Utilizing Large Language Models to Identify Evidence of Suicidality Risk through Analysis of Emotionally Charged Posts

Ahmet Yavuz Uluslu[*]
University of Zurich
ahmetyavuz.uluslu@uzh.ch

Andrianos Michail[*]
University of Zurich
andrianos.michail@cl.uzh.ch

Simon Clematide
University of Zurich
simon.clematide@cl.uzh.ch

Abstract

This paper presents our contribution to the CLPsych 2024 shared task, focusing on the use of open-source large language models (LLMs) for suicide risk assessment through the analysis of social media posts. We achieved first place (out of 15 participating teams) in the task of providing summarized evidence of a user's suicide risk. Our approach is based on Retrieval Augmented Generation (RAG), where we retrieve the top-k (k=5) posts with the highest emotional charge and provide the level of three different negative emotions (sadness, fear, anger) for each post during the generation phase.

1 Introduction

While healthcare systems are crucial in identifying suicide risk, the limited time available to clinicians often hinders a comprehensive assessment of all risk factors (Knipe et al., 2022). Expressions of suicidal thoughts are among the most significant warning signs. However, the standard practice of clinicians inquiring about these thoughts has not been reliably effective in predicting and preventing suicide (Hawton et al., 2022). It was revealed that the majority of patients who commit suicide had not reported suicidal thoughts to their healthcare providers (Chan et al., 2016).

The CLPsych 2024 shared task (Chim et al., 2024) addresses the significant challenge of generating supporting evidence for clinical assessments, with a specific focus on suicide risk assessment using open-source large language models (LLMs). This task concentrates on analyzing linguistic content from social media posts to substantiate the assigned suicide risk levels of individuals (Shing et al., 2018a). By examining users' posting activities on online forums, the goal is to extract, in an unsupervised manner, evidence within these posts that supports the pre-assigned risk levels.

Our approach aims to develop a scalable and efficient system that utilizes the state-of-the-art open-source LLM Mistral 7B (Jiang et al., 2023) for mental health assessment. It uses 4-bit quantization and Retrieval Augmented Generation (RAG) (Lewis et al., 2020) to effectively select the most emotional and relevant extracts from the user history with minimal resource requirements. We use emotional insights, which have been shown to correlate with mental illnesses such as depression, to improve our task by recognizing emotional patterns that could indicate suicidality (Zhang et al., 2023). We engage with both Task A (Highlighting Suicidal Evidence) and Task B (User's Summarized Suicidal Evidence). The two main contributions of this work, which were instrumental in achieving the top performance in Task B, are as follows:

- We retrieve the top-k ($k = 5$) emotionally charged user posts to include as context to the model to summarize evidence of suicidal risks.
- We enriched the prompt context with the regression-predicted percentage levels of three different negative emotions (sadness, fear, anger) alongside the selected posts.

2 Related Work

The CLPsych Shared Tasks 2019 (Zirikly et al., 2019) and 2022 (Tsakalidis et al., 2022) were mainly focused on suicide risk prediction and mood swing detection, which was predominantly considered a multi-class classification problem. The top approaches in the previous shared task have predominantly utilized transformer-based models and multitask learning, yet the capabilities of prompting-based approaches in this context remains largely unexplored. There is a growing interest in the responsible use of LLMs in healthcare, including in psychotherapy and mental health as-

[*]Equal contribution

Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pages 264–269
March 21, 2024 ©2024 Association for Computational Linguistics

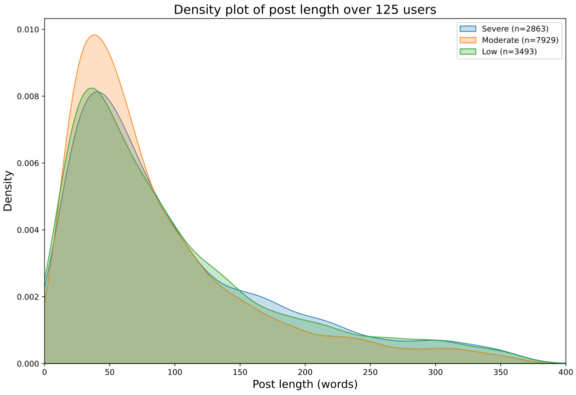

Figure 1: Kernel density estimation plot, grouped by post's author assigned risk level. n denotes the number of posts among all subreddits written by users of the assigned risk level. On average, the 125 users contributed to 131 posts each.

sessment (Stade et al., 2023). The recent advancements have seen a significant increase in the zero-shot classification abilities of LLMs, alongside a deepened understanding of mental health issues (Xu et al., 2023). These models are increasingly recognized for their effectiveness in extracting information, especially in identifying mental health crises. They have demonstrated the ability to generate explainable findings and exhibit reasoning capabilities, which significantly enhances their utility in mental health assessments (Yang et al., 2023). This evolution in the capabilities of LLMs sets a new precedent for our approach and underscores the potential of these models in contributing to mental health assessments.

