

The approach we use to train the classifier is also called the supervised learning approach (Sebastiani 2002) because the machine learning process is supervised by a set of manually classified data. By using this method, the 2174 downloaded training set sentences from the breast cancer support message board are first classified manually, yielding 1545 informational support sentences and 629 emotional support sentences. To create a uniform-distributed training dataset, 629 out of the 1545 informational support sentences are randomly selected to couple with the emotional support sentences, resulting in 1258 training sentences. This also sets the baseline result for this experiment: By always guessing one type of support would result in a 50% classification accuracy rate. These manually classified sentences are then sent to the next stage to pre-process.

Stage 2: Pre-Process

In this stage, sentences for training are sent to the pre-processor to 1. remove the stopwords, because too many stopwords that are not helpful in support classification will decrease the accuracy of the resulting classifier, and 2. stem words occurring in each sentence into its basic form. For example, "walking" is stemmed to "walk," "action" is stemmed to "act." The purpose of stemming is to increase the accuracy of classification. Without stemming, words such as "walk," "walks," "walked," and "walking" are treated as different words, which is unnecessary and redundant, and will negatively affect the trained classifier. We adopt the widely used porter algorithm (Porter 1980) to do the stemming task.

Stage 3: Train the Classifier

In the training process, the pre-processed manually classified sentences are fed to the learner software tool. The learner then extracts the features of the sentences based on the three pre-determined document characteristics

[1] http://www.nlm.nih.gov/research/umls/

mentioned above. With sentential features and the class each of the sentences belongs to as inputs, the SVM algorithm is "supervised" to learn to separate sentences of the two classes, resulting in the SVM classifier being capable of classifying sentences of messages in the online breast cancer support message board into informational or emotional support.

Stage 4: Evaluate the Classifier

When the training process is over, the next step is to evaluate the classifier by checking the accuracy of classification results it generates. A classifier that generates results with higher accuracy means the choices of document features and type of classifier are appropriate, and the classification results are more reliable. The evaluation method used here is cross-validation (Sebastiani 2002). By using cross-validation, the whole training dataset is first evenly divided into pre-determined number of subsets. Each of these subsets is then by turns treated as the (evaluating) testing set classified by the (evaluating) classifier trained with the rest of the training set through the same process described above. The resulting overall accuracy is the average accuracy of the results generated by these evaluating classifiers.

In this study, a 10-fold cross-validation method is used, by which ten evaluating SVM classifiers are trained and tested separately. The result of the automated support classification on the all 1258 training set sentences yields 87.5% of average accuracy, which is much improved over the baseline accuracy rate of 50%. The training process can also be tuned to acquire classifiers generating results with higher accuracy by adopting more sophisticated document characteristics. This issue will be discussed in the last section.

Stage 5: Classify Message Contents with Trained Classifier

Now that the automated support classifier is generated, it can be applied to classify contents of the testing set messages downloaded from the online breast cancer support message board. The advantage of the automated analysis method is that the only human intervention required is the initial manual labeling of the training dataset. Once the classifier is trained, it can be applied to analyze any amount of testing data in the same domain, which also leads to another advantage of this method: a relatively larger volume of message contents can be analyzed, resulting in a more holistic view of entire message board.

In the next section, a case study will be used to demonstrate how the automated support classifier is applied to a widely studied topic – gender differences on types of support sought/provided in online cancer support communities.

