Condition Unknown: Predicting Patients' Health Conditions in an Online Health Community

Changye Li
University of Minnesota
200 Union St SE, Minneapolis,
MN 55455
lixx3013@umn.edu

Haiwei Ma
University of Minnesota
200 Union St SE, Minneapolis,
MN 55455
maxxx979@umn.edu

Zachary Levonian
University of Minnesota
200 Union St SE, Minneapolis,
MN 55455
levon003@umn.edu

Svetlana Yarosh
University of Minnesota
200 Union St SE, Minneapolis,
MN 55455
lana@umn.edu

CSCW'18, November 3–7, 2018, Jersey City, NJ, USA.
© 2018 Copyright is held by the owner/author(s).
ACM ISBN 978-1-4503-6018-0/18/11.
https://doi.org/10.1145/3272973.3274075

Abstract

Online health communities rely on information about their users to provide services to members. We partner with the online health community CaringBridge.org to infer the health condition that users are discussing from their early writing on the site. We utilize the self-reported health condition data that is provided by users to train machine learning classifiers to predict the health condition of non-reporting users. An analysis of the classifier's errors reveals that users frequently discuss multiple health conditions. We present models with explainable features, enabling us to extract words for the enrichment of consumer health vocabularies and to support future designs connecting patients.

Author Keywords

Online Health Communities; Peer Health Support

ACM Classification Keywords

Human-centered computing [Collaborative and social computing]: Empirical studies in collaborative and social computing.

Introduction and Background

Online health communities (OHCs) have become a significant source for patients looking for support [7]. Previous studies have used the classification of

Site	*Total*	*Filtered*
Reported	230,568	129,542
Non-reported	357,642	88,148

Table 1: CaringBridge Site Ground Truth Overview

HC	*Full Name*
CA	Cancer
ST	Surgery/Transplantation
IJ	Injury
CS	Cardiovascular/Stroke
NC	Neurological Condition
IC	Infant/Childbirth
CI	Congenital/Immune Disorder
CN	Condition Unknown
CU	Custom
OT	Other

Table 2: Health Condition Full Name

HC	*Count*	*%*
CA	82122	63.39
ST	12260	9.46
IJ	9892	7.64
CS	9683	7.48
NC	7113	5.49
IC	6933	5.35
CI	1539	1.19

Table 3: CaringBridge Reported Sites Ground Truth

health-related text to predict health information or severity [1]. However, less research studies health condition classification in OHCs. Unlike social media, OHCs face challenges acquiring the information about their users [6] required to make effective outreach decisions with other community partners and to connect users to each other.

A tension occurs when new users sign up for profiles: OHCs want the joining process to be fast, but they also want to request user information [3]. One key piece of user information is the patient's health condition; if reported by the user, it becomes much easier to match patients with similar conditions and to recommend relevant health information [2, 4]. However, not every user report their health condition when joining the site. It is thus useful to accurately infer patients' health conditions from their early text posts on the site. We analyze data from one OHC, CaringBridge, to develop a predictive model of health conditions.

In the following sections, we report the details of our current work. First, we describe our partnership with CaringBridge; then we share the details of the classifier we built to predict health condition. Next, we report initial results on the model evaluation and conduct a preliminary error analysis. Finally, we provide design implications from our work and research.

Partnership Description

CaringBridge.org is an online health community that enables patients and non-professional caregivers to create personal, private blogs to facilitate support by writing about patients' health journeys. We partnered with CaringBridge to identify the health conditions that the site's users discuss and to identify patterns in the user's choice to explicitly report their health condition. This work was conducted under CaringBridge's terms of service and was reviewed by an IRB. While CaringBridge shares their user's data willingly, we consider that our model surfaces data not explicitly provided by the patient. Future work is needed to identify the motivations of users choosing not to explicitly report their health condition before our model could be ethically deployed.

Platform Description

We use the following language to refer to CaringBridge content:

- Health condition (HC): used here to refer to general categories of health conditions. This taxonomy of health conditions was defined by CaringBridge. The health conditions and their abbreviations used throughout the text are presented in Table 2. We omit the categories CN, CU, and OT as they contain a noisy mixture of several unrelated health conditions.
- Site: A public or private personal blog provided by CaringBridge to share one's health journey. When creating a site, the site creator can choose to report at most one HC. We refer to sites that report a health condition as "reported" sites, whereas we refer to sites that do not report a health condition as "non-reported" sites. Sites contain written journal entries, referred to here as "journals". We omit inactive sites from our analysis, defining active sites as having written a site description and at least three journals.

Dataset Overview

In the CaringBridge dataset, only 39.2% of total sites are reported sites. After omitting inactive and noisy sites, the ground truth number of reported and non-reported sites

HC	F_1	*Recall*
CA	0.95	0.91
ST	0.65	0.66
IJ	0.86	0.91
CS	0.71	0.73
NC	0.55	0.60
IC	0.91	0.95
CI	0.40	0.66
avg.	0.86	0.86

Table 4: Model Performance Summary

reduces to 129,542 and 88,148 respectively, as shown in Table 1. Reported sites are stratified split into a training and test set at the ratio of 70/30. As shown in Table 3, HCs are imbalanced: more than half of reported sites are Cancer.

Method

We assume that users who have the same health condition are likely to use similar words with similar frequency in their health journals. We tokenized the text of the journals, removing stopwords and lemmatizing words to a common form using NLTK's WordNet interface. We transformed the pre-processed data into unigram TF-IDF matrices, then we trained our model using a Linear Support Vector Machine (SVM) with stochastic gradient descent (SGD) model from the SCIKIT-LEARN package with 10-fold cross validation on the training set. We evaluated our model using weighted F_1 score and recall as our key performance metrics, focusing on one-vs-rest multiple classification. Due to the class imbalance, we apply repeated oversampling on the minority classes. Experiments with other modeling and sampling choices were found to be inferior.