3 Data and Tasks

We use the Reddit suicidality dataset provided by the organizers of the 2019/2024 CLPsych Workshop (Shing et al., 2018b; Zirikly et al., 2019; Chim et al., 2024). Our team's utilization of this data and our participation in the associated tasks adhere to the ethical review standards outlined by the organizers. The dataset comprises posts from 125 Reddit users on various subreddits where each user has at least one post in r/SuicideWatch. All users are categorized by experts into four risk levels: *No Risk*, *Low Risk*, *Moderate Risk*, and *Severe Risk*. The distribution of the length of posts is shown in Figure 1.

The participants of the shared task were asked to contribute to the following two methods of extracting evidence of suicidality from the users' posts:

System: You are a suicide prevention therapist expert.

You are performing psychological analysis of suicidality risks of the author of the following Reddit post:
title: {postTitle}
content: {postContent}
It is estimated that the author has at least {riskLevel} risk of suicidality.
Extract ALL relevant evidence (snippets) of at least {riskLevel} suicidality exclusively from the post's content.
Formatting: Strictly provide each snippet and only the snippets delimited by a semicolon(';')

Figure 2: Task A: Zero-shot prompt template (Step 1) given to Mistral7B for the extraction of relevant spans.

- **Task A – Highlighting Suicidal Evidence:** focuses on extracting highlights (snippets) exclusively from r/SuicideWatch posts that have been assigned a risk level by an expert.
- **Task B – User's Summarized Suicidal Evidence:** Using any content available from a person, the task is to find evidence for a person's suicidality risk level and report it, either extractive or abstractive.

4 Methods

4.1 Model & Computational Resources

Our approach exclusively employs the open-source LLM Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) in zero-shot setting with 4-bit quantization. Detailed parameters used for the text generation can be found in Table A1. All of our experiments, including inference over all posts and users, were carried out locally for a total of less than two hours using a MacBook Pro with an M2 Pro with a 10-core GPU.

4.2 Task A – Highlighting Suicidal Evidence

To extract and highlight relevant snippets from a user's post, we deploy a multistep procedure:

1. Prompting the LLM to extract relevant passages from the text (see Figure 2).
2. Prompting to remove unwanted text output (e.g., explanations) from Step 1 and reorganize snippets (see Figure A1).
3. Segmenting the snippets and applying up to 4-character replacement string substitutions

Team	Rank	Recall	Precision	Weighted Recall	Harmonic Mean
sophiaADS	1	0.944	0.906	0.489	0.924
UoS NLP	2	0.943	0.916	0.527	0.929
UniBuc Arch	3	0.939	0.890	0.390	0.914
SWELL	7	0.915	0.892	0.542	0.903
Our Official Submission	8	0.910	0.916	0.742	0.913
MHNLP	9	0.910	0.888	0.197	0.909

Table 1: Shared Task Results for Task A – Highlighting Suicidal Evidence. Our team name is UZH_CLyp and Rank denotes the subtask's ranking that is based on the primary score, Recall.

to restore the exact text form of the original post.[1] Unmatched evidence, often arising from rewritten or reordered user texts as well as from hallucinated insertions, is then discarded.

This method of using LLMs for an extractive task, while laborious, was explored to determine the feasibility of accomplishing this task solely through the use of generative AI. However, in doing so, we required two separate rounds of inference and an additional string matching, compromising the efficiency of our solution.

The primary evaluation metric for this task is a variation of BERTScore (Zhang et al., 2019), focusing on recall, weighted recall and precision, benchmarked against snippets extracted by human experts. Our submission for Task A achieved 8th place out of the 15 best team submissions (42 submissions in total). Detailed results for Task A can be found in Table 1.

> *System: You are a suicide prevention therapist expert.*
>
> You are performing psychological analysis of suicidality risks in online forums. Here are the most emotional posts of the same author for analysis:
> {postTitle, content and estimations(in percentages) of sadness, fear and anger separated by new lines, for all five posts retrieved by the highest sadness estimation}
> Aspects of text to consider are the emotions, cognitions as well as behaviours and mentions of the author related to things like self-harm or suicide.
> It was confirmed that the author has a {riskLevel} risk of suicidality. Provide your hypothesis of {riskLevel} suicidality from the post contents and general online behaviour.

