Exploring the Suitability of Transformer Models to Analyse Mental Health Peer Support Forum Data for a Realist Evaluation

Matthew Coole[1], **Paul Rayson**[1], **Zoe Glossop**[2], **Fiona Lobban**[2], **Paul Marshall**[2], **John Vidler**[1]
[1]UCREL Research Centre, Computing and Communications, Lancaster University, UK
[2]Spectrum Centre, Health and Medicine, Lancaster University, UK
Lancaster University
{m.coole, p.rayson, z.glossop, f.lobban, p.marshall4, j.vidler}@lancaster.ac.uk

Abstract

Mental health peer support forums have become widely used in recent years. The emerging mental health crisis and the COVID-19 pandemic have meant that finding a place online for support and advice when dealing with mental health issues is more critical than ever. The need to examine, understand and find ways to improve the support provided by mental health forums is vital in the current climate. As part of this, we present our initial explorations in using modern transformer models to detect four key concepts (connectedness, lived experience, empathy and gratitude), which we believe are essential to understanding how people use mental health forums and will serve as a basis for testing more expansive realise theories about mental health forums in the future. As part of this work, we also replicate previously published results on empathy utilising an existing annotated dataset and test the other concepts on our manually annotated mental health forum posts dataset. These results serve as a basis for future research examining peer support forums.

Keywords: mental health, peer support, transformers, machine learning

1. Introduction

Our project, related to improving the understanding and analysis of mental health peer online forums (iPOF[1]), aims to understand how mental health peer online support forums work and how we can improve them. The project is undertaking a realist evaluation (Pawson and Tilley, 1997), drawing together existing knowledge in a realist synthesis (Pawson et al., 2005), generating programme theories in the form of a series of Context, Mechanism, and Outcome (CMO) configurations. CMO configurations are short explanatory statements that articulate how health and social care programmes achieve their impacts. An example of one of the project's many CMOs is:

When forums bring together people with shared mental health experiences (context), forum users will have access to mental health narratives that resonate with their own (mechanism), leading to improved sense of social connectedness (outcome).

The project's CMOs will be evaluated using a mixed methods approach including surveys, qualitative interviews, linguistic analysis and NLP techniques. These different approaches will be triangulated together to create insights and recommendations that can be used by forum hosts and commissioners to build and manage more effective online peer support communities. This paper presents a set of NLP models built as an initial means of evaluation of some of our CMO configurations. Some of the features we need to investigate have been studied before in NLP. However, some are completely novel, and no comparative evaluations are possible.

The data collected in the iPOF project stems from several forum partners who agreed to participate in the project. There are eight forum partners in total, each participating at different levels. Language data has been gathered from seven forum partners; this includes any posts and replies made by users to the forums. There are several ethical considerations when gathering data, as some forums are openly accessible, some are closed but free to sign up to, and some require application or referral. This led to different procedures for different forum users to give informed consent for their data to be collected:

- opt-out: The project was advertised on open forums and users who did not wish their data to be collected could opt-out.
- implicit opt-in: Some closed forums collect consent to research from users at sign-up.
- explicit opt-in: Some closed forums (without a research consent option) required the collection of explicit consent from individual users or where forum use is dependent on giving research consent.

Each forum was given a bird name code by which it was referred to as part of our anonymisation procedures. In this paper, we will utilise the Starling corpus. The Starling corpus was selected to build the dataset and models as it was one of the larger and open forums participating in the project.

Based on the initial set of CMO configurations, we identified a set of common features or concepts which were crosscutting across multiple pro-

[1]https://www.lancaster.ac.uk/ipof/

Proceedings of CL4Health 2024 @LREC-COLING-2024, pages 184–188

gramme theories. Thus, we annotated data and built models to predict when users shared a lived experience narrative in their posts and when a post showed connectedness, empathy and gratitude. These language models and experiments provide one avenue to test our programme theories and allow us to show that NLP methods can be used to scale up qualitative analysis to a much larger dataset than would otherwise be possible. We show that it is possible to develop NLP models for these concepts, and this will facilitate further work on other concepts and features that are important to a wider set of CMOs and programme theories. This will subsequently feed into guidelines for the description and improvement of peer online support forums, particularly in the areas of mental health, but potentially more widely in the future.

2. Related Work

Empathy and its detection has been explored extensively in recent NLP literature. Empathy in news articles through shared tasks has resulted in several approaches based on BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) models and its variants (RoBERTa, ALBERT) (Tafreshi et al., 2021) as well as exploring differences between demographics (Guda et al., 2021). BERT-based models, as well as LSTMs (Long short-term memory), have also been applied in the context of medical texts (Dey and Girju, 2023) as well as in online cancer survivor forum posts (Hosseini and Caragea, 2021) which also provides a dataset[2] with labelled texts for those seeking and giving empathy.

