Identifying Emotional and Informational Support in Online Health Communities

Prakhar Biyani[1] **Cornelia Caragea**[2] **Prasenjit Mitra**[1] **John Yen**[1]
(1) College of Information Sciences and Technology, The Pennsylvania State University, USA
(2) Department of Computer Science and Engineering, University of North Texas, USA
pxb5080@ist.psu.edu, ccaragea@unt.edu, {pmitra,jyen}@ist.psu.edu

Abstract

A large number of online health communities exist today, helping millions of people with social support during difficult phases of their lives when they suffer from serious diseases. Interactions between members in these communities contain discussions on practical problems faced by people during their illness such as depression, side-effects of medications, etc and answers to those problems provided by other members. Analyzing these interactions can be helpful in getting crucial information about the community such as dominant health issues, identifying sentimental effects of interactions on individual members and identifying influential members. In this paper, we analyze user messages of an online cancer support community, Cancer Survivors Network (CSN), to identity the two types of social support present in them: *emotional* support and *informational* support. We model the task as a binary classification problem. We use several generic and novel domain-specific features. Experimental results show that we achieve high classification performance. We, then, use the classifier to predict the type of support in CSN messages and analyze the posting behaviors of regular members and influential members in CSN in terms of the type of support they provide in their messages. We find that influential members generally provide more emotional support as compared to regular members in CSN.

1 Introduction

Increasingly more people turn to online health communities (OHCs) to seek social support during their illnesses (LaCoursiere, 2001; Beaudoin and Tao, 2007). When people suffering from a serious disease such as cancer or AIDS *interact* with other people who have experienced similar medical conditions, they feel emotionally supported. In addition, through these interactions, people can obtain important information about the disease, e.g., about various medications, symptoms, and side-effects. Although authoritative health-related web sites contain the information they search for, obtaining this information directly from people in OHCs adds substantial value to it. Previous studies showed that obtaining social support in OHCs can help people feel better (Dunkel-Schetter, 1984; Maloney-Krichmar and Preece, 2005; Beaudoin and Tao, 2007; Vilhauer, 2009; Qiu et al., 2011).

As a result of online interactions in OHCs, a huge volume of user-generated content exists today on various issues/problems related to specific diseases. This content comprises of important information such as people's experiences with diseases, recommendations and feedbacks about certain medications or medical procedures, and emotional support in the form of encouragement, sympathy, and success stories. Mining this content can prove to be very useful in obtaining crucial insights into community dynamics such as identifying dominant health issues or the effects of social support on community members, identifying influential members, as well as designing smart information retrieval systems for users.

In this study, we focus on an online cancer support community, the Cancer Survivors Network[1] (CSN) of the American Cancer Society. We analyze user messages of CSN to identify the two most important types of social support present in them: informational and emotional support (Davison et al., 2000). Emotional support comprises of seeking or providing caring/concern, understanding, empathy, sympathy,

[1] http://www.csn.cancer.org

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 827–836, Dublin, Ireland, August 23-29 2014.

encouragement, affirmation and validation. In contrast, informational support comprises of seeking or providing knowledge such as advice, referrals, and suggestions (Bambina, 2007). We further explore the relation between the type of support present in messages and users' influence in the community.

Identifying the type of support in user messages in an OHC can potentially be used in many important applications including the following:

1. **Identify influential members in OHCs**: Every community has a set of members who influence (a much larger set of) other members in the community. These members are called leaders or influential members. The attributes of a leader in a community depends upon the community's nature (QA, Twitter, OHC, forum, blogsite, etc.). For example, high activity may not be an indicator of high influence in the blogosphere (Agarwal et al., 2008) and high popularity does not necessarily imply influence in Twitter (Romero et al., 2011). In OHCs, bringing positivity in the community and answering members' concerns effectively by posting messages that contain certain type of support (informational or emotional) may be an indicator of influence.
2. **Improve information search in OHCs**: Interactions in OHCs contain valuable information in the form of people's experiences, advice, referrals, pertaining to diseases, medications, side-effects, etc. Users embed this information often in messages containing other types of support, of which emotional support constitutes a major part. To efficiently search OHCs for this information, it must be separated from emotional support. Hence, identifying the type of support in user messages can help improve search and retrieval in OHCs.
3. **Understand social relationships in OHCs**: Emotional support is one of the dimensions of social tie strength between members in a social network (Gilbert and Karahalios, 2009). Previous studies have shown that members receiving emotional support in OHCs are more likely to remain in the community for a longer period of time as compared to members receiving informational support (Wang et al., 2012). Identifying emotional and informational support can help understand the social dynamics of an OHC. For example, it would be interesting to see if there is a correlation between the social tie strength of members and the type of support present in their interactions.