Figure 3: Task B: Zero-shot prompt template given to Mistral7B to generate the summarized evidence.

4.3 Task B – User's Summarized Suicidal Evidence

Emotion Regression Models In addition to using our generative Suicidal Evidence predictions, we apply Encoder Transformer models to regress the emotional load of a text. We fine-tuned the Muppet RoBERTa (Aghajanyan et al., 2021) Large Encoder models on the SemEval2018 Affect dataset (Mohammad et al., 2018) to function as Emotion Regressors for emotions like anger, sadness, and fear, each with its separately trained model. These models are fine-tuned using the Head First Fine-Tuning method as described in Michail et al. (2023). The Pearson r correlation coefficient on its test set is 0.856, 0.832 and 0.808 for anger, fear and sadness respectively. Before prompting Mistral7B, we compute predictions with our emotion regression models for all posts of the studied users.

Prompting To generate the summarized evidence, we perform a zero-shot query by concatenating the title, the post and the predicted emotions to the five most sad posts (in descending order), similar to a Retrieval Augmented Generation (RAG) approach (Lewis et al., 2020). In addition to the post information, we provide the model with the following system message "You are a suicide prevention therapist expert", and some information and hints about aspects to consider when performing the task. Figure 3 presents the complete prompt template.

5 Results

The official results of Task B are shown in Table 2. Our submission (UZH_CLyp) achieved first place out of the 42 submitted runs according to the official scores. The official score of the Shared Task attempts to measure agreement between the model-generated summary and the human expert analysis (Chim et al., 2024) using a model trained on Natu-

[1] This was necessary for the submission as the LLM would fix small grammar errors, typing nuances or irregular punctuation usage.

Team	Rank	Mean Consistency ↑	Max Contradiction ↓
Our Official Submission	1	**0.979**	0.064
Our Ablation: No Emotion Regression	-	0.976	0.074
SBC	2	0.976	0.079
SWELL	3	0.973	0.081
UniBuc-Arch	4	0.973	0.081
SKKU-DSAIL	5	0.970	0.096
Our Ablation: No Emotion RAG	-	0.947	0.120
sophiaADS	12	0.944	0.175

Table 2: Shared Task Results for Task B – User's Summarized Suicidal Evidence. Italics denote our additional evaluations for the ablation study. Our team name is UZH_CLyp and Rank denotes the Subtask's ranking that is based on the primary score, Mean Consistency.

ral Language Inference (Wang et al., 2021). This agreement is measured with two scores: the **Mean Consistency**, defined as the average sentence-level probability of consistency, 1 - P(Contradiction), and the **Max Contradiction**, which represents the average maximum probability of a contradiction occurring, $\max(P(\text{Contradiction}))$. Our submission performed best within Task B among all submissions.

It is worth noting that the top teams achieve promising performance, with only minor differences between them. However, another interesting insight from this generally high performance is that it demonstrates the ability of today's LLMs to generate analyses that align closely with the assessments of human experts.

5.1 Ablation Study

To better understand the relevant factors for the performance of our system, we asked the organizers to evaluate two additional post-submission runs for our ablation study. We have included the results of these variants in the main results table 2.

In the *No Emotion Regression* ablation experiment, we omitted information about the emotionality levels of each post. This leads to a very minor performance decrease, showing that the actual predictions of the emotions are not crucial to the model.

In the *No Emotion RAG* ablation experiment, we omitted information about the emotions and also replaced the retrieval procedure that selected the most emotional posts with a heuristic to retrieve the five longest posts. This results in a large performance decrease and showcases the value of the emotion regressions as a selection criterion for retrieving relevant posts.

6 Conclusion

The system presented in this paper demonstrates the potential of using open-source LLMs to identify evidence of suicidality utilizing emotionally charged social media posts. Our main innovation is the combination of RAG with emotional regression of posts. This technique was found to be effective, as evidenced by the first-place performance in Shared Task B and the insights from our ablation experiments. Our results highlight the ability of current LLMs to accurately summarize evidence of a user's suicide risk from online posts that closely align with human expert assessments.

In conclusion, this study highlights the potential of LLMs in healthcare, particularly for mental health assessments. While the approach shows promise, especially in suicide risk analysis from social media posts, it also poses challenges, such as the risk of inaccurate content generation. Future research should aim to enhance the accuracy of these models and consider the ethical implications of applying AI in sensitive health contexts. This research opens up new possibilities for the application of LLMs in mental health services, suggesting a path towards integrating them with traditional healthcare methods for more effective outcomes.

