Linguistic characteristics of eating disorder questions on Yahoo! Answers – content, style, and emotion

Jung Sun Oh jsoh@pitt.edu
Daqing He daqing@sis.pitt.edu
Wei Jeng wej9@pitt.edu
Eleanor Mattern emm100@pitt.edu
Leanne Bowler lbowler@sis.pitt.edu

University of Pittsburgh
135 North Bellefield Avenue
Pittsburgh, Pennsylvania

ABSTRACT

Social Q&A provides the possibility of looking into how people verbally express their information needs in natural language. In this study, we analyzed linguistic properties of different types of questions on the topic of eating disorders in Yahoo! Answers. Using term frequency analysis, Part-of-Speech (POS) analysis, and sentiment analysis, we examined linguistic content, linguistic style, and emotional expressions in two broad categories of eating disorder questions from Yahoo! Answers – socio-emotional questions and informational questions.

Overall, the results of this study show that the language used in these two categories of questions are substantially different, suggesting the different nature of the needs that underlie these questions. Socio-emotional questions take similar characteristics to personal narratives, focusing on past experiences and emotions. The heavy use of negative emotion words in this question type, along with other distinct linguistic characteristics, suggests that a key motivation of users asking this type of question is to work through their emotions related to the given health issue (eating disorders). On the other hand, informational questions show traits of relatively complex, precise, and objective writing, and reflect much varied interests with regard to the topic of eating disorders.

All in all, this study demonstrates that the combination of simple text analytic techniques reveals much about the linguistic characteristics associated with different kinds of questions, and thereby shed lights on the nature of the needs underneath the questions.

Keywords

Social Q&A, Health Information Behavior, Eating disorders, Language Use Analysis

ASIST 2013, November 1-6, 2013, Montreal, Quebec, Canada.

INTRODUCTION

For the past several years, social Question and Answer (Q&A) has attracted substantial attention from researchers in a variety of fields (Gazan, 2011). The tremendous amount of publicly available question/answer data created by real users holds potential for advancing many areas of research, studies of user information behavior being one of them.

Our research team has been investigating health information seeking behavior of teens and young adults in the context of social Q&A with a particular health topic, eating disorders. In our previous study (Bowler et al., 2012), we took the first step towards understanding teens' use of social Q&A for health information. Through a content analysis of eating disorder questions collected from *Yahoo! Answers*, we identified a range of needs and motivations appearing in the questions, and developed a taxonomy of question types consisting of five overarching themes - *Seeking Information, Seeking Emotional Support, Seeking Communication, Seeking Self-Expression,* and *Seeking Help to Complete a Task*.

Having identified the broad categories of questions, we now proceed to further understand the characteristics of questions in different categories and the needs that underlie them. One approach we are exploring is to apply an array of linguistic analysis methods to examine the words people use in their questions, in order to uncover cognitive, social, and affective aspects of the underlying needs. Previous studies of language use (reviewed in the next section) have shown that the way people use words in natural language is determined, to a large extent, by social, situational, and psychological factors and, therefore, the occurrences and distributions of certain words in texts can be used to probe the writer' situation and state of mind at the time of writing. In this study, we will adopt the same approach and scrutinize the use of words in our eating disorder questions, focusing on three word classes – content words (words representing themes), function words (words concerning linguistic styles, e.g. preposition, conjunction), and emotion words. Through the analysis of their distributions across question categories, we aim to address the following questions: How do questions in different categories differ in

their use of content, style, and emotion words? What can we learn about the underlying needs and motivations of users from the observed patterns of word use? Can we infer cognitive, social and affective aspects of the askers' needs from their question language?

In this paper, we will discuss the background and justification of our approach, and present a case study using a small dataset of eating disorder questions.

LITERATURE REVIEW

Health Information Seeking in Social Media

According to a recent Pew Internet report, seeking health information is among the Internet activities that are "becoming more uniformly popular across all age groups"(Zickuhr, 2010). The survey results show that seeking health information has become the third most popular activity, following email and search engine use.

With the prevalence of online health information seeking, researchers in the medical field started to notice the growing role of social media in health communication (Chou, Hunt, Beckjord, Moser, & Hesse, 2009), and have investigated its impact on patients (Jaloba, 2009; Ressler et al., 2012). At the same time, researchers have attended to the potential value of social media as a source for studying people's behavior and attitudes regarding health-related topics (Chou et al., 2011; Robillard et al., 2013). For instance, Rollibard et al. (2013) conducted a content analysis of questions collected from Yahoo! Answers, in order to find out the kinds of information being sought by public users with regard to a particular health topic, gene therapy, and to assess the public opinions and attitudes expressed towards the topic. They found a wide range of issues of interest to public users related to the topic, including ethical concerns. As a concluding remark, they claimed that user-generated contents in a social media can be a rich source for research into health-information seeking behavior.

This study joins the body of research on health information seeking in social media. The health topic of our interest, eating disorders, is closely tied to a particular user group: teens and adolescents. The tendency of relying heavily on Internet for obtaining health information appears to be more salient in this user group (Rideout, 2001) and especially for sensitive topics like eating disorders. The teens' and adolescents' hesitance to discuss sensitive health issues with those close to them or health professionals has been noted repeatedly by researchers (Ackard & Neumark-Sztainer, 2001; Eysenbach, 2008; Katzman et al., 2010). We may speculate that a significant portion of information seeking regarding eating disorders is likely to take place on Internet or through social media, due to a desire for anonymity. We believe that non-obtrusive observation afforded by social Q&A data is particularly valuable for this health topic, eating disorder, and the user group, teens and young adults.

