Using AI to Enhance Peer-to-Peer Moderation in Mental Health Forums

Yada Pruksachatkun[1], Sachin Pendse[1], Amit Sharma[2]

[1]) New York University Center for Data Science [yp913@nyu.edu]
[1]) Microsoft Research India [t-sapen@microsoft.com]
[2]) Microsoft Research India [amshar@microsoft.com]

Mental health is an integral part of wellbeing, mentioned prominently in the United Nations sustainable development goal of "good health and well-being". In many parts of the world, however, access to mental health support services is limited. In the United States, the US Department of Health and Human Services has projected that between 40 and 45 million people who may have needed help did not receive help [1]. Further, over 80% of the global population with mental health disorders live in developing countries, where access to mental healthcare is worse. Shortage of clinical support has partly led many people to participate in online mental health forums for seeking help. In forums such as Talklife [2], 7CupsofTea [3], and subreddits on Reddit such as r/depression [**?**], people can post and interact with others to provide peer support. This proposal aims to amplify the effectiveness of peer support for mental health through online communities.

In controlled settings, online peer-to-peer support has been shown to be effective in helping people in emotional and mental distress [4] [5] [6]. However, in real-world platforms that often have thousands to millions of users, our analysis reveals that a majority of posts by people do not receive adequate support. Currently, platforms such as 7CupsOfTea and TalkLife use human moderators to read through and identify threads in need of more expert intervention [7]

As noted by Choudhury and Kiciman [8], AI-based tools to support volunteers in moderating can be useful by automatically identifying detection of threads in need of human moderation. By doing so, human moderators can focus on helping people who have not been helped by peers instead of having their focus diluted by scanning and flagging threads. This will in turn better ensure the quality of responses in mental health forums. However, a key challenge in automatic detection is in quantifying posts that have not received effective support.

Previously, our team has worked on machine learning algorithms to detect threads that have a higher probability of needing expert moderators with input from clinical psychologists and mental health professionals. We focused on detecting moments of change (MOC), which we define as positive change in sentiment for the OP on a topic that caused the OP distress, over the course of a conversation in mental health forum TalkLife. We chose to focus on moments of change as it is indicative of the emotional state and trajectory of a person using a mental health forum. It is also indicative of how effective a thread is, since threads that are more effective will achieve a moment of change for the OP. Using gradient boosted trees (XGBoost model) with linguistic and metadata-level features, we were able to accurately detect when moments of change occur in threads (or do not) with an accuracy of 0.88, and predict if a post by the OP will contain a moment of change with an accuracy of 0.92. We also created culture-specific datasets and trained specifically on Talklife data over Indian and non-Indian datasets, where we found that MOC-detection is sensitive to cultural differences. Our results are more accurate than typical baselines, such as using the number of replies to a post as an indicator of MOC.

1 A Proposal for an AI-based platform for community moderation

Our models can be used to predict moments of change in the thread and post-level of mental health forums. These models can be used in two ways: internally by the mental health forum companies, and externally by public volunteers. Internally, our models can help online forums to route threads in need of moderation to experienced counsellors. We plan on creating a web application suite to enable volunteers to identify threads that are in need of additional outside moderation. The suite will include two parts: a platform that will link the user to threads that are in need of further attention and a visualization platform to broadly see the state of various forums. Externally, we can open up the same platform to the regular users of mental health forums, through which volunteers who would like to help others can find forums to contribute to. However, a key point of consideration when making a public tool will be in ethics and privacy, as well as ensuring that the volunteers have sufficient training and context to provide effective support.

Specifically, the web application will allow volunteers to scroll through a RSS-like feed of threads that are in need of moderation, and filter based on topic or category. There will also be visualizations to see the effectiveness of threads across various communities and sub-forums in a forum website. This will take the form of a map visualization, with various forums displayed in various parts of the graph similar to Figure 1. These visualizations will allow companies and volunteers of a mental health forum platform to gauge which sub-forums and sub-communities to focus their moderation efforts to.

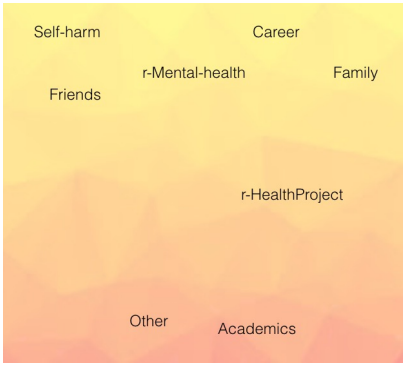

Figure 1: An example of a gradient visualization of the effectiveness of various sub-forums in a platform. In this visualization, yellow means more effective and red means less effective.

As a next step, we hope to partner with digital mental health companies to help them increase the effectiveness of support on their platforms. For continuing work on building AI models, we have established data partnerships with Talklife and 7CupsofTea. With this solution we hope to create ways that AI can help route community moderation effectively to ensure that people who come online for help receive the help they seek.

References

[1] "National projections of supply and demand for selected behavioral health practitioners: 2013-2025." [Online]. Available: https://bhw.hrsa.gov/sites/default/files/bhw/health-workforce-analysis/research/projections/behavioral-health2013-2025.pdf

[2] "Talklife," https://talklife.co/.

[3] "7cupsoftea," https://www.7cups.com/.

[4] K. M. Griffiths, A. J. Mackinnon, D. A. Crisp, H. Christensen, K. Bennett, and L. Farrer, "The effectiveness of an online support group for members of the community with depression: a randomised controlled trial," *PloS one*, vol. 7, no. 12, p. e53244, 2012.

[5] R. R. Morris, S. M. Schueller, and R. W. Picard, "Efficacy of a web-based, crowdsourced peer-to-peer cognitive reappraisal platform for depression: randomized controlled trial," *Journal of medical Internet research*, vol. 17, no. 3, 2015.

[6] K. O'Leary, S. M. Schueller, J. O. Wobbrock, and W. Pratt, "suddenly, we got to become therapists for each other: Designing peer support chats for mental health," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 331.

[7] "Safeguarding." [Online]. Available: https://talklife.co/safeguarding/

[8] E. Choudhury, Munmum de Kiciman, *Integrating Arti cial and Human Intelligence in Complex, Sensitive Problem Domains: Experiences from Mental Health.* AI Magazine, 2018.