Detecting and using gratitude for extrinsic analysis has been explored for various reasons. Recently, many techniques have been applied to the detection of gratitude. BiLSTMs (Bidirectional LSTM) have been utilised for detection in obituaries (Sabbatino et al., 2020), RNNs (Recurrent Neural Networks) in online question and answer dialogues (Noseworthy et al., 2017), SVMs (Support Vector Machines) in online tweets (Danescu-Niculescu-Mizil et al., 2013) as well as BERT based models in online therapy texts (Burkhardt et al., 2022). In health forums specifically, KNN (K-Nearest Neighbours) and Naive Bayes have been applied (Sokolova and Bobicev, 2013).

Unlike empathy and gratitude, little previous work explores the automatic detection of lived experience narratives from online texts. The closest existing approach that could be comparable is the detection of personal medical disclosures (Valizadeh et al., 2021; Arseniev-Koehler et al., 2018). However, personal medical disclosures do not fall into the same category of interest as discussion around lived experience, which go far beyond such disclosures and may include many things such as coping strategies, treatment experience and emotional support received from family and friends. Connectedness, as it is defined within the CHIME (Connectedness, Hope & optimism, Identity, Meaning, Empowerment) framework (Leamy et al., 2011), has yet to be explored through automatic detection.

Our review of existing literature around the detection of our four key concepts (connectedness, empathy, experience and gratitude) shows the varying degrees to which these concepts have been explored, from richly with available labelled datasets (empathy) to poorly defined vague concepts that need refinement and investigation (connectedness).

3. Methodology

3.1. Data

The dataset we chose to annotate for the concepts of connectedness, lived experience, and gratitude was the Starling corpus (for empathy, we made use of an existing dataset). This corpus was built from threads of a popular social media site with a sub-section for supporting mental health. This forum was selected from our forum partners as it was open with consent from users falling into the opt-out category (no users opted out). The dataset for this forum consists of 47k posts in 10k threads from 6k users. The tokenised dataset contains roughly 5.5M words.

3.2. Annotation

The Starling dataset was manually annotated for the key concepts connectedness, lived experience and gratitude. The annotated dataset was generated by sampling a random 2,000 posts and annotating each post for the three key concepts. A primary annotator tagged all 2,000 posts, and a secondary annotator also annotated a 5% subset to ensure the annotation guidelines were clear and to check for inter-annotator agreement (94.6%). For the empathy annotations, we made use of an available dataset annotated on cancer forum data (Hosseini and Caragea, 2021) as the medium of an online support forum is comparable to our own data, we attempt to replicate their results.

3.2.1. Connectedness

Connectedness is a very difficult concept to identify and annotate in online text. For our purposes, we relied on a practical definition from the CHIME framework (Jagfeld et al., 2021). This meant we

[2]https://github.com/Mahhos/Empathy

were looking for people discussing connecting with others. This may include:

- Discussion of peer support groups (their availability or experience of)
- Supportive relationships with family and friends, as well as intimate relationships
- Support from professionals, including helpful (and unhelpful) interactions with therapists and healthcare professionals.
- Discussion of actively participating in the community or online group.

3.2.2. Lived Experience (Narratives)

Lived experience narratives were annotated on the basis that the post included a personal experience of the user relating to their mental health or their experience caring for another with a mental health problem. This may include:

- Personal encounters or challenges with mental health services or treatments.
- Use of medications and their effects.
- Coping mechanisms employed by the individual.
- Interactions, both helpful and unhelpful, with friends and family.
- Personal accounts of emotional, professional, or personal issues stemming from their mental health condition.

3.2.3. Gratitude

Gratitude was annotated with particular emphasis on when users were showing gratitude to others in the forum. This was often shown by replies with very simple terms ("thanks :)", "cheers!"), but the annotators also attempted to pick out more subtle indicators where only looking for specific wording may miss e.g. "That's really helpful! :)". Posts where specific terms of gratitude were used that appeared to merely be expressions of politeness were also not annotated as gratitude. This was very common in the data after initial posts on threads where terms of gratitude were used simply as a sign off to the post.

These concepts were annotated in this way as an initial exploration into some of the concepts that appear in our programme theories. These concepts will need to be developed further based on further refinement of what they mean within the context of different CMO configurations and based upon how their automatic detection can be used to test our theories. The concept of connectedness, in particular, is highly likely to require refinement and further annotation before it can be applied to theory testing.