Language Use Analysis

While there exists a wide array of linguistic analysis with varying levels of sophistication, some researchers in the field of social linguistics have been advocating for a simple approach based on counting and categorizing words. They argue that the words people use in their daily speaking or writing reveal a great deal of information about their situation and the state of mind, often in an unexpected way (Boals, 2005; Chung & Pennebaker, 2007). In a series of studies on natural language use, using various kinds of texts (speech transcripts, emotional narratives, journals, and so on), they have demonstrated that aggregate counts of word categories in a text, regardless of the context in which individual words occur, closely correlate with various psychological, cognitive, and even biological measures (Pennebaker et al., 2003). In their analysis, they distinguish two broad categories of words – content words and function words (Tausczik & Pennebaker, 2010).

Content words generally include nouns, regular verbs, adjectives and adverbs. Taken together, content words represent the theme or topic of the text. It is then natural that most content analysis approaches focused on this type of content-heavy words. Similarly, automatic indexing and information retrieval techniques have revolved around the idea that certain words in a text better represent its content or topic. In general, frequently appearing terms, excluding stopwords, are assumed to be more important.

While content words represent *what* people are talking or writing about (linguistic content), function words, also called style words, are related to *how* they are saying it (linguistic style). Function words include pronouns, prepositions, articles, conjunctions, and auxiliary verbs. Note that these words usually fall under the category of 'stopwords,' words that are considered as of little to no value in indexing or content analysis. However, a wide range of word use studies consistently found the important role of function words, not only in understanding the psychological state of the speakers (writers), but also uncovering social factors or situations affecting their thoughts and feelings. Among others, pronoun uses have been studied heavily, and linked to depression or mental distress (first person singular pronoun), isolation or group identity (first person singular vs. plural pronouns), social support, engagement, or awareness (second and third person pronouns) (Boals, 2005; Chung & Pennebaker, 2007; Pennebaker et al., 2003). Other function words are also found to be useful in detecting different characteristics in writing. For instance, the use of conjunctions (e.g. "and," "but," "also") can be an indication of coherence in narratives, as they are used to combine or juxtapose multiple thoughts (Graesser et al., 2004, as cited in Tausczik & Pennebaker, 2010). The use of prepositions, on the other hand, suggests that information given in the text is more complex and concrete, as they are often used to qualify or further specify what is being described (Tausczik & Pennebaker, 2010).

In summary, the studies of word use demonstrate that relatively simple text analysis methods can reveal much about the message without getting deeply into examining intricate semantic and/or syntactic structure of language. The analysis strategy that we adopt in this study was informed by the word use analysis described above.

Similar technical approaches were adopted in previous studies for different purposes. Harper et al. (2009), in their investigation of three social Q&A sites (*Yahoo! Answers*, *Answerbag*, and *Ask Metafilter*), identified two broad categories of questions – informational questions and conversational questions – and investigated the problem of detecting question type for a given question using machine learning algorithms. Some linguistic categories, such as interrogative words and personal pronouns, were included among a number of textual features examined for the classification purpose. They found that pronouns are highly predicative of different types of questions. More specifically, they reported that the personal pronoun "I" appeared highly in informational questions while "you" is used more often to address the readers. Liu et al. (2011) also used linguistic features as well as statistical features as the learning features of their machine learning algorithms. The task for their classifier is to distinguish questions asked by healthcare professionals from those asked by healthcare consumers. They identified four linguistic categories – interrogative words, personal pronouns, indefinite pronouns, and auxiliary verbs – as potentially useful features for the classification purpose. Classification performance was measured using different combinations of the linguistic features and statistical features (e.g. word length, sentence length).

Unlike Harper et al. (2009) and Liu et al. (2011), who analyzed the word use and some linguistic categories for the purpose of automatic classification of questions, we are interested in exploring whether and how the language in questions reflect different nature of needs behind the questions or different situations where the needs arose.

Sentiment Analysis

Emotion plays an important role in people's information behavior as a motivating factor (Nahl, 2007). In the context of health information seeking, where user behavior is often interpreted as emotion-focused coping strategies in face of a health-threatening situation (Lambert & Loiselle, 2007), the emotional dimension was deemed to be more important. In light of that, we are interested in examining the degree to which teens express emotions in the questions and how they differ by question categories. Sentiment analysis was chosen as a method for the inspection. It should be noted here that, with sentiment analysis, we focus only on the valence dimension of emotions, which varies from positive to negative (Russell, 1980), rather than discrete emotions.

Sentiment analysis, also known as opinion mining or subjectivity analysis, encompasses various techniques to detect sentiments expressed in natural language text, mostly focusing on polarity (positive or negative) of sentiments (Liu, 2010; Pang & Lee, 2008). Broadly speaking, there are two main approaches in sentiment analysis: lexicon-based approach and machine-learning approach (Taboada et al., 2011). The lexicon-based approach relies on complied lists of words, in which sentiment orientation of each word is defined. The sentiment of a text is basically determined based on the occurrences of these words. The machine-learning approach, on the other hand, typically involves a set of texts with known polarity and a learning algorithm to identify features associated with a particular sentiment in the training set. The identified features are then used to detect the sentiment in other texts.