Limitations

While our approach leveraging open-source LLMs shows promising results in both Task A and Task B of the CLPsych 2024 Shared Task, it is important to recognize inherent limitations when using LLMs in sensitive contexts of mental health assessment. We acknowledge the possibility of hallucinations and generation of inaccurate content, which can lead to misinterpretation of a user's mental state. During a manual inspection, we inspected the hallucination factor in scenarios where the model encounters posts with low/medium pre-assigned risk levels. In these cases, it often fails to pick up relevant clinical cues such as an intent to self-harm, thus underlining the crucial role of pre-assigned risk levels in guiding the model's explanation. Furthermore, the shared task is assessed using automated metrics, which may lead to significant discrepancies between these results and the evaluations of human expert annotations.

Ethics

We used publicly available data that was stripped of identifiable information and collected in a non-intrusive manner for mental health research. All researchers working on the project have signed a non-disclosure agreement with the dataset providers. The data was stored securely at the storage services of the Department for Computational Linguistics at the University of Zurich and was only accessible to the parties involved during the project. The open-source language models used in the project were hosted locally without any potential data disclosure to third parties. The results of this work are intended for fellow researchers in the fields of computational linguistics and psychology to improve mental health assessment technology. It is part of the growing body of mental health research aimed at applications to improve well-being. However, it should not be used without collaboration with clinical practitioners.

Acknowledgements

Ahmet Yavuz Uluslu was partially supported by the Innosuisse innovation project AUCH (103.188 IP-ICT). We are thankful to the organisers of the CLPsych 2024 Shared Task and the American Association of Suicidology for preparing the dataset.

References

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Melissa KY Chan, Henna Bhatti, Nick Meader, Sarah Stockton, Jonathan Evans, Rory C O'Connor, Nav Kapur, and Tim Kendall. 2016. Predicting suicide following self-harm: systematic review of risk factors and risk scales. *The British Journal of Psychiatry*, 209(4):277–283.

Jenny Chim, Adam Tsakalidis, Dimitris Gkoumas, Dana Atzil-Slonim, Yaakov Ophir, Ayah Zirikly, Philip Resnik, and Maria Liakata. 2024. Overview of the clpsych 2024 shared task: Leveraging large language models to identify evidence of suicidality risk in online posts. In *Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Keith Hawton, Karen Lascelles, Alexandra Pitman, Steve Gilbert, and Morton Silverman. 2022. Assessment of suicide risk in mental health practice: shifting from prediction to therapeutic assessment, formulation, and risk management. *The Lancet Psychiatry*, 9(11):922–928.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Duleeka Knipe, Prianka Padmanathan, Giles Newton-Howes, Lai Fong Chan, and Nav Kapur. 2022. Suicide and self-harm. *The Lancet*, 399(10338):1903–1916.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Andrianos Michail, Stefanos Konstantinou, and Simon Clematide. 2023. UZH_CLyp at SemEval-2023 task 9: Head-first fine-tuning and ChatGPT data generation for cross-lingual learning in tweet intimacy prediction. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1021–1029, Toronto, Canada. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018a. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, pages 25–36.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018b. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages

25–36, New Orleans, LA. Association for Computational Linguistics.

Elizabeth C Stade, Shannon W Stirman, Lyle H Ungar, Cody L Boland, H. A Schwartz, David B Yaden, João Sedoc, Robert DeRubeis, Robb Willer, and johannes C Eichstaedt. 2023. Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, et al. 2022. Overview of the clpsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 184–198.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. *CoRR*, abs/2104.14690.

Xuhai Xu, Bingshen Yao, Yuanzhe Dong, Hong Yu, James Hendler, Anind K Dey, and Dakuo Wang. 2023. Leveraging large language models for mental health prediction via online text data. *arXiv preprint arXiv:2307.14385*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore. Association for Computational Linguistics.

Tianlin Zhang, Kailai Yang, Shaoxiong Ji, and Sophia Ananiadou. 2023. Emotion fusion for mental illness detection from social media: A survey. *Information Fusion*, 92:231–246.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 24–33.

A Appendix – Supplementary Material

> *System: You are a helpful assistant.*
>
> We have the original text and a set of extracted snippets mixed in text. We want to extract ONLY all snippets (not numbered) without any further discussion or comments
> Original: {postContent}
> Mixed Extracted Snippets: {Step 1 Output}
> Follow the following format for all snippets (each on a new line): \nsnippet text as presented in the original

Figure A1: Task A: Zero-shot prompt template (Step 2) given to Mistral7B to extract the relevant spans.

Model Parameter	**Value**
Temperature	0.8
Top-P	0.8
Top-K	40
Max Tokens	512
Context Size	4096

Table A1: The main parameters used for Mistral7B (mistral7binstructv0.2.Q4_K_M)