Model	Precision	Recall	F1-Score	Support
Connectedness				
Naive Bayes	0.57	0.57	0.57	64
SVM	0.47	0.49	0.35	64
Random Forest	0.68	0.65	0.61	64
Distilbert	0.84	0.83	0.82	64
RoBerta	0.82	0.81	0.81	64
Empathy				
Naive Bayes	0.75	0.44	0.44	1001
SVM	0.71	0.49	0.50	1001
Random Forest	0.66	0.53	0.55	1001
Distilbert	0.79	0.79	0.79	1001
RoBerta	0.82	0.81	0.82	1001
Experience				
Naive Bayes	0.75	0.64	0.58	243
SVM	0.72	0.72	0.71	243
Random Forest	0.76	0.76	0.76	243
Distilbert	0.86	0.86	0.86	243
RoBerta	0.93	0.93	0.93	243
Gratitude				
Naive Bayes	0.81	0.81	0.81	91
SVM	0.84	0.80	0.79	91
Random Forest	0.92	0.91	0.91	91
Distilbert	0.98	0.98	0.98	91
RoBerta	0.98	0.98	0.98	91

Table 1: Model metrics (macro averages)

3.3. Models

Our annotated dataset was undersampled on the most common class to create three balanced datasets for connectedness, lived experience and gratitude. An existing dataset for empathy was used. A 60/20/20 train/test/validate split was used to ensure there was sufficient test data for some of the less frequently annotated concepts.

To compare our results, we also created baselines using Naive Bayes, SVM and Random Forest methods. For these baseline methods, we used word level TF-IDF (Term Frequency Inverse Document Frequency) vectors to train and used the popular SKLearn library[3].

The BERT-based models we chose to use were DistilBERT (Sanh et al., 2019) and RoBERTa using the transformers library[4]. These were hyperparameter tuned using Optuna[5] with 50 runs varying learning rate, training and evaluation batch sizes. The final models were then trained over 30 epochs to assess their training loss and learning rate (see Figure 1).

4. Results

The results of our best model runs (with tuned hyperparameters) on our validation set are shown in Table 1. These results were taken when the best metric (F1 score) was found on our final training run (up to 30 epochs). In some cases, F1 score began to deteriorate (possibly as the result of overfitting).

[3]https://scikit-learn.org/stable/index.html

[4]https://huggingface.co/docs/transformers/index

[5]https://optuna.org/

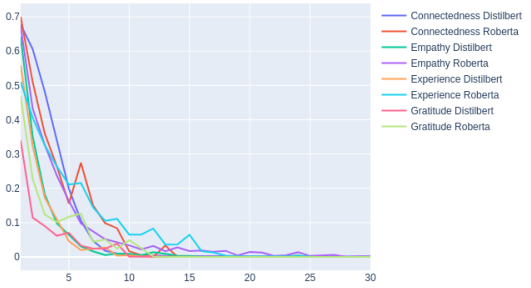

Figure 1: Model loss during training

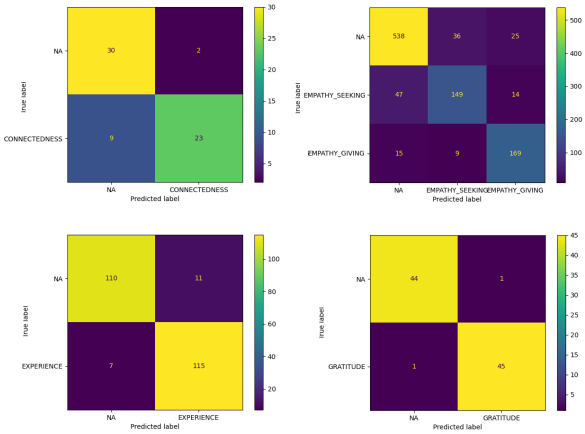

Figure 2: Confusion Matrices

We find that across all four concepts, BERT-based models outperform all baseline models, often by a substantial margin. Confusion matrices are shown in Figure 2.

Connectedness appeared to show little difference in performance between DistilBERT and RoBERTa, but they did both outperform Random Forest, which was the best baseline model. It could be that for connectedness (as this concept needs to be developed further and there is a low support) in the future these results may change. RoBERTa achieved the highest F1 score in the detection of experience; this concept was the only one where there was a notable difference between RoBERTa and DistilBERT. Gratitude is the only concept where one of the baseline models was comparable to the BERT-based models. Random Forest achieved excellent F1 scores for the detection of gratitude on our dataset, but DistilBERT and RoBERTa still proved to be the most effective in the detection of this concept.

Our experiments looking at the concept of Empathy using a pre-existing dataset achieved similar results to the original paper, which utilised the original BERT model. Tuned RoBERTa and DistilBERT models achieved F1 scores of 0.82 and 0.79 respectively, which are comparable to the original published score of 0.74 using BERT. The baseline models were also similarly comparable to the original publication. 0.44, 0.50, 0.55 for Naive Bayes, SVM and Random Forest respectively vs 0.39, 0.59, 0.55.