Although sentiment analysis has been used mainly for commercial goals to collect consumer feedback on products or brands, there have been studies applying sentiment analysis to Social Q&A data. For instance, Li et al. (2008) proposed to automatically detect subjectivity orientation (subjective vs. objective) of a question to find answers that better match the intent of the question. Using a supervised learning method, they identified features (character, word, and POS n-grams) that are useful for detecting subjectivity orientation of questions and answers. Denecke and Nejdl (2009) applied sentiment analysis to social media contents in the medical domain. They noted a need for more sophisticated search mechanisms for user-generated medical contents that take into account the expertise of the author and the type of information given. With that goal in mind, they compared medical Q&A portals, medical weblogs, medical reviews, and Wikis. Sentiment analysis was performed on blog posts in order to consider informative posts apart from affective posts. These studies use sentiment analysis as part of algorithms to classify questions in social Q&A based on their expressed sentiment. Our approach differs from these in that our purpose in analyzing sentiments is to assess and compare the degree to which emotions are expressed in questions in different categories, and to explore the role of emotions in questions asked in social Q&A. With the availability of user generate contents on a variety of topics and in diverse contexts, Thelwall et al. (2010) noted the potential value of sentiment analysis for studying user behaviors, especially for understanding the affective dimension of information seeking. This study attempts to exploit the potential.

METHODS

Data set and question categories

In this study, we use a dataset constructed in our previous study of eating disorder questions on Yahoo! Answers (Bowler et al., 2012), in which we collected a total of 2,230 questions on the topic of eating disorders from *Yahoo! Answers*, and subsequently built two datasets based on different criteria. A classification scheme was derived inductively through content analysis of these two datasets, with five overarching themes - *Seeking Information, Seeking Emotional Support, Seeking Communication, Seeking Self-Expression,* and *Seeking Help to Complete a*

Task (For a detailed description of the datasets and a discussion of question categories and subcategories in our scheme, see Bowler et al., 2012).

In this study, we used the dataset containing the 180 longest questions from the initial set of 2,230 questions (the number of words per question in this subset ranges from 439 to 1,466) and used the coding results from the previous study to examine characteristics of questions in different categories. Among the 180 questions, we removed those questions where the topic of eating disorders was not a main interest or a central theme but rather mentioned in a cursory manner. The final dataset for this study includes 72 questions that two coders agreed were closely tied to the topic of eating disorders.

Due to the small size of the dataset, we divided the questions into two broad categories – informational and socio-emotional – by aggregating the original coding. Socio-emotional category includes subcategories of *Seeking Emotional Support*, *Seeking Communication*, and *Seeking Self-Expression*; informational category includes *Seeking Information* and *Seeking Help to Complete a Task.* Out of the total of 72 questions (ones that are agreed as relevant by two coders), 40 questions were categorized as *socio-emotional* questions and 16 were categorized as *informational* questions. The remaining 16 were categorized as *mixed*, meaning that coding results were split between socio-emotional and informational. This high proportion of socio-emotional questions may be an artifact of choosing the longest questions in the initial dataset.

Language analysis

To explore differences in word use between questions in different categories, we examined three different aspects of word use: linguistic contents (content words), linguistic style (function words), and subjectivity/sentiment expression. As for the methods, term frequency analysis, Part-of-Speech (POS) analysis, and sentiment analysis were used to capture those aspects. All the analysis was done in *R*, a statistical computing environment (http://www.R-Project.org).

In order to examine the words that represent central themes in question sets, we conducted an analysis of term frequency focusing on content words. Using the *tm* package in *R*, we created the question corpus with relevant metadata including the assigned category for each question. After a series of preprocessing, including removal of stopwords, we built a document-term matrix (DTM) holding the term frequency of each term *t* for each document *d*. In order to reduce the effect of document length, each term frequency measure in a document was normalized using the log average of all the term frequencies in the document.

In order to examine the use of function words in questions, it is necessary to first identify the linguistic category of each and every word in the question sentences. For that purpose, we conducted a Part-of-Speech (POS) analysis using the *KoRpus* package in *R*, which is a *R* wrapper for the *TreeTagger* program (Schmid, 1994). We then compared the distribution of POS tags across question types, and examined the differential use of the words in certain linguistic categories.

The emotional aspects expressed in questions were explored with sentiment analysis. We used a lexicon-based approach, which is basically matching terms in the corpus of interest against a list of subjective terms with pre-coded polarity. The lexicon-based sentiment analysis has been used widely in previous studies. We identified three commonly used lexicons in sentiment analysis and merged them for our analysis. The three lexicons we used were Harvard General Inquirer (Stone et al., 1966), Opinion Lexicon (Hu & Liu, 2004), Subjectivity Lexicon (Wilson et al., 2005). The resulting list includes 3,172 positive terms and 5,281 negative terms in total. Using the list of positive terms and negative terms and a simple algorithm, we assign subjectivity scores and sentiment scores to questions, in order to compare the extent to which emotional or subjective expressions appear in different types of questions.

RESULTS

Basic statistics per question set

Before going into the main analysis, we first inspected some descriptive statistics about the questions in each set, shown in Table 1. Socio-emotional questions tend to be longer and contain more proportion of stopwords (most of which can be regarded as function words as discussed in the Literature Review section). In Table 1, a clear trend of increased number and length of sentences can be seen when comparing informational questions to socio-emotional questions with the mixed questions in between. In contrast, average length of words is longest in informational questions and shortest in socio-emotional questions. It is interesting to see that the mixed questions (questions in which category decisions of the coders were split) stand in the middle on every measure, suggesting that these questions take characteristics of both questions, not only semantically but also in linguistic features. In order to make a clearer comparison between the two main categories, we excluded the mixed questions from the subsequent analyses.