In addition, we calculate the most informative features by ranking the SVM coefficients assigned to the words for each HC. These words are represented by their TF-IDF score, which means that if a word is widely used in one HC, but rarely used in other HCs then that word is rated as "informative" for that HC.

Result and Limitations

Table 4 indicates that the classifier achieves an average weighted F_1 score and recall of 0.86. The mean of 10-fold cross validation weighted F_1 score on the training set is also 0.86. Weighted F_1 score decreases to 0.65 if the omitted categories OT and CU are included. Overall, the model performs badly on ST, NC, and CI. The most informative unigram features are given in Table 5, which indicates a new source to enrich consumer health vocabulary, a lay language that is used by non-professionals to describe their health issues [5]. We conduct an analysis of the classifier's errors by randomly selecting 40 misclassified sites in HCs: 20 false positives and 20 false negatives. These 240 sites are notated as the "error set."

We found that 132 of 240 sites discussed multiple HCs. For example, 20 of 40 CA patients wrote details of surgeries, 18 of 40 ST patients stated that they had surgeries due to other health issues, e.g., they had Cancer or Cardiovascular problem. Similarly, 22.5% of IJ patients indicated that their injuries caused brain problem, which is a sub-category of NC. Moreover, in 65% of the misclassified CI sites, newborn patients are reported to have congenital heart issues, which is an overlap of CS and IC. Hence, we can conclude that CaringBridge patients discuss multiple health conditions. Moreover, we found 35% of sites in the error set were given true predictions. For example, six CA sites wrote surgery plans and details in their early posts while making no mention of cancer. Our model predicts these sites as ST, considered an error, whereas the predictions should be true given the context. We found 27% of sites in the error set are true misclassifications.

Our model has several limitations that can be addressed by future studies. First, while we utilize unigram features for their explainability, many other feature sets are possible. In particular, more complex representations of the text may better handle contextual word use. For example, "little" is an informative word in IC to refer to

HC	*Top 10 Unigram Features*
CA	cancer, chemo, leukemia, treatment, chemotherapy, oncologist,radiation, stage, node, breast
ST	transplant, donor, fusion, recovery, craniosynostosis, list, liver, curve, knee, kidney
IJ	accident, injury, fracture, burn, trauma, broken, break, injure, fell, neck
CS	stroke, attack, bypass, heart, cardiac, aneurysm, blockage, suffer, massive, speech
NC	seizure, als, al, parkinson, tumor, alzheimer, epilepsy, dementia, brain, headache
IC	milk, contraction, nicu, weigh, pregnancy, bear, baby, little, ultrasound, preemie
CI	lupus, ms, defect, cf, autoimmune, congenital, immune, cardiologist, hlhs

Table 5: Top 10 Most Informative Words, Calculated During the Classification Process Using Test Set

newborn patients as shown in Table 5. However, if a site from another HC frequently uses this word in its early posts, such as "... have little concern about ...", it is likely to be falsely classified as IC. We found four predicted IC sites in the error set facing similar issues. We anticipate that the use of "black box" deep learning approaches may address this concern. Second, we did not investigate the applicability of these models for OHCs other than CaringBridge. We encourage other researchers to triangulate our results in their work with other communities.

Implications and Conclusion

We present machine learning models to accurately classify health condition from CaringBridge posts. In partnering with CaringBridge to identify the health condition of sites, we provide the groundwork for CaringBridge to reach out to community partners appropriate to the health conditions of its users and to design the site to better connect users with related conditions. We extract features from our models to identify words used by site authors that uniquely predict the reported health condition, providing data for the expansion of consumer health vocabulary lexicons. Our models' explainable features may be useful for the creation of future health tools designed to connect authors writing about the same health condition.

An additional implication of our work is that OHCs, especially for OHCs who provide blog services to patients like CaringBridge, should consider allowing users to report multiple conditions as a single health condition category may be insufficient. Future work is necessary to understand the behavior of users who choose to report or not report their health condition and to begin the design of holistic health tools that take into account patients' health conditions.

Acknowledgements

We would like to thank the CaringBridge management team and the Minnesota Supercomputing Institute (MSI) at the University of Minnesota. This work is supported by CaringBridge and the National Science Foundation (grant 1464376).

References

[1] Chancellor, S., Lin, Z., Goodman, E. L., and et al. Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. In *Proc. of CSCW'16* (New York, NY, USA, 2016), 1171–1184.

[2] Frost, Jeana H & Massagli, M. P. Social uses of personal health information within patientslikeme, an online patient community: what can happen when patients have access to one another's data. *JMIR 10*, 3 (2008).

[3] Gross, R., and Acquisti, A. Information revelation and privacy in online social networks. In *Proc. of Privacy in the Electronic Society*, ACM (2005), 71–80.

[4] Hara, Noriko & Foon Hew, K. Knowledge-sharing in an online community of health-care professionals. *Information Technology & People 20*, 3 (2007), 235–261.

[5] He, Z., and et al. Enriching consumer health vocabulary through mining a social q&a site. *J. of Biomedical Informatics 69*, C (May 2017), 75–85.

[6] Newman, M. W., and et al. It's not that i don't have problems, i'm just not putting them on facebook: Challenges and opportunities in using online social networks for health. In *Proc. of CSCW'11* (New York, NY, USA, 2011), 341–350.

[7] Sadah, S. A., Shahbazi, M., Wiley, M. T., and Hristidis, V. Demographic-Based Content Analysis of Web-Based Health-Related Social Media. *JMIR 18*, 6 (2016), e148.