Hi X, I had a bilateral with radical on the right and prophylactic on the left. I think all you can do is gentle exercises to strengthen your back (yoga). There are also herbal painkillers that work well too. I just tolerate the pain and consider it a signal of my new limit and go down to rest. You want to talk, anytime! We are all there with you.

Table 1: A user message. Sentences in grey and black fonts are informational and emotional, respectively.

We model the task of identifying the two types of supports as a binary classification problem. Specifically, we classify each sentence in a user message as containing either emotional or informational support [2]. Table 1 shows a user message containing emotional and informational supports. We use several features computed from sentences of messages such as unigrams, part-of-speech tags, lexicon-based features and word patterns for the classification. After building the classification model, we predict the amounts of the two supports in all CSN messages and explore the following research question:

RQ: Do influential members of CSN post one of the two types of supports significantly more compared to regular members?

We analyze messages posted by regular members and messages posted by certain members, identified as *influential* by the CSN community managers and two staff members who monitor the contents of the CSN on a full time basis, for the type of support (informational and emotional) present in them. Using the classification model, we calculate the amounts of the two supports posted by influential members and regular members and compare them across the two populations (For details, see Section 3.1).

Previous works on analyzing social support in OHCs have mainly been in the field of social science (Eriksson and Lauri, 2000; Rodgers and Chen, 2005; Høybye et al., 2005; Pfeil and Zaphiris, 2007; Beaudoin and Tao, 2007; Buis, 2008; Han et al., 2011). These works used manual techniques for identifying the type of support in user messages and hence, are limited to a small number of messages as

[2] Although a sentence may belong to both the classes, we did not find such cases in our data.

compared to the real world data. In contrast, the current work builds machine learning classifiers that can automatically predict the type of support in messages. Also, to the best of our knowledge, there have been no reported works on analyzing the relationship between users' influence and the type of support present in their messages in OHCs. Next, we review related works.

2 Related Work

Many studies in social science have focused on analyzing social support in user messages of OHCs (Coursaris and Liu, 2009; Han et al., 2011; Pfeil and Zaphiris, 2007), finding impacts of social support on users (Eriksson and Lauri, 2000; Rodgers and Chen, 2005; Buis, 2008; Høybye et al., 2005), identifying information needs of users in OHCs (Rozmovits and Ziebland, 2004), etc. Among various types of social supports, emotional support and informational support have received major attention. In this section, we first review social science works on analyzing online social support, discuss works on identifying the type of social support, and, finally, compare the current problem with **subjectivity analysis**.

LaCoursiere (2001) presented an integrated theory conceptualizing online social support. She defined three channels through which online social support occurs: 1) *perceptual*: individual feeling the need of social support arising due to emotional states such as stress, etc, 2) *cognitive*: individual seeking information about certain medical entities such as procedures, medication, etc, 3) *transactional*: individual evaluating the received social support. In our case, these channels correspond to emotional support and informational support. Høybye et al. (2005) conducted a qualitative study to analyze the effects of online social support by interviewing women with breast cancer who used an online support group and found that the women were empowered by the exchanges of knowledge and experience within the online support group. Rodgers et al. (2005) conducted a longitudinal content analysis of messages of participants in a breast cancer discussion board to analyze changes in affect/sentiment of the participants towards breast cancer and found that a positive shift in sentiment occurred over the period of time. Pfeil and Zaphiris (2007) analyzed messages of SeniorNet forum to extract language patterns used to provide empathic support. Budak and Agrawal (2013) interviewed participants of group chats in Twitter and found that informational support is more important than emotional support in educational Twitter chats.

All the above works used manual methods of data preparation such as interviews with users of support groups, manual coding of messages to identify emotional and informational support and performed further qualitative and/or quantitative analyses based on that data. Since, manual methods have serious limitations in terms of scalability, the number of messages used for analysis in these studies is too small compared to the real world data which contains millions of messages. To address these limitations, we develop automatic methods for identifying the type of support in user messages in an online cancer support group using machine learning. We develop a classifier that learns on a smaller set of manually labeled messages and makes predictions on a much larger set of messages with a very high accuracy.