	Infor-mational	Mixed	Socio-emotional
Avg. no. of words per question	613.50	650.69	760.10
Avg. % of stopwords per question	59.4%	64.4%	69.7%
Avg. word length	4.46	4.19	3.95
Avg. no. of sentences per question	44.50	41.25	46.68
Avg. sentence length	14.90	17.32	21.93

Table 1. Descriptive statistics

Content words – term frequency analysis

The analysis of term frequency allows us to see what people are talking about with related to their questions on eating disorder. In order to focus on content words, stopwords were removed in this part of analysis. By comparing the frequent terms across question categories, we can find out the words that typify a certain type of question.

Figure 1 shows two Wordclouds created from the result of term frequency analysis, visually presenting terms based on their prominence in the respective set. A larger font size indicates a higher frequency of the term. Note that frequency analysis was done after stemming, but for a better reading of the results, stems were replaced by complete words after computing frequencies. Among the possible extensions of the stem word, a stem word was mapped to either the shortest or the most prevalent form found in the corpus.

While there is some obvious and expected overlap such as 'eat', 'disorder', or 'teen', the two wordclouds in Figure 1 show quite different patterns of word usage in the two question sets. For example, words referring to family members stand out in the socio-emotional set, whereas more general reference terms, such as 'girl' or 'model', appear more prominent in the informational set.

In order to further examine the differential word use by question category, the chi-square analysis was performed to find out those terms that are significantly more frequent in one set than in the other. More specifically, we calculated chi-square value for each term in our dataset using the chi-square function in the R *corpora* package, which is based on Stefan Evert's algorithm for comparing word frequency in text corpora (Evert, 2005). For this test, the raw frequency counts were used instead of log-normalized measures because the test itself takes the discrepancies in document sizes into account and compares the frequencies proportionally.

The result of chi-square analysis shows that there are 44 words that are significantly more frequent in socio-emotional questions, and 108 words appearing more frequently in informational questions (at $p < .05$). Table 2 presents the top 40 words that are significantly more frequent in socio-emotional questions based on the chi-square value, and Table 3 shows the same for informational questions.

1-10	11-20	21-30	31-40
time	live	smoke	ate
friend	gluten	random	school
try	hate	argue	deal
lot	mom	doctor	refuse
sister	bit	yell	sexual
day	sorry	scare	site
dad	mean	facebook	home
life	father	kcal	stupid
mother	move	actual	start
feel	sometime	pretty	boyfriend

Table 2. Top 40 words in socio-emotional questions

1-10	11-20	21-30	31-40
model	girl	result	paper
beauty	anorexia	skinny	feet
retouch	muscle	specify	fashion
image	jasmine	characterized	osteoporosis
magazine	loss	stamp	pressure
disorder	children	contest	diet
nervosa	women	mary	behavior
body	natural	pageant	woman
media	build	bulimia	caus
thin	celebrity	society	injury

Table 3. Top 40 words in informational questions

The results confirmed the initial observation made in Figure 1 Wordclouds. Words denoting social relations show a high

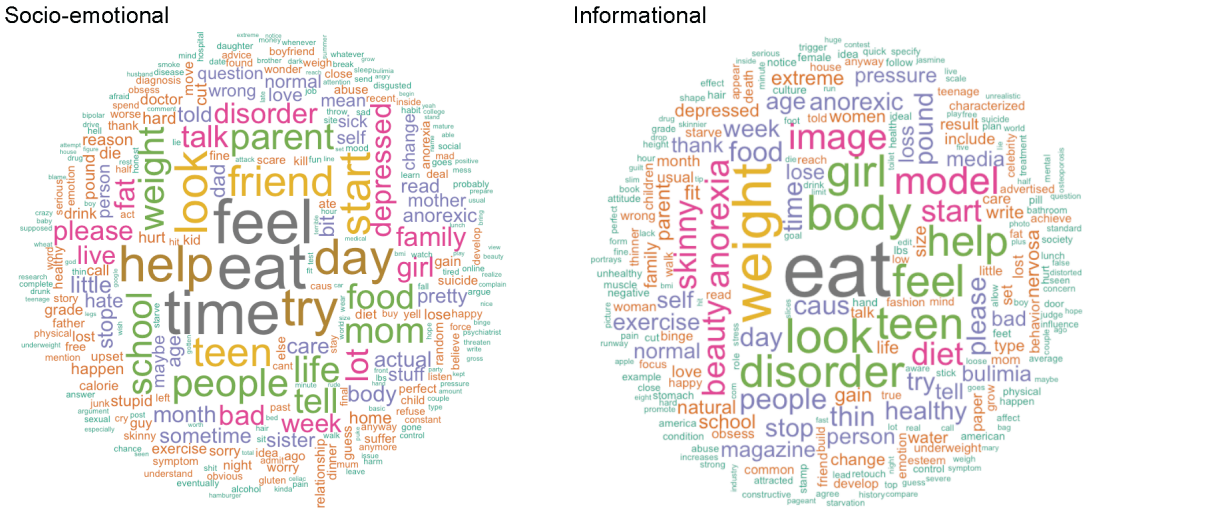

Figure 1. Wordclouds of two question sets

frequency in the socio-emotional question set. Words for family relations – sister, dad, mother, mom, father, and parent – all have high chi-square value (all significant at $p < .05$). The word 'friend' ranked second in the order of chi-square value, and 'boyfriend' was also significantly more frequent in this set. Words referring to social settings such as home or school were also high. Another group of words that appear more frequently in socio-emotional questions includes those words related to feeling or subjective evaluation (e.g. hate, sorry, scare, stupid, upset).

