

Investigating the Potential of Extracting Answers to Health-Related Quality of Life Questionnaires From Patient Community Data

David M. Schmidt, MSc¹, Brian Po-Han Chen, MSc², Deborah Kuk, MSc², Yogesh Vohra, MS, PharmD, PhD², Valmeek Kudesia, MD², Philipp Cimiano, PhD¹
¹Center for Cognitive Interaction Technology, Bielefeld University, Bielefeld, Germany, ²Inspire, Arlington, VA, USA; correspondence: rwe@inspire.com



Introduction

- Health-related quality of life (HRQoL) questionnaires capture the impact of disease and treatment on the well-being and functioning of a patient
- Collection of data imposes significant burden on patients and health care personnel
- Many patients share detailed information about their condition and treatment experiences on online patient fora
- We investigate to what extent HRQoL information can be gained from the analysis of posts by patients in online healthcare communities

Methods

- Participants from the Inspire.com patient online forum were recruited using the following eligibility criteria:
 - Has a female breast cancer diagnosis
 - Age 18 years or older
 - Residency in the United States
 - Has at least one post or comment in the “Breast Cancer” and “Advanced Breast Cancer” Inspire.com communities
- Participants completed the EORTC QLQ-C30 and QLQ-BR23 questionnaires [1], which consists of 30 general and 23 additional breast cancer-specific questions
- Participants’ online posts in the last 24 months were coded by human coders to identify parts of the text providing answers to one of the 53 questions
- “Did you feel ill or unwell?” used as a fallback for HRQoL-relevant content not fitting in any other category
- Applied classification metrics from machine learning research were used to assess alignment between posts and survey responses for data close to the survey conduction time (6 months) and all data (24 months)
- Data was reduced to binary answers: “Yes” or “No”
- True Positives are defined as instances where both the online post and the corresponding questionnaire response indicate a “Positive” answer (e.g. “I had so much pain last week!” and “Yes” for “Have you had pain?”)
- False Negatives, False Positives and True Negatives are defined analogously
- Precision, Recall and F_1 scores were calculated for each question to determine how well the coded posts predict the questionnaire responses

Results

- 134 female breast cancer patients were recruited
- Of 2,669 annotated posts and comments, 613 (23.0%) contained information for at least one QoL question
- The data annotation yielded a substantial agreement (mean Fleiss’ Kappa of 0.5, with all values being non-negative).
- Answers were found in the coded data for 50 out of 53 QoL questions
- The information coded in the posts reliably predicted the answers given in the questionnaires (mean $F_1 = 0.7$)
- The five most frequently answered questions:
 - “Did you feel ill or unwell?” (304 of 2683)
 - “Did you worry?” (105 of 2683)
 - “Have you had pain?” (104 of 2683)
 - “Did you feel tense?” (85 of 2683)
 - “Were you limited in doing either your work or other daily activities?” (77 of 2683)

Results

Figure 1. Posts per user, with 123 users on the x-axis and (log-scaled) number of posts on the y-axis.