A recent work by Wang et al. (2012) is close to our work. They used a linear regression model to predict the amount of informational and emotional supports present in messages of a cancer forum. For a test message, the trained model predicts the amount of the two supports on a scale of $1 - 7$. Since a message may contain both types of support, it is generally difficult for human annotators to assess the amount of each support in an entire message on a particular scale for model training. In contrast, we label each sentence as belonging to either informational or emotional support class and identify the two types of support at sentence level in messages (using binary classification). Note that it is much easier and less ambiguous for a human annotator to identify the type of support present in a sentence (of a message) compared to giving a score to an entire message based on the amount of the two supports present in it.

Relationship with Subjectivity Analysis: Subjectivity analysis is an active area of research in computational linguistics. It essentially deals with separating subjective parts (e.g., expressing opinion, emotion, speculation and other private states of mind) from objective parts (presenting facts, verifiable information) of a text (Wiebe et al., 1999; Biyani et al., 2012a). It has been widely used in applications like opinion mining from product reviews (Liu, 2010), community question-answering (Li et al., 2008a; Stoyanov et al., 2005a; Somasundaran et al., 2007), summarization (Carenini et al., 2006; Seki et al., 2005), and finding opinionative threads in online forums (Biyani et al., 2014; Biyani et al., 2012b; Biyani et al., 2013a). Though the current work has some relation with subjectivity analysis in the sense that both are text classification, there are important differences between the two problems. The two classes in subjectivity analysis (subjective and objective) are different from the two types of support that we identify. While emotional support is subjective in nature, informational support is not necessarily objective as it also contains opinions of users. Also, social support in OHCs encompasses several types of supports such as understanding, caring, concern, sympathy, empathy, knowledge about medications, etc. which are generally not provided by users in other sites such as product reviews, question-answering sites, etc. These differences make the two problems different in both the nature and the approaches that can be used to address them. For example, we use certain word patterns to identify sympathy and affirmation and use the presence of terms related to cancer medications, procedures and side-effects for computing features for classification. These features have not been used in subjectivity classification.

3 Problem Formulation

Online health communities provide social support to its members of which emotional and informational supports constitute a major part and have received major attention as compared to other supports such as companionship, community building, network support, etc. (Bambina, 2007; Meier et al., 2007; Himle et al., 1991; Wang et al., 2012; Pfeil and Zaphiris, 2007). We focus on the two supports and follow their definitions as given by Bambina (2007) in their study of social supports expressed in a cancer support group. They define *emotional* messages as the messages that have the following supports: caring/concern, understanding, empathy, sympathy, encouragement, affirmation and validation. *Informational support* is defined as providing advice, knowledge and referrals. Since a user message often contains a mixture of these supports, we identify the two supports at sentence level. Table 1 contains a user message with sentences marked with the type of support in them. Specifically, given a sentence s, in a user message, we want to classify it into one of the two classes: emotional support or informational support. We use machine learning methods for classification. After training the classifier, we use it to predict the type of support in the sentences of user messages in CSN and address our research question outlined in Section 1. We present the details of the features used for classification in Section 3.2.

3.1 Research Question

To address the research question **(RQ)**, we need to compute the amounts of the two supports in the messages of regular and influential members and then compare the two amounts. Let u denote a user and M be the set of messages posted by her such that $M = \{m_1, m_2, ...m_p\}$ where p is the total number of messages in the set M. For a message $m_k \in M$, we compute its *emotional index*, $e_{uk} = n_{ek}/(n_k)$ where n_{ek} and n_k are the number of sentences containing emotional support and the total number of sentences in m_k. Since a sentence can belong to either emotional support or informational support class, informational index of m_k, $i_{uk} = 1 - e_{uk}$. The overall emotional index of u (e_u) is the average of the emotional indices of her messages: $e_u = \frac{1}{p}\sum_{k=1}^{p} e_{uk}$. The informational index of u, $i_u = 1 - e_u$. Since, the informational index can be derived from emotional index, we compute only emotional indices for all regular and influential members and compare them between the two user populations (regular and influential). We compute the emotional indices of regular members, E_R, and emotional indices of influential members, E_I. We compare the means of the two populations of emotional indices (μ_{Re} and μ_{Ie}) and test the null hypothesis (H_0) and the alternate hypothesis (H_1) as follows:
H_0: The two populations have equal means, i.e., $\mu_{Re} - \mu_{Ie} = 0$.
H_1: The two populations have significantly different means , i.e., $\mu_{Re} - \mu_{Ie} \neq 0$.