Case Study: Online Breast and Prostate Cancer Support Groups

The purpose of this case study is to show that our automated classification method produces results comparable to the traditional manual content analysis methods. To show this we take gender differences in online support communities as the topic of this case study. Previous research has indicated that men and women share different characteristics of communication in terms of behavioral or linguistic patterns (e.g., Coates 2004; Spence and Helmreich 1978; Tannen 1990). These gender differences also manifest in online user behavior (e.g., Boneva et al. 2001; Hargittai and Shafer 2006). In the healthcare domain, researchers of cancer support groups pointed out men and women show different tendencies in types of support sought/provided, online (Gooden and Winefield 2007; Klemm et al. 1999; Seale et al. 2006) or offline (Gray et al. 1996). More specifically, compared to men with cancer, women with cancer are more likely to seek/provide emotional support. On the other hand, men with cancer are more likely to seek/provide informational support when compared to women with cancer. In this case study, the proposed automated support classification method is applied to two large online cancer support message boards – breast cancer and prostate cancer support message boards hosted on the cancer support community mentioned in the previous section. These two types of cancer support groups are widely adopted in the study of gender differences since breast cancer support groups are dominated by women participants and the majority of participants of prostate support groups are men. In addition, breast cancer and prostate cancer have similar age of onset, morbidity and mortality rates, thus providing comparable sources of analysis (Gooden and Winefield 2007; Gray et al. 1996; Klemm et al. 1999; Seale et al. 2006). Furthermore, prior research in these two communities also has focused specifically on informational and emotional support, which is relevant for the current study. Our objective is to check if the results of automated support classification of messages in online breast/prostate cancer support message

boards also support previous results that woman exchange relatively more emotional support and men exchange relatively more informational support.

To do the support classification, first messages from both breast and prostate support message boards are downloaded using RSS feed. There are a total of 10000 messages (55175 sentences) from the breast cancer support message board, spanning a two-year period and 6184 messages (49174 sentences) from the prostate support message board, spanning a eight-year period downloaded. Please note that the 10000 messages from breast cancer support message board are not overlapped with the messages for training the classifier in the previous section and thus are unseen messages. Because in the previous section we have trained the breast cancer support classifier, the classifier is directly applied to the classification task. As to the automated classifier for prostate support message board, we again downloaded and manually labeled 492 messages (3958 sentences, which are disjoined from the 6184 testing set messages) and then chose 529 informational support and 529 emotional support sentences out of the labeled dataset to train the classifier. The resulting classification accuracy using 10-fold cross-validation is 88%.

Now that the breast cancer support classifier and the prostate cancer support classifier are available, they are used to classify those testing set messages, which is an automated process. The summarization of the findings in this case study is listed in table 1 below. As can be seen, compared to message contents in prostate cancer support message board, message contents in breast cancer support message board express relatively higher proportion of emotional support (53% compared to 31% in prostate cancer support message board). On the other hand, the proportion of informational support sentences is relatively higher in the prostate support message board (69% compared to 47% in breast cancer support message board). These results support findings generated through other qualitative methods. It also tells us that in the online breast cancer support message board, the amount of exchanged informational and emotional support are about the same, while in the online prostate cancer support message board, the amount of exchanged informational support is more than twice as exchanged emotional support, signifying the gender difference.

To summarize, this case study illustrates the use of the proposed research method to study gender differences in online support communities. By applying this method to analyze online breast cancer and prostate cancer discussion boards, message contents are classified automatically as either informational support or emotional support. The comparison of the analysis outcomes from the two online discussion boards shows that this case study concludes similarly to previous studies using manual content analysis methods. Different from previous studies in which data was collected from a period of days or months, this method analyzes data spanning over years, result in a more comprehensive view of the online support communities and thus a stronger support for the gender differences.

Table 1. Results of automated support classification of online breast and prostate cancer support community messages

	Num. of Messages	Num. of Sentences	Info. Support Sentences	Emo. Support Sentences	Percentage of Info. Support	Percentage of Emo. Support
Breast Cancer Support Board	10,000	55,175	25,816	29,359	47%	53%
Prostate Cancer Support Board	6,184	49,174	33,737	15,437	69%	31%

Limitations, Implications, and Future Work

In this article, a new method of analyzing messages of online support communities is proposed. This method helps classify online message contents into informational support and emotional support automatically with the results comparable to those acquired from other qualitative data analysis methods. By exempting from human intervention, this method gives timely analysis of messages from online support communities. This feature makes it possible to analyze larger volumes of messages than using other alternative methods, which not only render a more comprehensive view of supports sought/provided in a given online support community, but also enables the quick comparison of multiple online support communities which was a time-consuming and tedious job. Although this method seems promising, there are a couple of limitations requiring further investigation.