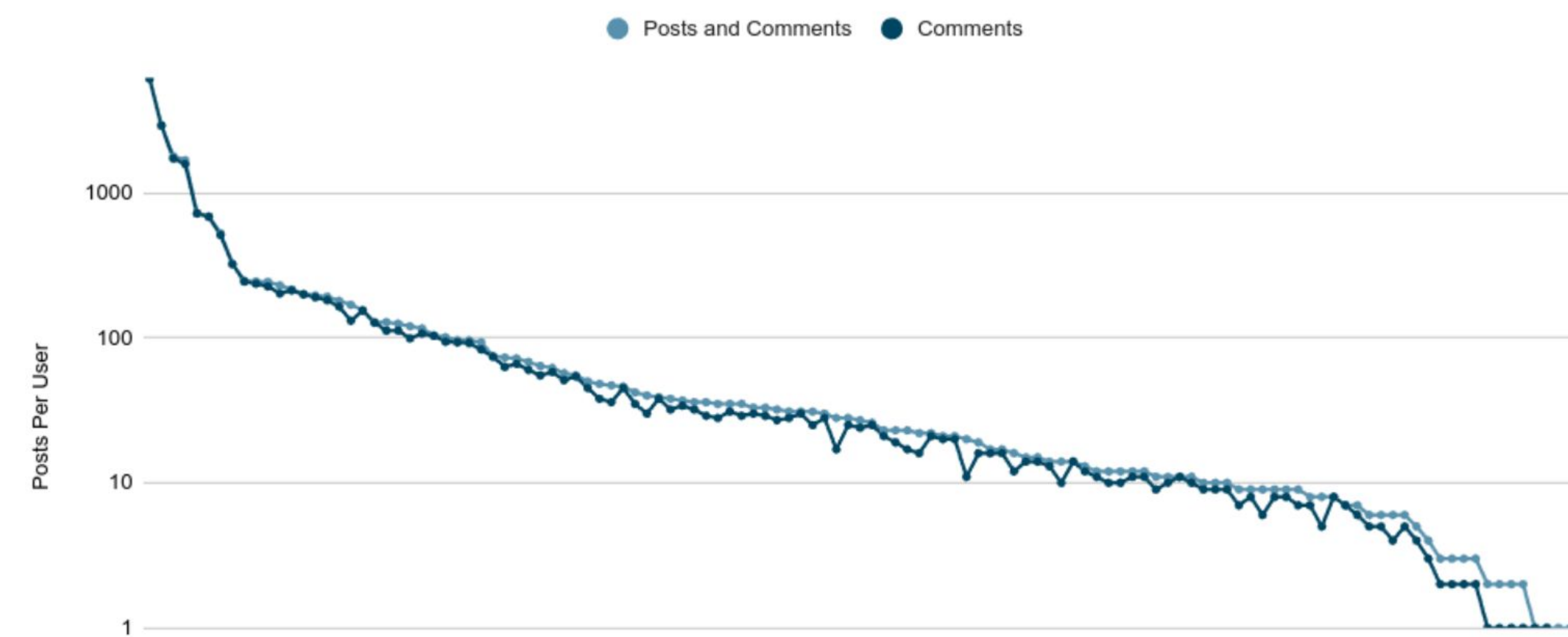


Figure 2. F_1 scores for the 10 most frequently coded questions, considering the last 6 months of data.