For one of the population indices to be significantly more than the other, we should have the null hypothesis rejected. We use one-sided t-test to conduct hypothesis testing and report the results in Section 4.5. Next, we discuss the features used in the classification.

3.2 Features for Classification

3.2.1 Words and POS tags

Words and their part-of-speech tags capture basic lexical properties of text and have been extensively used in text classification problems such as subjectivity classification and sentiment classification (Li et al., 2008b; Yu and Hatzivassiloglou, 2003; Biyani et al., 2013b). We use frequency of words and their POS tags in a sentence as features in our classification model.

3.2.2 Lexicon-based Features

Emotional support expresses caring, concern, sympathy, and other kinds of sentimental support whereas informational support provides knowledge about cancer medications, cancer reports, referrals, and other kinds of information (Bambina, 2007). Due to this difference in the nature of these supports, a sentence expressing emotional support is likely to contain emotional words which are subjective in nature and a sentence containing informational support is likely to have cancer-related keywords such as drug names, names of cancer procedures, etc. To capture this difference, we use frequencies of subjective words and cancer-related keywords as features. Specifically, we design five features to code frequencies of weak subjective words (**numWeak**), strong subjective words (**numStrong**), cancer drugs (**numDrug**), side-effects of cancer medications (**numSide**), and cancer procedures (**numProc**) respectively in a sentence. We use the subjectivity lexicon compiled from the MPQA corpus (Stoyanov et al., 2005b) to get weak and strong subjective words. We compile lexicon of cancer drugs [3], and CSN staff members helped get a list of side-effects and cancer procedures. Some of the side-effects of cancer medications are hair loss, neuropathy, fatigue, fibrosis, etc.

3.2.3 Linguistic Features

We analyzed user messages to find patterns that are expressive of emotional and informational support. We found that members, generally, use certain word patterns to express similar feelings. For example, to provide affirmation and sympathy, people use positive verbs such as *know, feel, understand, sense, support*, etc. in patterns <*I $posVerb*> and <*I $aux $posVerb*>, where *$posVerb* is a positive verb and *$aux* is an auxiliary verb from the set {can, could, do, would, will, may}. Some people use *"We"* instead of *"I"* in their messages to provide support such as "**we understand** *what you are going through*". To take into account such cases, we use the same patterns by replacing *"I "* with *"We"*. Hence, we get four patterns for emotional support. For providing informational support, people often use patterns such as <*You $advice*>, <*I $opinion*>, <*I $aux $opinion*> to provide advice and opinions. *$advice* is an auxilliary verb from the set {should, must, need, might}, *$opinion* is an opinion verb from the set {recommend, advise, suggest, advocate, request}, and *$aux* is an auxilliary verb. People also give information about their experiences using patterns such as <*I too*>, <*I also*> and <*I $pastVerb*> to tell their own experiences related to similar problems as that of the support seeker where *$pastVerb* is a past tense verb such as *underwent, undergone, experienced, had, found*, etc. So, we get six patterns for informational support. We design two features (**IsEmPattern** and **IsInPattern**) to encode presence (1) or absence (0) of the two types of patterns.

For a sentence, we also use its number of words (**numWords**) and its type, question sentence (**IsQues**) and/or exclamatory sentence (**isExclaim**), as features. To identify question sentences, we see if a sentence starts with any of the 5W1H words *(what, why, who, when, where, how)* or words in the set {do, does, did} or ends with a question mark.

4 Experiments

We now describe our data and the experimental setting, and present our results.

4.1 Data Preparation

We use data from a popular online cancer support community, the Cancer Survivors' Network (CSN), developed and maintained by the American Cancer Society. CSN is an online community for cancer patients, cancer survivors, their families and friends. Its features are similar to many online forums with dynamic interactive medium such as chat rooms, discussion boards, etc. Members of CSN post in discussion boards for seeking and sharing information about cancer related issues and for seeking and providing emotional support. To conduct our experiments, we used user messages in the discussion threads of the Breast Cancer sub forum of CSN that were posted between June 2000 to June 2012. Breast cancer is the largest among all the sub-forums of CSN. A dataset of 250, 868 messages posted by 5516 users in 22, 297 discussion threads is used in this study.