On the other hand, in informational questions, among the words with highest chi-square value are terms related to media or information bearing objects such as magazine, paper, article. Words referring to specific medical conditions (e.g. osteoporosis) or describing human body (muscle, thin, skinny) were also relatively more frequent in informational questions. Another interesting observation is that terms for various eating disorder conditions (e.g. anorexia, bulimia, binge) were frequently found in informational questions, but not in socio-emotional questions, even though those terms were included in our search terms.

Function words – POS analysis

In this part, we shift our attention to linguistic markers that represent different styles of communication. As discussed in the Literature Review above, function words include words in certain linguistic categories such as preposition, which can be identified by Part-of-Speech (POS) tagging. .

Tables 4 and 5 present Top 10 most frequent POS tags and their proportions in the socio-emotional question set and the informational question set, respectively. Note that the POS tags shown in Tables 4 and 5 are defined in the Penn Treebank tag set (Marcus et al.,1993). Comparing Table 5 and 6, we can see that the POS tags are not distributed consistently in the two sets.

Tag	Description	%
NN	Noun, singular or mass	11.83
PP	Personal pronoun	9.61
IN	Preposition or subordinating conjunction	7.82
RB	Adverb	6.68
DT	Determiner	5.88
JJ	Adjective	5.56
VBD	Verb, past tense	5.22
CC	Coordinating conjunction	4.69
VB	Verb, base form	4.35
NNS	Noun, plural	4.22

Table 4 Top 10 POS tags and their proportions in socio-emotional questions

Tag	Description	%
NN	Noun, singular or mass	13.62
IN	Preposition or subordinating conjunction	8.24
DT	Determiner	6.79
JJ	Adjective	6.77
NNS	Noun, plural	6.24
PP	Personal pronoun	5.50
RB	Adverb	5.05
CC	Coordinating conjunction	4.24
VB	Verb, base form	3.94
VBP	Verb, present tense, not 3rd person singular	3.91

Table 5 Top 10 POS tags and their proportions in informational questions

In order to further examine the differences in tag distributions between the two sets of questions, we did a pairwise comparison by calculating the following value for each tag:

$$S_{s,i}^{T} = \frac{P_s^T - P_i^T}{P_s^T + P_i^T}$$

where P_s^T is the proportion of the given tag T in the socio-emotional question set (i.e., the number of occurrences of the tag divided by the total counts of all the tags in the set), and P_i^T is the proportion of the tag in the informational question set. As the result, the tags that appear more often (in a higher proportion) in socio-emotional questions have positive values, while negative values are given to the tags appear more in informational questions. Figure 2 shows the value of S for different POS tags. POS tags that appear less than 0.2% in both sets were removed from the figure.

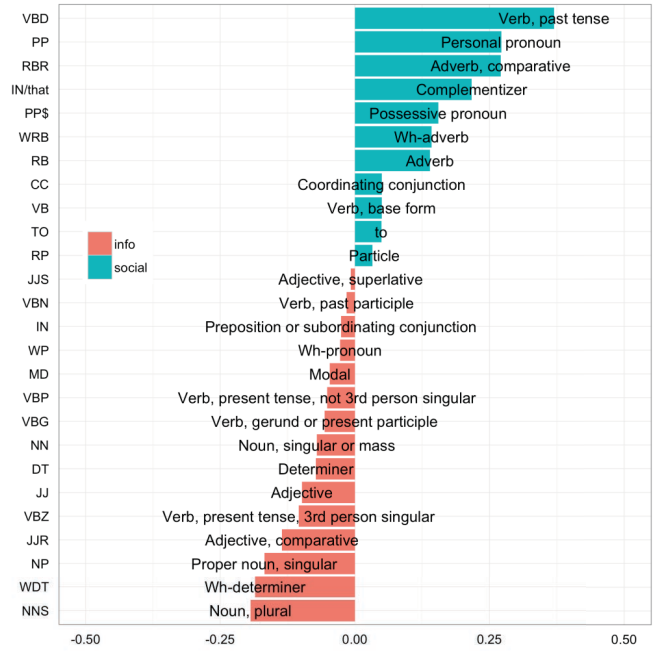

Figure 2. POS tag distribution

Several observations can be made in Figure 2:

Words in the noun class: Socio-emotional questions contain more pronouns. Both personal pronoun and possessive pronoun show a much higher proportion in socio-emotional questions. Proper nouns (both plural and singular) and common nouns, on the other hand, tend to be more frequent in informational questions.