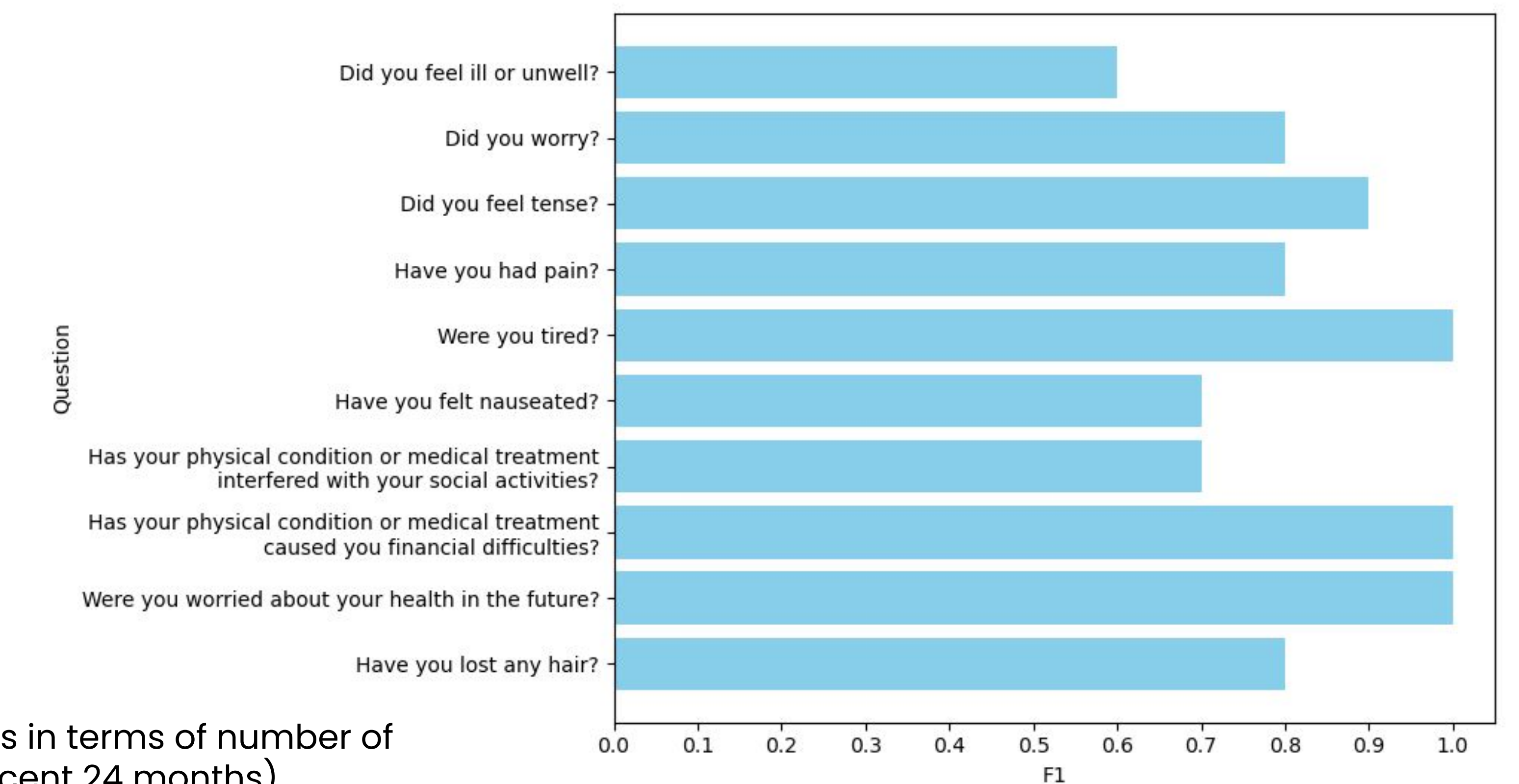


Figure 3. Portion of users answering the top 10 most frequently coded labels, considering the most recent 24 months of data.