To prepare the evaluation dataset for classification experiments, we randomly sampled 240 messages from 27 discussion threads. Since, our focus is on the messages that provide support, we do not consider

[3] http://www.cancer.gov/cancertopics/druginfo/alphalist

messages posted by thread starters in discussion threads as they seek support. We took help of three human annotators to tag all the sentences of all the messages in one of the two support classes. First, two annotators tagged all the sentences. The percentage agreement between them was 89%. For the remaining 11% sentences, majority vote was taken with the help of the third annotator. Following this tagging scheme, we obtained a total of 1066 sentences with 390 sentences in the informational support class and 676 sentences in the emotional support class. In many cases, members only write a few words, e.g., see you, bye, or their names at the end of a message. To deal with these situations, we filter out sentences that have less than four words.

4.2 Experimental Protocol

We experimented with various machine learning algorithms (Naive Bayes, Support Vector Machines, Logistic Regression, Bagging, Boosting, etc.) to conduct our classification experiments. Naive Bayes Multinomial gave the best performance with words & POS tags features, logistic regression with lexicon-based features and AdaBoost (with Decision Stump as the weak learner) with linguistic features. For combining the models built on the three types of features, we used the following three methods:

1. **Feature combination:** Classification model built on the feature set generated by combining the three types of features. It is denoted by **All**. We use Multinomial Naive Bayes for this model.
2. **Average confidence:** Ensemble of the three classifiers built on the three types of features respectively. The final confidence of the ensemble is calculated by taking average of the confidences outputted by the three classifiers. It is denoted by **AllAvgConf**.
3. **Highest confidence:** Similar to the **AllAvgConf** model but the final prediction of the ensemble is taken as the prediction of the most confident classifier of the three classifiers. More precisely, the prediction for an instance is given by the classifier that returns the maximum prediction confidence for one class or the other. It is denoted by **AllMostConf**.

We used Weka data mining toolkit (Hall et al., 2009) to conduct classification experiments. To evaluate the performance of our classifiers, we used macro-averaged precision, recall and F-1 score. We use F-1 score to compare performances of two classifiers and used 10-fold cross validation. A naive baseline that classifies all the instances in the majority class will have a macro-averaged precision, recall and F-1 score of 0.402, 0.634 and 0.492, respectively.

Model	**Precision**	**Recall**	**F-1**
Emotional support class			
Words & POS tags	0.855	0.858	0.857
Lexicon-based features	0.722	0.836	0.775
Linguistic features	0.698	0.837	0.761
All	0.862	0.861	0.862
AllAvgConf	0.848	0.893	0.87
AllMostConf	0.851	0.911	**0.88**
Informational support class			
Words & POS tags	0.753	0.749	0.751
Lexicon-based features	0.608	0.441	0.511
Linguistic features	0.569	0.372	0.45
All	0.76	0.762	0.761
AllAvgConf	0.797	0.723	0.758
AllMostConf	0.825	0.723	**0.77**
Overall			
Words & POS tags	0.818	0.818	0.818
Lexicon-based features	0.68	0.691	0.678
Linguistic features	0.651	0.667	0.647
All	0.825	0.825	0.825
AllAvgConf	0.829	0.830	0.83
AllMostConf	0.841	0.842	**0.84**

Table 2: Classification results.

4.3 Classification Results

Table 2 presents the results of the support classification experiments. The table reports precision, recall and F-1 score of different classification models for the individual classes and the overall result. Words & POS tags are the best performing features followed by lexicon-based features and linguistic features. Further, combining all the features (model denoted as “**All**”) improves the performance over individual feature types for both classes. We see that **AllMostConf** model is the best performing of all the models, particularly outperforming **All** and **AllAvgConf** models. This observation suggests that the three classifiers built on the three features types have different knowledge. For some instances, a particular classifier is more confident than the rest while for other instances, other classifiers are more confident. Hence, we see that taking prediction of the most confident classifier gives the best performance. It is interesting to note that combining the three classifiers' knowledge in this fashion is more effective than simply combining all the three types of features and train a single classifier on the combined feature set. We also note that all the models have better performance for the emotional support class than for the informational support class. This can be caused by the fact that there are significantly more number of instances in the former class and, hence, more patterns to learn for the class.