Verbs and verb tenses: Past tense verbs are distinctively more frequent in socio-emotional questions. The high relative frequency of past tense in socio-emotional question suggests that the askers' attention was focused on past events in describing their situation. In informational questions, present tense, gerund and verbs combined with modal auxiliary (e.g. can, could, may, should, will) appeared more often. Comparing the relative frequencies of VBZ (verb, present tense, third person singular) and VBP (verb, present tense, not third person singular), we can see informational questions tend to use more third person verbs, which may indicate objective description rather than subjective narrative. Conversely, when the present tense is used in socio-emotional questions, it is written from the viewpoint of the questioners themselves (first person) or of the audience (second person).

Interrogative words: There is also a difference in the use of interrogative words. While Socio-emotional questions have more Wh-adverb (e.g. how, where, why), informational questions have more Wh-pronoun (e.g. what, who, whom). Wh-determiner (e.g. which, whichever, what, whatever) is also high in informational questions.

Prepositions and conjunctions: It is not surprising that socio-emotional questions contain more connective words, considering that they tend to be longer than informational questions in our dataset. However, a slight, but notable difference was found in the use of prepositions and conjunctions in the two question sets. While coordinating conjunctions appear more in socio-emotional questions, prepositions or subordinating conjunctions are used more often in informational questions. This finding is important because the use of preposition and conjunction words is connected to some important properties of text such as narrative structure, cognitive complexity, or clarity [more in Discussion].

Sentiment analysis

This part of the analysis concerns itself with the extent to which subjective views and emotions are expressed in questions. The level of subjectivity and the overall sentiment in a question is measured with a pre-complied list of subjective terms with the polarity sign (positive/negative).

At first, for each question, the number of occurrences of positive and negative terms from the list was counted. Considering the variations in question lengths, the simple counts were then divided by the length of the question (the total number of words after removing stopwords), showing the fraction of negative/positive terms within the given question. These were called negative/positive scores. Using the positive and negative scores, two compound measures were defined. The subjectivity score, which is calculated simply by summing the positive score and the negative score for a given question, measures the extent to which the question contains subjective or emotional expressions. Sentiment score, on the other hand, represents the sentiment orientation of a question, i.e. whether the question overall is positive or negative. Sentiment score is calculated by subtracting the negative score from the positive score. A negative value of sentiment score indicates that there are more negative terms than positive terms in a given question. The larger the sentiment score value, the more positively oriented the question is.

Table 6 shows the average scores for all four measures – negative, positive, sentiment, subjectivity – for each question set. To compare the means between the two sets, t-test was done with the Welch correction of nonhomogeneity of variance.

	Socio-emotional	Informational	*t (df)*
Negative score	0.184	0.137	3.02 (21.50) **
Positive score	0.113	0.118	-0.37 (22.34)
Subjectivity	0.298	0.255	2.24 (23.21)*
Sentiment	-0.070	-0.019	-2.55 (20.63)*

* $p < .05$ **$p < .01$.

Table 6. Score comparison between the two sets

Overall, regardless of question categories, the fraction of negative words in a question turned out to be higher than that of positive terms (negative score of 0.184 vs. positive score of 0.113 in socio-emotional questions; negative score of 0.137 vs. positive score of 0.118 in informational questions). This might be simply a reflection of the fact that the lexicon itself includes a longer list of negative terms to be matched. Comparing the two question types, it becomes apparent that socio-emotional questions contain much more negative terms and less positive terms. The difference in the use of positive terms is significant at $p < .01$. This tendency is confirmed again in the compound measure of sentiment score. Although both socio-emotional questions and informational questions have negative values (meaning that the fraction of negative words is larger than the fraction of positive words in the questions), the magnitude is more than 3 times larger in socio-emotional questions. Therefore, we can conclude that, on average, informational questions are more positively oriented compared to socio-emotional questions.

Another striking observation here is the high volume of emotion words, indicated by a high subjectivity score, in both question sets. The subjectivity score of a question, as defined above, shows the fraction of both positive and negative words out of the total count of words (excluding stopwords) in the question. On average, about 30% of terms in socio-emotional questions and 25.5% of terms in

informational questions are subjective (emotional) terms. Pennebaker et al. (2003) provides a point of reference for a comparison with other kinds of texts with regard to the degree of emotional expressions. According to them, "in daily speech, emotional writing, and even affect-laden poetry, less than 5% of the words people use can be classified as emotional" (p. 571). This number, 5%, however, cannot be directly compared to our subjectivity score because it is most likely that they meant the percentage of emotion words over the entire text, without removing stopwords. To make the numbers comparable, we revisit the original length of each question before removing stopwords, and recalculate the fraction of emotion (sentiment) words (both positive and negative) with respect to the original length. The result shows that emotional words account for about 8.9% of words in socio-emotional questions and about 10% of words in informational questions on average (note that the percentage of emotional words is now higher in informational questions. It is due to the fact that, as shown in Table 1, socio-emotional questions contain a lot more stopwords). As can be seen, the new percentages in our question sets are still much higher that the number (5%) mentioned in Pennebaker et al (2003). While some of this difference may be attributed to the difference in the lexicon of emotional/sentiment words used in the analysis, it is still reasonable to conclude that emotional expressions are abundant in our dataset exceeding the level usually observed in other texts.

DISCUSSION

Using simple text analysis methods, we examined linguistic content, linguistic style, and emotional expressions in two different categories of eating disorder questions from *Yahoo! Answers*. In the following we will summarize some of the main findings and their implications.

