**README document for Improving Peer Online Forums (iPOF) Datasets**

**Forum Data Analysis**

**Prof. Fiona Lobban, Spectrum Centre, Department for Mental Health Research, Lancaster University.** [**F.lobban@lancaster.ac.uk**](mailto:F.lobban@lancaster.ac.uk)

**Date of data collection: 1/3/2022-31/12/2024**

This project is funded by the National Institute for Health and Social Care Research (NIHR), 134035. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Within the “WS2 Forums” folder, there are a set of subfolders, one per forum or subforum. See the raw data upload for the mapping between the bird names and the real forum names. Each forum folder contains one or more “tar.gz” files. Tarfiles can be expanded using 7-Zip or WinZip in Windows, or via the ‘tar’ command on Linux and MacOS command lines.

Each tarfile contains the results of data and format clean up performed by Matthew Coole and John Vidler of the UCREL NLP group (<https://ucrel.lancs.ac.uk/>) using bespoke Python scripts. The analysis scripts take the raw forum data and encode them with the Text Encoding Initiative format for Computer Mediated Communication (<https://wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication>). The encoded data is then annotated via a bespoke Natural Language Processing (NLP) pipeline which includes tokenisation, lemmatisation, part-of-speech tagging, semantic tagging, emotion and sentiment analysis. Further bespoke scripts perform frequency analysis of emotions and sentiments, key semantic tags, key words, n-grams, topic modelling, and post times (where this data is available to us). Results of these analyses can be found within the “Analysis Files” subfolder.

**This dataset should be kept closed and not shared. It is on PURE for storage only.**

Methodological information is available from the study protocol and other project publications:

Glossop, Z., Čermáková, A. L., Marshall, P., Rayson, P., Robinson, H., Semino, E., & Lobban, F. (2025). Who uses mental health support forums, and why? Triangulating findings from surveys, interviews, and forum posts. medRxiv. <https://doi.org/10.1101/2025.05.11.25327409>

Marshall, P., Booth, M., Coole, M., Fothergill, L., Glossop, Z., Haines, J., Harding, A., Johnston, R., Jones, S., Lodge, C., Machin, K., Meacock, R., Nielson, K., Puddephatt, J.-A., Rakic, T., Rayson, P., Robinson, H., Rycroft-Malone, J., Shryane, N., ... Lobban, F. (2024). Understanding the Impacts of Online Mental Health Peer Support Forums: Realist Synthesis. JMIR Mental Health, 11, Article e55750. <https://doi.org/10.2196/55750>

Coole, M., Rayson, P., Glossop, Z., Lobban, F., Marshall, P., & Vidler, J. (2024). Exploring the Suitability of Transformer Models to Analyse Mental Health Peer Support Forum Data for a Realist Evaluation. In D. Demner-Fushman, S. Ananiadou, P. Thompson, & B. Ondov (Eds.), Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024 (pp. 184-188). ELRA and ICCL. <https://aclanthology.org/2024.cl4health-1.22>

Coole, M., Marshall, P., Rayson, P., & Lobban, F. (2023). iPOF: Improving Peer Online Forums Update. Poster presented at Healthcare text analytics conference 2023, Manchester, United Kingdom.

Lobban, F., Coole, M., Donaldson, E., Glossop, Z., Haines, J., Johnston, R., ... & Wise, S. (2023). Improving Peer Online Forums (iPOF): protocol for a realist evaluation of peer online mental health forums to inform practice and policy. BMJ open, 13(7), e075142.

Lobban, F., Rayson, P., & Coole, M. (2022). iPOF: Improving Peer Online Forums. Poster presented at HealTAC 2022: 5th Healthcare Text Analytics Conference, United Kingdom. https://youtu.be/ryA\_obyJnfU

Ethical approval: Solihull Research Ethics Committee. 20 June 2022 (IRAS314029).