4.4 Informative Features

Next, we study the importance of individual features by measuring their chi-squared statistic with respect to the class variable. We, first, study the word features and then present rankings of the other types of features. Figure 1 shows a cloud of top 26 most informative words. The size of a word is proportional to its chi-squared statistic, i.e., bigger a word, more informative it is. We see that cancer specific keywords such as herceptin, tamoxifen, chemo, dose, stage, etc and words conveying emotions such as good, hope, glad, pain, hugs, etc are highly informative for the support classification. Since, chi-square method gives feature ranking for the class variable and not for individual classes, we compute word rankings for individual classes using $tf - idf$ scores of words. Specifically, for a term t and a class c, we compute the term frequency of t by counting its number of occurrences in the instances (sentences) belonging to c and multiply the term frequency with the inverse document frequency of t (calculated from the entire corpus) to get the $tf - idf$ score of t for c. Using this method, we calculated $tf - idf$ scores for all the words and ranked them according to their scores for the two classes. Figure 2 shows top ten $tf - idf$ ranked keywords for the two classes. We see that cancer-related keywords and words expressing emotions are among the top ten most informative words for the informational and the emotional support classes respectively. We also note that most of the top ten words for the two classes in Figure 2 are in the word cloud of the top 26 words computed using chi-squared method except "keep" for the emotional support class and "after", "first", "because" and "cancer" for the informational class. These words have semantic relationships with the classes. For example, "keep" is often used by support providers in phrases such as "**keep** *you in prayers*", "*may god* **keep** *you in good health*", etc to provide emotional support and "after" and "first" are used in the context of providing one's own experience related to cancer procedures, medications, etc such as "**After** *my* **first** *chemo, I did not feel light*".

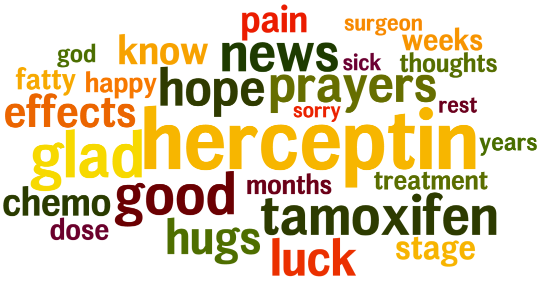

Figure 1: Top 26 words ranked by Chi-squared test.

Emotional support	Informational support
good	chemo
know	after
glad	radiation
news	first
hope	herceptin
keep	treatment
prayers	tamoxifen
luck	cancer
hugs	because
better	pain

Figure 2: Top ten words for the two classes ranked using tf-idf scheme.

We, now, discuss the ranking of non-word features: POS tags, lexicon-based and linguistic features. The chi-squared ranking for the lexicon-based and linguistic features is as follows: numStrong $>$ numWords $>$ isExclaim $>$ numDrug $>$ numSide $>$ numProc $>$ isInPattern $>$ isEmPattern $>$ numWeak $>$ isQues. The features on the right side of $>$ have higher rank than those on the left side. We see that the number of strong subjective words in a sentence is the most informative feature followed by number of words in a sentence. Among cancer-related terms, drug names are more informative than side-effects and cancer procedures. Also, informational support word patterns are more informative than word patterns capturing emotional support. It is interesting to note that isQues is the least informative feature, maybe due to the fact that, while providing support, people generally do not ask questions. The top 5 most informative POS tags are: cardinal number (CD) followed by singular noun (NN), participle verb (VBN), past tense verb (VBD) and preposition (IN).

4.5 Influence versus Support type

CSN managers provided a list of 62 influential members (IMs) for the breast cancer forum. IMs posted a total of $340,147$ sentences in $85,244$ messages and regular members posted $825,651$ sentences in $165,624$ messages in the breast cancer forum. As described in Section 3, we conduct statistical hypothesis testing on the two populations of emotional indices (regular members and IMs) to understand if there is a significant difference in their posting behaviors in terms of providing one of the two supports more often than the other. To test our hypothesis, we conducted one sided t-test on the two populations. We found that the mean of emotional indices of IMs (0.713) is significantly larger than that of the regular members (0.542). We also note that the posting behavior of regular members in CSN follows a power law distribution with most of the members posting very few messages ($mode = 1$, $median = 2$, $mean = 30$) and only a few members posting very many messages. To verify that this behavior does not have impacts on our hypothesis testing, we conducted three more t-tests between the two populations using a threshold on the number of messages that a member has posted. We used three threshold values on the number of messages: 1, 2, and 30 (as mode, median and mean values). For all the three t-tests, the null hypothesis was rejected at p-value< 0.001, suggesting that IMs posted significantly more emotional support than regular members. The values of Mean Emotional Indices corresponding to the three thresholds are 0.715, 0.719 and 0.746 for influential members and 0.564, 0.581 and 0.646 for regular members respectively.