The most frequent terms in our dataset correspond to the contributing factors of eating disorder surveyed in medical literature. Studies of eating disorders commonly report family dynamics (e.g. dysfunctional family, controlling or neglecting parents, communication issues) and peer influence as one of the most important factors leading to the development and persistence of eating disorders (Leon et al., 1994; Polivy & Herman, 2002). In the socio-emotional question set, terms denoting family relations as well as peers appear highly frequently. It is interesting to note that informational questions contain more terms that are related to some widely known socio-cultural factors of eating disorders (e.g. media influence, idealization of thinness, promotion of certain body image through celebrities or fashion models). It appears that while socio-emotional questions address more personal aspects, informational questions concerns more objective and depersonalized view of the issues.

The result of the POS analysis also leads to a similar inference. First, the use of words in the noun class shows a sizable difference between two question sets, with a much higher proportion of personal pronouns in socio-emotional questions, and a clear profusion of general nouns in informational questions. Second, socio-emotional questions are written predominantly in the past tense, and when written in the present tense, it is more often to talk about the questioners themselves (first person) or to address the audience (second person). In contrast, informational questions tend to use more third person verbs, suggesting relatively neutral or disengaged attitudes. In addition, while social questions tend to be longer and more verbose, connecting multiple ideas using coordinating conjunction words (e.g. and, but, so, yet), informational questions tend to include more specifications or detailed explanation about the information being sought, as evidenced by the abundant use of prepositions and subordinating conjunctions. These results, taken together, suggest that socio-emotional questions show similar characteristics to that of personal narratives and informational questions have traits of objective and precise writing. This observation is also corroborated by the results of sentiment analysis, in which socio-emotional questions turned out to have a significantly higher subjectivity score.

We found that emotions appear abundant in eating disorder questions in general, regardless of the question type, but negative emotion words are clearly more dominant in socio-emotional questions. Taken with the observation of the heavy use of past-tense verbs in socio-emotional questions, this finding indicates that questioners often share the past emotional experiences that are related, in their mind, to their current health issues. Rime et al. (2002), in their study of social sharing of emotions, explained why people share emotional experiences with others. Drawing on the cognitive dissonance theory (Festinger, 1960), they claimed that sharing emotions gives people an opportunity to work through their experiences and to gain a new cognitive view of their problem situation. They also noted social functions of sharing emotions, including an increased sense of bonding and social support. Our findings of the language use in socio-emotional questions suggest that in many cases these questions are not about seeking information or helps related to the given health issue per se, but rather about expressing their own thoughts and emotions in search for an explanation for their health problems.

In our previous study, we observed that there exists a wide array of social, informational, and affective needs manifested in the eating disorder questions posted to Social Q&A. We argued that social Q&A may serve as a ground for teens and young adults to work out their problems in a variety of ways including venting emotions, reflecting on their experiences, and looking for solutions (Bowler et al., 2012). The results of the current study further demonstrate distinctive characteristics of questions in different categories, which we believe reflect the nature of underlying needs, and support the previous finding.

The different characteristics of questions found in this study highlight potential values of this data in various areas and,

at the same time, point to a need for looking at the questions from different angles. As discussed above, our analysis uncovered that the common linguistic features shown in personal narratives appear in many questions, especially in the socio-emotional category. In those questions, users not only express their information needs (in a broad sense), but often put them in context by explaining their personal situations. This kind of question can be a valuable source for gaining insights on personal and contextual factors affecting health information seeking behavior. On the other hand, the long list of terms that are significantly more frequent in informational questions indicates that informational questions vary more in their contents. This suggests that informational questions may be more useful in probing the kinds of information being sought in relation to eating disorders or the range of issues associated with eating disorders in teens' mind than socio-emotional questions.

Although interesting findings were made, this study has its limitations. First and foremost, the dataset analyzed in this study is small in size and unbalanced in the composition of question categories. As this study set out to be an initial exploration, we chose the small set of the longest text-rich questions. These questions may well not be representative of all eating disorder questions from *Yahoo! Answers*, let alone other sites. It remains to be studied whether the distinctive characteristics we found in this particular set would generally present in a larger set that also includes much shorter questions. Second, our language analysis was admittedly at a crude level and left some aspects of word use unexplored. For instance, in previous linguistic studies, meaningful differences were shown in the use of different kinds of personal pronouns (e.g. first person singular, first person plural, second person, third person), yet our analysis did not differentiate them. We found that the overall use of personal pronouns appeared to be higher for socio-emotional questions, but further examination may uncover differential use of these words by their kind.

CONCLUSION

Social Q&A provides the possibility of looking into how people verbally express their information needs in natural language. In this study, we analyzed the linguistic properties of different types of questions on the topic of eating disorders.

One of the main purposes of this study was to assess the applicability and effectiveness of the linguistic approach to our study of health information seeking in social Q&A. The results of this study demonstrate that the combination of simple text analytic techniques reveals much about the linguistic characteristics associated with different kinds of questions, and thereby sheds lights on the nature of the needs underneath the questions.

Given the promising results and with the ideas on further sophistication of our approach obtained from this study, we have started to expand this study with a larger dataset, addressing the limitations of this study. There are a number of possible directions for further exploration. Due to the small size of the dataset in this study, we imposed a binary distinction of questions over our detailed taxonomy of eating disorder questions. A larger dataset would allow us to examine the thematic and linguistic features at the sub-category level. We also intend to analyze the language use in answers along with the questions, and see what patterns may emerge, whether and how they differ across question types, and what they tell us about answers and answerers.