5. Conclusion

We have presented our initial exploration of building models for the detection of four key concepts related to mental health peer support forums: connectedness, empathy, experience and gratitude. We have demonstrated that modern transformer-based models (DistilBERT and RoBERTa) outperform classic baseline models (Naive Bayes, SVM, Random Forrest). We have also replicated previous work utilising BERT models on an existing empathy dataset.

Future work will take these models and apply them across seven datasets that have been collected as part of the iPOF project. We will then explore ways that they can be used to test a set of realist theories by looking for correlations between these and other annotations, e.g. Does the sharing of lived experience lead to an increase in sentiment in a forum thread? Does receiving expressions of empathy contribute to how active a forum user becomes?

6. Ethics Statement

There are very important ethical issues in analysing forum posts. People often share details about the things that are causing them distress, in the hope that other people who have faced similar problems can help them. It is vital that the forum feels a safe space in which to do this. We do not want this research to jeopardise this feeling of safety in any way. Therefore, we have developed a comprehensive ethical framework for this study. This has been developed with input from legal, clinical, academic and lived expertise, and approved by the Health Research Authority (IRAS 314029). As the project progresses we may need to make changes to how the study is conducted. Any changes will be approved by the study sponsor and the ethics committee and will be updated online for information[6].

[6]https://www.lancaster.ac.uk/ipof/

7. Acknowledgements

This study is funded by the NIHR Health and Social Care Delivery Research (HS&DR) (NIHR134035). The study is hosted by Berkshire Healthcare NHS Foundation Trust. The sponsor is Lancaster University.

8. Bibliographical References

Alina Arseniev-Koehler, Sharon Mozgai, and Stefan Scherer. 2018. What type of happiness are you looking for?-a closer look at detecting mental health from language. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 1–12.

Hannah Burkhardt, Michael Pullmann, Thomas Hull, Patricia Areán, and Trevor Cohen. 2022. Comparing emotion feature extraction approaches for predicting depression and anxiety. In *Proceedings of the eighth workshop on computational linguistics and clinical psychology*, pages 105–115.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Priyanka Dey and Roxana Girju. 2023. Investigating stylistic profiles for the task of empathy classification in medical narrative essays. *arXiv preprint arXiv:2302.01839*.

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. Empathbert: A bert-based framework for demographic-aware empathy prediction. *arXiv preprint arXiv:2102.00272*.

Mahshid Hosseini and Cornelia Caragea. 2021. It takes two to empathize: One to seek and one to provide. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13018–13026.

Glorianna Jagfeld, Fiona Lobban, Paul Marshall, and Steven H Jones. 2021. Personal recovery in bipolar disorder: Systematic review and "best fit" framework synthesis of qualitative evidence–a poetic adaptation of chime. *Journal of affective disorders*, 292:375–385.

Mary Leamy, Victoria Bird, Clair Le Boutillier, Julie Williams, and Mike Slade. 2011. Conceptual framework for personal recovery in mental health: systematic review and narrative synthesis. *The British journal of psychiatry*, 199(6):445–452.

Jasy Suet Yan Liew and Howard R Turtle. 2016. Exploring fine-grained emotion detection in tweets. In *Proceedings of the NAACL student research workshop*, pages 73–80.

Fiona Lobban, Matthew Coole, Emma Donaldson, Zoe Glossop, Jade Haines, Rose Johnston, Steven H Jones, Christopher Lodge, Karen Machin, Paul Marshall, et al. 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.

Michael Noseworthy, Jackie Chi Kit Cheung, and Joelle Pineau. 2017. Predicting success in goal-driven human-human dialogues. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 253–262.

Ray Pawson, Trisha Greenhalgh, Gill Harvey, and Kieran Walshe. 2005. Realist review-a new method of systematic review designed for complex policy interventions. *Journal of health services research & policy*, 10(1_suppl):21–34.

Ray Pawson and Nick Tilley. 1997. *Realistic evaluation*. sage.

Valentino Sabbatino, Laura Bostan, and Roman Klinger. 2020. Automatic section recognition in obituaries. *arXiv preprint arXiv:2002.12699*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Marina Sokolova and Victoria Bobicev. 2013. What sentiments can be found in medical forums? RANLP.

Shabnam Tafreshi, Orphée De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. Wassa 2021 shared task: Predicting empathy and emotion in reaction to news stories. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–104. Association for Computational Linguistics.

Mina Valizadeh, Pardis Ranjbar-Noiey, Cornelia Caragea, and Natalie Parde. 2021. Identifying medical self-disclosure in online communities.