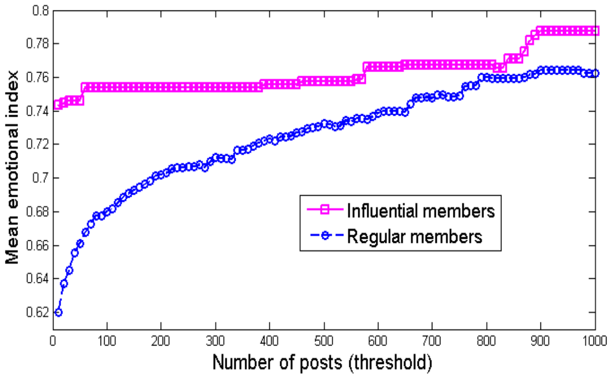

Figure 3: Plot showing the change in mean emotional indices of influential members (pink) and regular members (blue) with the threshold on the number of messages posted by them.

In our analysis, we observed an interesting behavior. As we increased the threshold, the mean of emotional indices also increased. To further investigate this finding, we plotted the means of emotional indices of regular members and IMs as the function of the threshold on the number of messages posted by them. We increased the threshold from 10 to 1000 in steps of 10. Figure 3 reports the finding. We see that the mean of emotional indices of regular members increase with the threshold suggesting that more active members post more emotional support as compared to the less active members. We also see that the mean of emotional indices of IMs is higher than that of regular members for all the thresholds. These interesting observations can be helpful in analyzing behavior of influential members in OHCs.

5 Acknowledgments

We would like to thank Iulia Bivolaru for her help with data preparation. This material is based upon work supported by the National Science Foundation under Grant No. 0845487.

6 Conclusion and Future Work

We identified two types of social support, emotional and informational, provided in user messages of an online cancer support community using machine learning classification models. We used three types of features and got the best results by using ensemble of the three classifiers built on the three individual feature types. Our models achieved strong results with over 80% F-1 score. We also found that influential members provide significantly more emotional support to the community as compared to regular members. The finding can be helpful in identifying properties of influential members in online health communities. In future, we plan to analyze effects of the two types of supports on OHCs' dynamics and use it to improve information search in OHCs.

References

Nitin Agarwal, Huan Liu, Lei Tang, and Philip S Yu. 2008. Identifying the influential bloggers in a community. In *Proceedings of the 2008 international conference on web search and data mining*, pages 207–218. ACM.

Antonina Bambina. 2007. *Online social support: The interplay of social networks and computer-mediated communication*. Cambria press.

Christopher E Beaudoin and Chen-Chao Tao. 2007. Benefiting from social capital in online support groups: An empirical study of cancer patients. *CyberPsychology & Behavior*, 10(4):587–590.

Prakhar Biyani, Sumit Bhatia, Cornelia Caragea, and Prasenjit Mitra. 2012a. Thread specific features are helpful for identifying subjectivity orientation of online forum threads. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 295–310.

Prakhar Biyani, Cornelia Caragea, Amit Singh, and Prasenjit Mitra. 2012b. I want what i need!: analyzing subjectivity of online forum threads. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 2495–2498.

Prakhar Biyani, Cornelia Caragea, and Prasenjit Mitra. 2013a. Predicting subjectivity orientation of online forum threads. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 109–120.

Prakhar Biyani, Cornelia Caragea, Prasenjit Mitra, Chong Zhou, John Yen, Greta E Greer, and Kenneth Portier. 2013b. Co-training over domain-independent and domain-dependent features for sentiment analysis of an online cancer support community. In *ASONAM*, pages 413–417. ACM.

Prakhar Biyani, Sumit Bhatia, Cornelia Caragea, and Prasenjit Mitra. 2014. Using non-lexical features for identifying factual and opinionative threads in online forums. *Knowledge-Based Systems*.

Ceren Budak and Rakesh Agrawal. 2013. On participation in group chats on twitter. In *Proceedings of the 22nd international conference on World Wide Web*, pages 165–176. International World Wide Web Conferences Steering Committee.

Lorraine R Buis. 2008. Emotional and informational support messages in an online hospice support community. *Computers Informatics Nursing*, 26(6):358–367.

G. Carenini, R. Ng, and A. Pauls. 2006. Multi-document summarization of evaluative text. In *EACL*, pages 305–312.

Constantinos K Coursaris and Ming Liu. 2009. An analysis of social support exchanges in online hiv/aids self-help groups. *Computers in Human Behavior*, 25(4):911–918.

Kathryn P Davison, James W Pennebaker, and Sally S Dickerson. 2000. Who talks? the social psychology of illness support groups. *American Psychologist*, 55(2):205.