First, since this method broadly classifies supports into either of the two classes – informational and emotional support, the analysis results lack details of message contents that can be acquired through existing manual

qualitative analysis methods. For example, using grounded theory to analyze online support message contents, Gooden and Winefield (2007) identified concepts "Facts about the disease" and "Dealing with effects of disease" as sub-categories of informational support, which is more difficult for this computerized method to accomplish and requires more sophisticated design of the classifier training process. Second, the proposed method does not differentiate support seeking and providing, which also results in less detailed insights. Third, by taking the sentence as the unit of analysis without relating them together, this method is unable to capture supports conveyed through multiple sentences. For example, there are cases when a sequence of sentences as a whole conveys a different type of support from what each single sentence literally represents. Last, from the preliminary experiment results, the automated classification generates an accuracy rate of 87.5%, which is good though not great, leaving space for improvement. To conclude, compared to traditional qualitative content analysis methods, although the proposed research method is able to analyze large amounts of data, its generated results contain less detailed insights of message contents and may be less accurate. This shortcoming, however, would expect to draw researchers, particularity from computer science or information system disciplines, to discover better machine learning approaches with the aim to generate analysis results as detailed as human scrutiny. In short, the purpose of the proposed automated classification method is not to replace but to complement existing qualitative methods. It provides a more holistic view of the target online communities and also could used to verify findings generated from traditional methods. Despite these limitations, this method nevertheless opens doors to various research topics and thus has tremendous implications on different research disciplines in the studies of online support communities. This method itself also provides a new way of doing qualitative content analysis. Some of these implications are briefly described below.

Empirically, with this automated message classification method, researchers can study and compare dynamics of support behavior differentiated by various stress events or demographic distributions such as gender and ages on online support communities without much difficulty. Furthermore, researchers can study the trend of informational/emotional supports sought/provided, either by individual users or the whole online support community over a longer period of time. Last, based on the result of automated support classification, many analysis methods can be applied for further understandings. For example, social network analysis can be conducted to compare the support structures of different support types for a given online support community. This study can also be furthered to investigate the recent trend of consumers' involvement of knowledge co-creation with healthcare providers on online support communities (Winkelman and Choo 2003) by studying the interplay of structure of social network formed, type of support sought/provided and the knowledge creation processes, which is also the main theme of our ongoing project.

In the aspect of designing online support web sites, online support communities such as PatientsListMe.com allow their members to maintain and visualize their health status chronologically as charts in their user profiles. By using this automated support classification method, the trend of types of support sought/provided can also be analyzed and visualized to help users keep track of their activities. In the community level, a chart depicting different types of support provided in each discussion board or discussion thread can also help support seekers/providers and lurkers have the preliminary idea of the intended discussion board/thread.

Technically, this method can be the onset of a new research topic. Efforts can be taken to devise different machine learning strategies to generate more sophisticated text classifiers to help acquire more detailed support information such as increasing the accuracy of the automatic analysis, more fine-grained support classification, or using a different unit of analysis such as paragraph or term level analysis. We are also working on enhancing the accuracy and capabilities of the classifier by incorporating more document content features such as syntactic features, and by using a more fine-grained support classification such as the five types of support proposed by Cutrona and Russell (1990).

Last and probably a more important implication is that the proposed method alone provides qualitative researchers an alternative choice when analyzing document contents. The applications of this method are not just restricted to study online support message boards. Any document content that needs to be analyzed and categorized according to its content can benefit from the advantages brought about by this method – capable of analyzing large amounts of data while relieving from human intervention. This method thus is particularly useful and a necessity when the required amount of data exceeds human abilities to analyze, or needs much effort and time to generate the desired outcome, such as messages spanning multiple years in online support communities presented in this article. As mentioned above, this proposed method has its advantages and limitations and is meant to complement current qualitative research methods. By providing an alternative method choice to researchers, the aim is to facilitate the acquisition of knowledge on the dynamics of human behavior or other human generated contents.

References

Berkman, L. F., and Glass, T. 2000. "Social Integration, Social Networks, Social Support, and Health," in *Social Epidemiology*, L. F. Berkman and I. Kawachi (eds.), New York: Oxford Press, pp. 137-173.

Boneva, B., Kraut, R., and Frohlich, D. 2001. "Using E-Mail for Personal Relationships: The Difference Gender Makes," *American Behavioral Scientist* (45:3), pp. 530-549.