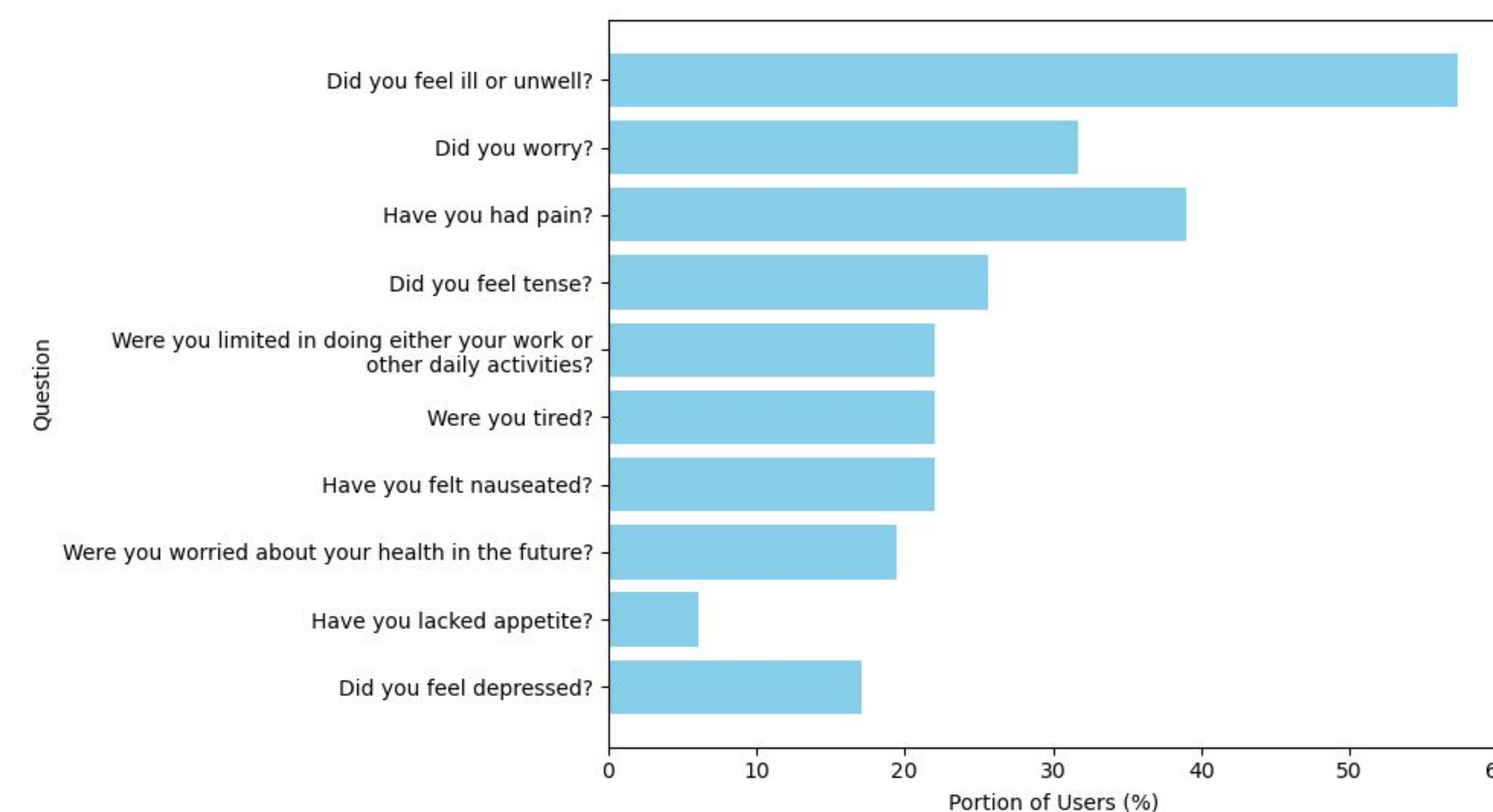


Table 1. Top 10 most frequent labels in terms of number of posts that are annotated (most recent 24 months).

Question	Number of Posts	Portion of Posts	Portion of Users
Did you feel ill or unwell?	301	11.0%	57.3%
Did you worry?	105	3.9%	31.7%
Have you had pain?	102	3.7%	39.0%
Did you feel tense?	84	3.1%	25.6%
Were you limited in doing either your work or other daily activities?	77	2.8%	22.0%
Were you tired?	55	2.0%	22.0%
Have you felt nauseated?	51	1.9%	22.0%
Were you worried about your health in the future?	41	1.5%	19.5%
Have you lacked appetite?	32	1.2%	6.1%
Did you feel depressed?	32	1.2%	17.1%

- Some questions (e.g., 39% for “Have you had pain?”) are answered by large parts of the user base, while others are rarely discussed
- Post counts follow a long-tail distribution
- Nevertheless, F_1 scores (harmonic mean of precision and sensitivity) of over 0.5 for all top 10 questions indicate high predictive power of posts against actual HRQoL survey answers

Conclusion

- 23% of patients’ posts contain information that can be mapped to at least one HRQoL question
- HRQoL information in posts is reliable, corresponding to the questionnaire answers to a large extent
- Analysis of online posts is not meant to replace the collection of questionnaire data but represents a low-burden alternative to gain HRQoL insights
- Future work should investigate how HRQoL information from online posts can complement existing HRQoL assessment methodologies and whether the information can be extracted automatically

Limitations

- Limited population from single online forum, might be affected by selection bias
- Intimate topics rarely discussed due to public nature of online fora
- Unequal distribution of posts among forum users, thus degree of extracted information very much depends on specific users
- Found signs of recall bias, e.g., mentioning worries but answering the opposite in the survey

Download this poster!



Table 2. Examples for confusion matrix values of “Have you had pain?”

Have you had pain?	Post	Survey Answer
True Positive	I received my third round of treatment a few days ago. The effects of the first round are still lingering, and it's becoming increasingly difficult to deal with both the physical and emotional pain. The fourth round will be my final one, but I haven't yet decided what path I will choose after that.	Yes
False Positive	I received my third round of treatment a few days ago. The effects of the first round are still lingering, and it's becoming increasingly difficult to deal with both the physical and emotional pain. The fourth round will be my final one, but I haven't yet decided what path I will choose after that.	No
False Negative	I received my third round of treatment a few days ago. I have no pain at all, it is unbelievable! The fourth round will be my final one, but I haven't yet decided what path I will choose after that.	Yes
True Negative	I received my third round of treatment a few days ago. I have no pain at all, it is unbelievable! The fourth round will be my final one, but I haven't yet decided what path I will choose after that.	No

[1] Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, Filiberti A, Flechtner H, Fleishman SB, de Haes JC. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993 Mar 3;85(5):365–376. PMID:8433390