REFERENCES

Boals, A. (2005). Word use in emotional narratives about failed romantic relationships and subsequent mental health. *Journal of Language and Social Psychology*, *24*(3), 252-268.

Bowler, L., Oh, J. S., He, D., Mattern, E., & Jeng, W. (2012). Eating disorder questions in yahoo! Answers: Information, conversation, or reflection? In *Proceedings of the Annual Meeting of the American Society for Information Science and Technology*.

Chou, W. -Y. S., Hunt, Y. M., Beckjord, E. B., Moser, R. P., & Hesse, B. W. (2009). Social media use in the united states: Implications for health communication. *Journal of Medical Internet Research*, *11*(4).

Chou, W. -Y. S., Hunt, Y., Folkers, A., & Augustson, E. (2011). Cancer survivorship in the age of youtube and social media: A narrative analysis. *Journal of Medical Internet Research*, *13*(1).

Chung, C. K., & Pennebaker, J. W. (2007). The psychological function of function words. *Social Communication: Frontiers of Social Psychology*, 343-359. .

Denecke, K., & Nejdl, W. (2009). How valuable is medical social media data? Content analysis of the medical web. *Information Sciences*, *179*(12), 1870-1880.

Evert, S. (2005). *The Statistics of Word Cooccurrences: Word Pairs and Collocations.* Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Gazan, R. (2011). Social Q&A. *Journal of the American Society for Information Science and Technology*, *62*(12), 2301-2312.

Harper, F. M., Moy, D., & Konstan, J. A. (2009). Facts or friends?: Distinguishing informational and conversational questions in social Q&A sites. In *Proceedings of the 27th international conference on human factors in computing systems* (pp. 759-768).

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 168-177).

Jaloba, A. (2009). The club no one wants to join: Online behaviour on a breast cancer discussion forum. *First Monday*, *14*(7-6).

Katzman, D.K., Kanbur, N.O., and Steinegger, C.M. (2010). "Medical screening and management of eating disorders in adolescents." In W.S. Agras (ed.), *The Oxford handbook of eating disorders*, (267-291). New York: Oxford University Press.

Keel, P.K. (2010). "Epidemiology and course of eating disorders." In W.S. Agras (ed.), *The Oxford handbook of eating disorders* (25-32). New York: Oxford University Press.

Lambert, S. D., & Loiselle, C. G. (2007). Health information seeking behavior. *Qualitative Health Research*, *17*(8), 1006-19.

Leon, G. R., Fulkerson, J. A., Perry, C. L., & Dube, A. (1994). Family influences, school behaviors, and risk for the later development of an eating disorder. *Journal of Youth and Adolescence*, *23*(5), 499-515.

Li, B., Liu, Y., & Agichtein, E. (2008). CoCQA: Co-training over questions and answers with an application to predicting question subjectivity orientation. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 937-946).

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing,* 627-666.

Liu, F., Antieau, L. D., & Yu, H. (2011). Toward automated consumer question answering: Automatically separating consumer questions from professional questions in the healthcare domain. *Journal of Biomedical Informatics*, *44*(6), 1032-8.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, *19*(2), 313-330.

Nahl, D. (2007). The centrality of the affective in information behavior. In D. Nahl & D. Bilal (Eds.), *Information and emotion: The emergent affective paradigm in information behavior research and theory* (pp. 3-37). Medford, NJ: Information Today.

Neumark-Sztainer, D. and Hannan, P.J. (2000). "Weight-related behaviors among adolescent girls and boys: Results from a national survey." *Archives of Pediatrics & Adolescent Medicine* 154(6): 569-577.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1-135. .

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*(1), 547-577.

Polivy, J., & Herman, C. P. (2002). Causes of eating disorders. *Annu. Rev. Psychol*, *53*, 187-213.

Ressler, P. K., Bradshaw, Y. S., Gualtieri, L., & Chui, K. K. H. (2012). Communicating the experience of chronic pain and illness through blogging. *Journal of Medical Internet Research*, *14*(5).

Rideout, V. (2001). *Generation RX.com: How young people use the internet for health information.* Henry Kaiser Family Foundation: Menlo Park, CA. Available at http://www.kff.org/entmedia/20011211a-index.cfm (retrieved on September 4, 2011).

Rimé, B., Corsini, S., & Herbette, G. (2002). Emotion, verbal expression, and the social sharing of emotion. In S. R. Fussell (Ed.), *The verbal communication of emotions: Interdisciplinary perspectives* (pp. 185-208). Lawrence Erlbaum Associates Publishers.

Robillard, J. M., Whiteley, L., Johnson, T. W., Lim, J., Wasserman, W. W., & Illes, J. (2013). Utilizing social media to study information-seeking and ethical issues in gene therapy. *Journal of Medical Internet Research*, *15*(3), e44.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161-1178.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing* (Vol. 12, pp. 44-49).

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis. In Cambridge, MA: MIT press.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267-307.

Tausczik, . R., & Pennebaker, . W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24-54.

Thelwall, M., Wilkinson, D., & Uppal, S. (2010). Data mining emotion in social network communication: Gender differences in myspace. *Journal of the American Society for Information Science and Technology*, *61*(1), 190-199.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 347-354).

Zickuhr, K. (2010). *Generations 2010*. Washington D.C.: Pew Internet & American Life Project. Available at http://pewinternet.org/Reports/2010/Generations-2010.aspx (retrieved on March 4, 2013).