Boyatzis, R. E. 1998. *Transforming Qualitative Information: Thematic Analysis and Code Development*, London: Sage.

Braithwaite, D. O., Waldron, V. R., and Finn, J. 1999. "Communication of Social Support in Computer-Mediated Groups for People with Disabilities," *Health Communication* (11:2), pp. 123-151.

Chang, C.-C., and Lin, C.-J. 2001. "LIBSVM: A Library for Support Vector Machines," Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Coates, J. 2004. *Women, Men and Language: A Sociolinguistic Account of Gender Differences in Language*, (3rd ed.), London: Pearson ESL.

Cobb, S. 1976. "Social Support as a Moderator of Life Stress," *Psychosomatic Medicine* (38), pp. 300-314.

Cohen, S., and Wills, T. 1985. "Stress, Social Support, and the Buffering Hypothesis," *Psychological Bulletin* (98), pp. 310-357.

Corbin, J., and Strauss, A. C. 2007. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*, (3rd ed.), London: Sage.

Coulson, N. S. 2005. "Receiving Social Support Online: An Analysis of a Computer-Mediated Support Group for Individuals Living with Irritable Bowel Syndrome," *CyberPsychology & Behavior* (8:6), pp. 580-584.

Cutrona, C. E., and Russell, D. W. 1990. "Type of Social Support and Specific Stress: Theory of Optimal Matching," in *Social Support: An Interactional View,* B. R. Sarason, I. G. Sarason, and G. R. Pierce (eds.), New York: John Wiley, pp. 319-366.

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

De Wever, B., Schellens, T., Valcke, M. and Van Keer, H. 2006. "Content Analysis Schemes to Analyze Transcripts of Online Asynchronous Discussion Groups: A Review," *Computers & Education* (46:1), pp. 6-28.

Eysenbach, G., and Till, J. E. 2001. "Ethical Issues in Qualitative Research on Internet Communities," *British Medical Journal* (323:7321), pp. 1103-1105.

Fox, S., and Fallows, D. 2003. "Internet Health Resources: Health Searches and Email Have Become More Commonplace, but There is Room for Improvement in Searches and Overall Internet Access," *Pew Internet & American Life Project*.

Gooden, R. J., and Winefield, H. R. 2007. "Breast and Prostate Cancer Online Discussion Boards: A Thematic Analysis of Gender Differences and Similarities," *Journal of Health Psychology* (12:1), pp. 103-114.

Gray, R., Fitch, M., Davis, C., and Phillips, C. 1996. "Breast Cancer and Prostate Cancer Self-Help Groups: Reflections on Differences," *Psycho-Oncology* (5:2), pp. 137-142.

Hargittai, E., and Shafer, S. 2006. "Differences in Actual and Perceived Online Skills: The Role of Gender," *Social Science Quarterly* (87:2), pp 432-448.

Heaney, C. A., and Israel, B. A. 2002. "Social Networks and Social Support," in *Health Behavior and Health Education: Theory, Research, and Practice*, K. Glanz, B. K. Rimer, and F. M. Lewis (eds.), San Francisco, CA: Jossey-Bass, pp. 185-209.

Joachims, T. 1998. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," in *Proceedings of the European Conference on Machine Learning (ECML'98)*, Chemnitz, Germany, pp. 137-142.

Klemm, P., Reppert, K., and Visich, L. 1998. "A Nontraditional Cancer Support Group: The Internet," *Computers in Nursing* (16:1), pp. 31-36.

Klemm, P., Hurst, M., Dearholt, S. L., and Trone, S. R. 1999. "Gender Differences on Internet Cancer Support Groups," *Computers in Nursing* (17:2), pp. 65-72.

Lakey, B., and Cohen, S. 2000. "Social Support Theory and Measurement," in *Social Support Measurement and Intervention: A Guide for Health and Social Scientists*, S. Cohen, L. G. Underwood, and B. Gottlieb (eds.), New York: Oxford University Press, pp. 29-52.