Christine Dunkel-Schetter. 1984. Social support and cancer: Findings based on patient interviews and their implications. *Journal of Social Issues*, 40(4):77–98.

Elina Eriksson and Sirkka Lauri. 2000. Informational and emotional support for cancer patients relatives. *European Journal of Cancer Care*, 9(1):8–15.

Eric Gilbert and Karrie Karahalios. 2009. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 211–220. ACM.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. 2009. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Jeong Yeob Han, Dhavan V Shah, Eunkyung Kim, Kang Namkoong, Sun-Young Lee, Tae Joon Moon, Rich Cleland, Q Lisa Bu, Fiona M McTavish, and David H Gustafson. 2011. Empathic exchanges in online cancer support groups: distinguishing message expression and reception effects. *Health communication*, 26(2):185–197.

David P Himle, Srinika Jayaratne, and Paul Thyness. 1991. Buffering effects of four social support types on burnout among social workers. In *Social Work Research and Abstracts*, volume 27, pages 22–27. Oxford University Press.

Mette Terp Høybye, Christoffer Johansen, and Tine Tjørnhøj-Thomsen. 2005. Online interaction. effects of storytelling in an internet breast cancer support group. *Psycho-Oncology*, 14(3):211–220.

Sheryl Perreault LaCoursiere. 2001. A theory of online social support. *Advances in Nursing Science*, 24(1):60–77.

B. Li, Y. Liu, A. Ram, E.V. Garcia, and E. Agichtein. 2008a. Exploring question subjectivity prediction in community qa. In *SIGIR*, pages 735–736. ACM.

Baoli Li, Yandong Liu, and Eugene Agichtein. 2008b. Cocqa: co-training over questions and answers with an application to predicting question subjectivity orientation. In *EMNLP '08*, pages 937–946.

B. Liu. 2010. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing,*, pages 978–1420085921.

Diane Maloney-Krichmar and Jenny Preece. 2005. A multilevel analysis of sociability, usability, and community dynamics in an online health community. *TOCHI*, 12(2):201–232.

Andrea Meier, Elizabeth J Lyons, Gilles Frydman, Michael Forlenza, and Barbara K Rimer. 2007. How cancer survivors provide support on cancer-related internet mailing lists. *Journal of Medical Internet Research*, 9(2).

Ulrike Pfeil and Panayiotis Zaphiris. 2007. Patterns of empathy in online communication. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 919–928. ACM.

Baojun Qiu, Kang Zhao, P. Mitra, Dinghao Wu, C. Caragea, J. Yen, G.E. Greer, and K. Portier. 2011. Get online support, feel better – sentiment analysis and dynamics in an online cancer survivor community. In *SocialComm' 11*, pages 274–281.

Shelly Rodgers and Qimei Chen. 2005. Internet community group participation: Psychosocial benefits for women with breast cancer. *Journal of Computer-Mediated Communication*, 10(4):00–00.

Daniel M Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A Huberman. 2011. Influence and passivity in social media. In *Machine learning and knowledge discovery in databases*, pages 18–33. Springer.

Linda Rozmovits and Sue Ziebland. 2004. What do patients with prostate or breast cancer want from an internet site? a qualitative study of information needs. *Patient education and counseling*, 53(1):57–64.

Y. Seki, K. Eguchi, N. Kando, and M. Aono. 2005. Multi-document summarization with subjectivity analysis at duc 2005. In *DUC*. Citeseer.

S. Somasundaran, T. Wilson, J. Wiebe, and V. Stoyanov. 2007. Qa with attitude: Exploiting opinion type analysis for improving question answering in on-line discussions and the news. In *ICWSM*.

V. Stoyanov, C. Cardie, and J. Wiebe. 2005a. Multi-perspective question answering using the opqa corpus. In *EMNLP 2005*, pages 923–930. ACL.

Veselin Stoyanov, Claire Cardie, and Janyce Wiebe. 2005b. Multi-perspective question answering using the opqa corpus. In *HLT-EMNLP '05*, HLT '05, pages 923–930, Stroudsburg, PA, USA. ACL.

Ruvanee P Vilhauer. 2009. Perceived benefits of online support groups for women with metastatic breast cancer. *Women & health*, 49(5):381–404.

Yi-Chia Wang, Robert Kraut, and John M Levine. 2012. To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 833–842. ACM.

J.M. Wiebe, R.F. Bruce, and T.P. O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *ACL*, pages 246–253. ACL.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics.