Mental-Mentor: Assessing Utility Of Comments On Mental-Health Forums

Cheng-Chen Huang[1], Avanthika Rajesh[2], Vaibhav Garg[1]

[1*]Department of Computer Science, Virginia Tech, Innovation Campus, Alexandria, Virginia, United States.
[2]Department of Computer Engineering, Virginia Tech, Blacksburg, Virginia, United States.

Contributing authors: jjooeeyh@vt.edu; avanthikarajesh@vt.edu; vaibhavg@vt.edu;

Abstract

Warning: This paper may contain triggering language for some readers, especially for people dealing with mental health issues such as depression and anxiety.
Mental health has turned out to be a serious problem these years. People sometimes leverage social media to share their traumas, emotions, feelings, and in turn seek help from fellow platform users. Reddit is one such platform to facilitate such support seeking. In this domain, the majority of prior research focuses on classifying posts that describe authors' depression or anxiety. However, there has not been much work to assess the quality of comments on such posts.
We build a computational tool, Mental-Mentor that predicts the utility of the comment based on two factors: (1) if the comment grabs attention of the original poster and (2) if the comment is helpful to the original poster.
Mental-Mentor consists of two concatenated models: one-class autoencoder followed by fine-tuned Tiny-Llama. Autoencoder model predicts if the comment can grab original poster's attention and the Tiny-Llama predicts if it's helpful or not.
On five-fold cross validation, the autoencoder and Tiny-Llama achieve average macro F1 score of 93.34% and 72.00%, respectively. Through Mental-Mentor, we also assessed the utility of comments on a new subreddit called r/Anxiety. We found that 21.5% of comments collected from this subreddit are having low utility scores. This suggests support-seeking can be improved on such subreddits, which may also accelerate future research in this domain.

Keywords: NLP, Mental Awareness,Models, Machine Learning, Anxiety, Depression, LLM, Social Media Analysis

1 Introduction

Background. According to Forbes, 36.2% of people between 18 and 25 years of age, 29.4% of people between 26 and 49 years of age and 13. 9% of people over 50 years of age suffer from mental illness [1] . Reddit, a popular social networking website, has multiple subreddits (specialized forums) for mental health related discourse. On such subreddits, users (called *orignial posters or OP*) post their problems anonymously and receive comments from community members (called *helpers*). Subreddits such as `r/Anxietyhelp` and `r/depression_help` are classic examples of such support-seeking communities.

Problem Settings. Previous studies have focused on analyzing the content of the posts written by OPs [2–4]. However, in order to aid the OP, the comments received should be supportive and helpful. Receiving comments with low utility may not serve the purpose of these support-seeking communities. To assess the utility of comments received by OPs, we pose following research questions.

RQ1: How can we predict if a comment can grab OP's attention?
RQ2: How can we predict if a comment can be helpful to the OP?

We define ***utility*** of a comment based on two attributes. First, if the comment is able to grab OP's ***attention***. Second, if the comment is able to possibly ***help*** the OP. If OP has received numerous comments to their post, they may overlook some of them. Hence, we first investigate if the given comment can grab OP's attention. Once it grabs attention, we check if it can be helpful to the context provided by the OP in their post.

To answer RQ1 and RQ2, we stick to two subreddits, namely, `r/Anxietyhelp` and `r/depression_help`. Using posts and comments in these subreddits, we train our computational tool, Mental-Mentor, consisting of an autoencoder and Tiny-Llama. The autoencoder predicts if the given comment can catch OP's attention or not (RQ1), whereas the fine-tuned Tiny-Llama predicts the helpfulness of the comments grabbing attention (RQ2). Mental-Mentor generates utility score for each given comment. If the comment is unable to grab attention, it is assigned a utility of -1. If it is able to grab attention but still may not be helpful, it is assigned 0. On the other hand, if it is able to grab attention and also helpful, it is assigned 1. We consider -1 and 0 as low utility scores, however 1 as a high score.

Contributions and Findings. We developed Mental-Mentor, which can predict the utility of each comment based on the content in the posts and the following comments. Mental-Mentor can be leveraged on multiple mental-health related subreddits to assess utility of the support being provided. We showcase one such use case by applying it on another subreddit, r/Anxiety. Our Mental-Mentor found that 21.5% of the new comments are having low utility score. In other words, roughly 1 in 5 posted comments is having low utility, leaving scope of improvement for helpers on such forums.

To the best of our knowledge, there have been no prior computational studies assessing the comments on such support-seeking communities. We do so by the notion of utility score, as described above.

2 Data Collection

We utilized Reddit's official API to obtain the posts and comments from r/Anxietyhelp and r/depression_help. For each post, we also collected its meta-data such as the title written by its OP. Finally, we collected 583 455 comments from r/depression_help and 268 020 comments from r/Anxietyhelp.

Since we were dealing with only textual data, we discarded posts or comments containing only images such as visual memes. Moreover, comments containing irrelevant links were also discarded. After this filtering process, we were left with a total of 834 738 comments for our training purpose.

Data Annotation. We treat this problem as multi-class comment classification, where each comment can fall into one of the following three classes. Based on the following process, we assign utility scores to 834 738 comments, which forms as a training set for our tool, MENTAL-MENTOR.

(1) Utility score is -1: The comments which OP did not find important to respond to fall in this category. For example, the comment shown below, "Ok thanks", may not grab OP's attention. Hence, the low utility score of -1. We consider all these comments for which OP did not reply, hence the OP did not resonate enough as -1 label.

Post Content: "...I switched from lexapro to Buspirone and it doesn't control my anxiety as well..."

Comment: "Ok thanks"

(2) Utility score is 0: The comments that can attract the attention of the OP (meaning the OP replied) but they may not find them helpful. For example, the OP's response to the below comment shows that they did not understand it well. hence, the utility score is 0.

Post Content: "I'm so close to Shutting all of my Socials Down and Forget everyone as They have forgotten me, I have No Friends No Family Nobody..."

Comment: "I'm just on fb to be a sh*t stirrer and read about the favored people's lives. I like how much it stings. Praying for God to end me quick and soon."

OP's Reply: "Can you elaborate a bit? What's the deal on FB and you? Context please so I can get what you mean?"

(3) Utility score is 1: The comments that can grab the OP's attention and even be helpful as described by the OP's response

Post Content: "...I know it's gross, and I do try to maintain it, but I struggle a lot with self neglect along with other mental health issues. I attempt and keep failing...."

Comment: "Hello! Hygiene care is seen by many as a simple chore or task but for some it can be a monumental task where to have to use a lot of willpower..."

OP's Reply: "Thank you so much! That seems so simple and clever in hindsight. I shall try this!"

The OP's response clearly reflects this comment to be helpful, hence it is annotated with a high utility of 1.

3 Methodology

We concatenated the autoencoder model and fine-tuned TinyLlama to solve RQ1 and RQ2, respectively. The architecture is shown in Fig. 1. The autoencoder first classifies if the comment will be responded to by the OP or not. If the output autoencoder predicts that the comment is not going to be responded by the original post author, it is directly predicting this comment as utility score is -1. If the autoencoder predicts that the comment will be responded by the original post author, the model then can pass the comment to the prompted LLM. Prompted LLM will predict if the incoming comment is helpful to the OP or can generate resonance OP. If the prompted LLM predicts that the comment cannot generate resonant with OP or not make OP feels helpful, it's predicting this comment as label == 0. If the second model predicts that the comment can generate resonance with OP or is helpful to OP, prompted LLM will predict this comment as label == 1.

The dataset shows that the portion between three classes are highly unbalanced. There are 833 134 utility score is -1 comments, 505 utility score is 0 comments, and 1099 utility score is 1 comments. In this case, we decided to apply **Autoencoder algorithm** for utility score is -1 or not prediction. We used prompted LLM for the 0 and 1 utility score prediction.

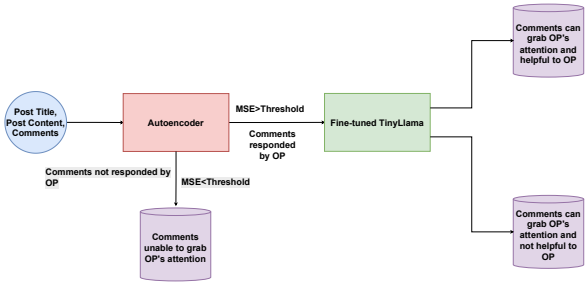

Fig. 1 The architecture of MENTAL-MENTOR, showing the two concatenated models.

Autoencoder. Autoencoder is a one-class deep neural network. It learns the pattern to compress and reconstruct the data. Autoencoder contains two parts: encoder and decoder. The encoder is responsible for compressing the input data, and the decoder is responsible for reconstructing the compressed data. Our autoencoder is a 5 layers structure architecture and the nodes in each layer is 1536-768-384-768-1536. We can use evaluation metrics such as MSE to evaluate the difference between the original data and the reconstructed data.

We introduced the MSE threshold to achieve classification. We filtered only the comments that are not capturing OP's attention as the training data, applied Universal Sentence Encoder(USE) for word embedding as USE captures the meaning of the comments, then mapped to a length is 512 vector per feature. The autoencoder tried to reconstruct the input comments then calculate the MSE compared to the original input data. We have applied 5 folds cross-validation for getting a more robust result. The average MSE mean of the 5 folds is 0.000237, and the average MSE standard deviation

is 0.000172. The higher the MSE, the lower is the autoencoder able to reconstruct, the higher the possibility that the input comment will be replied by OP. On the other hand, the lower the MSE, the higher the autoencoder is able to reconstruct the comment, the lower the comment will be replied by OP. Once we have determined the threshold, we can use the threshold to decide whether the MSE belongs to high or low.

We had leveraged Youden Index [5] then created an algorithm to determine threshold x^*. If model returns the MSE x satisfies $x > x^*$, then the input is able to capture OP's attention then can be passed to next model. Otherwise, the Label -1 is returned.

Prompt LLM. In our study, we leveraged replied comments labeled as 0 and 1 to analyze the relevance and helpfulness of responses. The labels were assigned based on whether the original poster (OP) engaged with a response, with:

- Label 1: OP engaged with the comment, indicating relevance/helpfulness.
- Label 0: OP did not engage with the comment, suggesting irrelevance or low perceived usefulness.

In our study, we employed few-shot prompting to analyze the relevance and helpfulness of responses in the subreddits `r/Anxiety_help` and `r/depression_help`. We analyzed replied comments labeled as 0 and 1, where label 1 indicated engagement from the original poster (OP), suggesting that the response was meaningful or valuable, while label 0 signified a lack of interaction. To train the models effectively, we incorporated a small set of labeled examples as prompts, enabling them to recognize patterns in conversational engagement. Several pre-trained language models, including BERT, XLNet, RoBERTa, and TinyLlama, were employed to determine their ability to identify comments that foster engagement. Fine-tuning was performed on our dataset, and model performance was assessed using Recall, F1-Score, and Accuracy. This approach allowed us to examine how different architectures interpret textual cues and contextual depth to predict whether a response is likely to sustain interaction. The results provide insights into the capacity of transformer-based models to navigate online peer-support discussions and discern which responses contribute to continued dialogue.

Model Evaluation on Original Poster (OP) Comment Interactions. We experimented with multiple transformer-based models such as BERT [6], XLNet [7], RoBERTa [8], and TinyLlama [9]. For the training and testing purposes, we used 128 tokens as maximum sequence length, 5 epochs, 16 as batch size, and 2e-5 as learning rate. Both BERT and TinyLlama have high performance, but BERT is facing an overfitting challenge. The accuracy, precision, recall, and F1 score have a significant drop in test data. We ended up selecting TinyLlama as our fine-tuned LLM model. TinyLlama is a compact 1.1B parameter language model pretrained on approximately 1 trillion tokens.

Selecting the final model for deployment, we prioritized recall and F1 score as our key evaluation metrics, as they are more indicative of a model's ability to correctly identify comments that are meaningful and engaging to the original poster. Recall reflects how well the model identifies all relevant comments, while F1-score balances recall with precision, giving a more holistic view of overall performance.

4 Results

On five-fold cross-validation, the autoencoder model shows 88.27% average macro recall along with 93.34% average macro F1 score. Moreover, the fine-tuned Tiny-Llama shows 73.18% average macro recall and 72.00% average macro F1 score.

To assess utility of unseen comments, we applied MENTAL-MENTOR on another subreddit, `r/Anxiety`. Using the same Reddit API, we collected comments from r/Anxiety for about two days. As a result, we could collect 3714 comments. After applying MENTAL-MENTOR, we found 21.5% of these comments are having low utility score (score of -1 and 0), meaning these comment either did not grab OP's attention or did grab attention but still were not helpful. We investigated these predictions by manually checking 100 comments randomly sampled. We found that 75 of these 100 comments were correctly predicted to have low utility.

5 Related Work

LLMs (large language models) have been applied for mental health related diagnosis such as detecting mental illness or depression [10]. Kim et al. [11] also point out that LLMs (large language models) are better than health care professionals in identifying Obsessive-Compulsive Disorder (OCD). Balani and De Choudhury [12] build a computational model to predict the lavel of self-disclosure in a Reddit post. Joshi et al. [13] have apply deep learning to achieve feature extraction through user behavior on social media like posts.

Other research works assess the impact of social media on users' mental health. Saha et al. [14] use casual inference technique to understand the long-term impact of online mental health communities. On the other hand, Yang et al. [15] use TinyLlama to not only identify the type of mental illness but also provide reasoning for such predictions.

In general, peer support through multiple mechanisms (both online and offline) is proven to be an effective way to help others. O'Leary et al. [16] compare and list several mediums, such as texting, email, phone call, instant messaging, which are often used to share experience and support people in need.

To our best knowledge, none of these studies focus on the interaction between comment and OP. Instead of diagnosing mental illness, we aim to analyze whether the comments posted in response to help-seeking content are meaningful, helpful, and likely to generate engagement with the original poster.

6 Conclusion

Prior research mostly focuses on the post content, however assessing the comment is also essential in order to facilitate support on mental-health related subreddits. For such assessment, we come up with the notion of assigning utility score to each comment.

Using the comments from r/Anxietyhelp and r/depression_help, we built MENTAL-MENTOR, which predicts the utility of the given comment. This tool consists of deep autoencoder framework followed by fine-tuned Tiny-Llama. Using the comments from

another subreddit, r/Anxiety, we demonstrated how MENTAL-MENTOR can be applied to assess the utility of comments on multiple platforms.

In future, MENTAL-MENTOR can be used to suggest helpers about how to draft their comments in order to reach high utility scores. Such high utility will result in providing comments to the OP that are both helpful and can grab attention. Our work also paves the way for future research to explore the type of language used in high utility comments. Empathetic and comforting language could some of the possible areas to explore. In turn, all these works would improve the way support is shared within such mental-heath communities.

References

[1] Forbes: Mental Health Statistics by Age. Resources Related to Mental Health Statistics. Accessed 2025-05-11 (2025). https://www.forbes.com/health/mind/mental-health-statistics/

[2] Kim, J., Lee, J., Park, E., Han, J.: A Deep Learning Model for Detecting Mental Illness from User Content on Social Media. Scientific Reports **10**(1), 11846 (2020)

[3] Ghosh, S., Anwar, T.: Depression Intensity Estimation via Social Media: A Deep Learning Approach. IEEE Transactions on Computational Social Systems **8**(6), 1465–1474 (2021)

[4] Hasib, K.M., Islam, M.R., Sakib, S., Akbar, M.A., Razzak, I., Alam, M.S.: Depression Detection from Social Networks Data Based on Machine Learning and Deep Learning Techniques: An Interrogative Survey. IEEE Transactions on Computational Social Systems **10**(4), 1568–1586 (2023)

[5] Schisterman, E.F., Perkins, N.J., Liu, A., Bondell, H.: Optimal Cut-point and Its Corresponding youden index to Discriminate Individuals Using Pooled Blood Samples. Epidemiology **16**(1), 73–81 (2005)

[6] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805 (2018)

[7] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: Xlnet: Generalized Autoregressive Pretraining for Language Understanding. arXiv preprint arXiv:1906.08237 (2019)

[8] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A Robustly Optimized Bert Pretraining Approach. arXiv preprint arXiv:1907.11692 (2019)

[9] Zhang, P., Zeng, G., Wang, T., Lu, W.: Tinyllama: An open-source small language model. arXiv preprint arXiv:2401.02385 (2024)

[10] Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., Ghassemi, M., Dey, A.K., Wang, D.: Mental-llm: Leveraging large language models for mental health prediction via online text data. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **8**(1) (2024) https://doi.org/10.1145/3643540

[11] Kim, J., Leonte, K.G., Chen, M.L., Torous, J.B., Linos, E., Pinto, A., Rodriguez, C.I.: Large language Models Outperform Mental and Medical Health Care Professionals in Identifying Obsessive-compulsive Disorder. NPJ Digital Medicine **7**(1), 193 (2024)

[12] Balani, S., De Choudhury, M.: Detecting and characterizing mental health related self-disclosure in social media. In: Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems. CHI EA '15, pp. 1373–1378. Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2702613.2732733 . https://doi.org/10.1145/2702613.2732733

[13] Joshi, D.J., Makhija, M., Nabar, Y., Nehete, N., Patwardhan, M.S.: Mental health analysis using deep learning for feature extraction. In: Proceedings of the ACM India Joint International Conference on Data Science and Management of Data. CODS-COMAD '18, pp. 356–359. Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3152494.3167990 . https://doi.org/10.1145/3152494.3167990

[14] Saha, K., Suh, J., Kıcıman, E., De Choudhury, M.: Measuring the Causal Impact of Online Mental Health Communities. Proceedings of the International AAAI Conference on Web and Social Media **16**(1), 671–682 (2022) https://doi.org/10.36190/2022.58

[15] Yang, K., Zhang, T., Kuang, Z., Xie, Q., Huang, J., Ananiadou, S.: Mentallama: Interpretable Mental Health Analysis on Social Media with Large Language Models. In: Proceedings of the ACM Web Conference 2024, pp. 4489–4500 (2024)

[16] O'Leary, K., Bhattacharya, A., Munson, S.A., Wobbrock, J.O., Pratt, W.: Design Opportunities for Mental Health Peer Support Technologies. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. CSCW '17, pp. 1470–1484. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/2998181.2998349 . https://doi.org/10.1145/2998181.2998349