Lieberman, M. A., and Goldstein, B. A. 2005. "Self-Help On-Line: An Outcome Evaluation of Breast Cancer Bulletin Boards," *Journal of Health Psychology* (10:6), pp. 855-862.

Maloney-Krichmar, D., and Preece, J. 2005. "A Multilevel Analysis of Sociability, Usability, and Community Dynamics in an Online Health Community," *ACM Transactions on Computer-Human Interaction* (12:2), pp. 1-32.

Nigam, K., Mccallum, A. K., Thrun, S., and Mitchell, T. 2000. "Text Classification from Labeled and Unlabeled Documents using EM," *Machine Learning* (39), pp. 103-134.

Pang, B., and Lee, L. 2008. "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval* (2:1-2), pp. 1-135.

Pfeil, U. 2009. "Online Support Communities," in *Social Computing and Virtual Communities*, P. Zaphiris and C. S. Ang (eds.), Chapman & Hall, pp. 121-150.

Pfeil, U., and Zaphiris, P. 2007. "Patterns of Empathy in Online Communication," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, San Jose, California, pp. 919-928.

Pfeil, U., and Zaphiris, P. 2009. "Investigating Social Network Patterns within an Empathic Online Community for Older People," *Computers in Human Behavior* (25:5), pp. 1139-1155.

Porter, M. F. 1980. "An Algorithm for Suffix Stripping," *Program* (14:3), pp. 130-137.

Riloff, E., and Wiebe, J. 2003. "Learning Extraction Patterns for Subjective Expressions," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'03)*, Sapporo, Japan, pp. 105-112.

Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E. 1998. "A Bayesian Approach to Filtering Junk E-Mail," in *AAAI-98 Workshop on Learning for Text Categorization, AAAI Technical Report WS-98-05*, Madison, WI, pp. 55-62.

Schaefer, C., Coyne, J. C., and Lazarus, R. S. 1981. "The Health-Related Functions of Social Support," *Journal of Behavioral Medicine* (4:4), pp. 381-406.

Scott, M. 2004. *WordSmith Tools version 4*, Oxford: Oxford University Press.

Seale, C., Ziebland, S., and Charteris-Black, J. 2006. "Gender, Cancer Experience and Internet Use: A Comparative Keyword Analysis of Interviews and Online Cancer Support Groups," *Social Science & Medicine* (62:10), pp. 2577-2590.

Sebastiani, F. 2002. "Machine Learning in Automated Text Categorization," *ACM Computing Surveys* (34:1), pp. 1-47.

Spence, J. T., and Helmreich, R. L. 1978. *Masculinity and Femininity: Their Psychological Dimensions, Correlates, and Antecedents*, Austin: University of Texas Press.

Tannen, D. 1990. *You Just Don't Understand: Women and Men in Conversation*, New York: William Morrow.

Thoits, P. A. 1995. "Stress, Coping, and Social Support Processes: Where are We? What Next?," *Journal of Health and Social Behavior* (Extra Issue), pp. 53-79.

Uden-Kraan, C. F. v., Drossaert, C. H. C., Taal, E., Shaw, B. R., Seydel, E. R., and Laar, M. A. F. J. v. d. 2008. "Empowering Processes and Outcomes of Participation in Online Support Groups for Patients with Breast Cancer, Arthritis, or Fibromyalgia," *Qualitative Health Research* (18:3), pp. 405-417.

Vapnik, V. 1999. *The Nature of Statistical Learning Theory*, (2nd ed.), Springer.

Wellman, B., Salaff, J., Dimitrova, D., Garton, L., Gulia, M., and Haythornthwaite, C. 1996. "Computer Networks as Social Networks: Collaborative Work, Telework, and Virtual Community," *Annual Review of Sociology* (22:1), pp. 213-238.

White, M. H., and Dorman, S. M. 2000. "Online Support for Caregivers: Analysis of an Internet Alzheimer Mailgroup," *Computers in Nursing* (18:4), pp. 168-179.

Winkelman, W. J., and Choo, C. W. 2003. "Provider-Sponsored Virtual Communities for Chronic Patients: Improving Health Outcomes through Organizational Patient-Centered Knowledge Management," *Health Expectations* (6:4), pp. 352-358.