Using Large Language Models (LLMs) to Extract Evidence from Pre-Annotated Social Media Data

Falwah AlHamed[1,3], **Julia Ive**[2], and **Lucia Specia**[1]

[1]Department of Computing, Imperial College London, London, UK
[1]{f.alhamed20,l.specia}@imperial.ac.uk
[2]Queen Mary University of London, London, UK
[2]j.ive@qmul.ac.uk
[3]King Abdulaziz City for Science and Technology(KACST), Riyadh, Saudi Arabia

Abstract

For numerous years, researchers have employed social media data to gain insights into users' mental health. Nevertheless, the majority of investigations concentrate on categorizing users into those experiencing depression and those considered healthy, or on detection of suicidal thoughts. In this paper, we aim to extract evidence of a pre-assigned gold label. We used a suicidality dataset containing Reddit posts labeled with the suicide risk level. The task is to use Large Language Models (LLMs) to extract evidence from the post that justifies the given label. We used Meta Llama 7b and lexicons for solving the task and we achieved a precision of 0.96.

1 Introduction

In today's world, many people use social media platforms. These platforms allow individuals to express themselves openly, sharing daily details about their activities and thoughts. Researchers have been studying social media data for years to understand users' mental health.

Natural Language Processing (NLP) is often applied to social media data in research that focuses on classifying the presence or absence of depression (Boinepelli et al., 2022; Chancellor and De Choudhury, 2020). Researchers also examine how to detect the transition from depression to suicidal ideation (De Choudhury et al., 2016; Gong et al., 2019; Matero et al., 2019; Sawhney et al., 2020).

In this paper, we explain our approach to the CLPsych 2024 shared task. The goal of this shared task is to utilize Large Language Models (LLMs) to detect textual cues that support the designated Suicide Risk Level, which may be classified as Low, Moderate, or High. This "evidence" could be provided in two ways, either highlighting (or "extracting") relevant spans within the text or by providing a summary, aggregating evidence that supports the assigned suicide risk level.

2 Dataset

Data used for this shared task was from (Zirikly et al., 2019; Shing et al., 2018). It was pulled from Reddit. This well-known social media platform contains communities known as "subreddits", each of which covers a different topic. Access has been granted to the UMD Suicidality Dataset v2, encompassing multiple Reddit users and their corresponding posts on the platform, along with the associated Suicide Risk Level labels. The dataset incorporates annotations for Suicide Risk Levels across subsets of posts within the r/SuicideWatch subreddit, categorized as follows:

- No Risk (or "None"): Absence of evidence indicating the person (post author) is at risk of suicide;
- Low Risk: Some factors may suggest a level of risk, but the likelihood of suicide is deemed low;
- Moderate Risk: Indications exist that the person could genuinely be at risk of attempting suicide;
- High ("Severe") Risk: The belief that the person is at a high risk of attempting suicide in the near future.

This shared task exclusively concentrates on the assessment of Low, Moderate, and High risk levels. It is essential to note that, although the term "suicidal crisis" was not employed in the original risk labeling by (Shing et al., 2018), the High category closely aligns with this concept, denoting an acute situation necessitating immediate intervention. All authors have signed the Data User Agreement (as requested by the organisers) to have access to the dataset.

Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pages 232–237

3 Methods

In this section, we will describe the methods we developed to address the shared tasks. The main instruction for this task was to use Large Language Models (LLMs) to extract the evidence. LLMs have demonstrated superior performance in understanding human language and generating text resembling it.

3.1 Task Description

The task (Chim et al., 2024) is to detect textual cues that support the annotated Suicide Risk Level to Reddit users who wrote posts on the r\suicideWatch subreddit. The "evidence" critical to our analysis can be presented through two approaches. The first method involves providing a comprehensive summary. This entails aggregating and synthesizing the identified evidence into a cohesive overview that captures essential information throughout the text. The second method centers on highlighting or extracting specific, relevant spans within the text, focusing on essential details that contribute to the assigned suicide risk level. This method includes extracting key textual segments indicative of the individual's suicide risk level. The granularity of requirements of the tasks is as follows: the risk is annotated at user-level, the summary evidence is required at user-level, and the highlights evidence is required at post-level. Some rules were identified for accomplishing this task, including that the summary does not exceed 300 words, and that highlights are extracted as exact quotes from the posts. In addition, it is not allowed to use APIs as transmitting the data to other servers raises a concern of data leaks. Thus, we are not allowed to use OpenAI GPT models [1] or Google Bard model [2].

3.2 Model

We used the open-source Meta Llama 2 7B chat LLM (Touvron et al., 2023). Llama is built on a transformer architecture and underwent pretraining on openly accessible online data sources. Subsequently, the fine-tuned model, Llama Chat, utilizes publicly available instructional datasets, incorporating input from over 1 million human annotations. The Hugging face library (Wolf et al., 2019) is used, namely the `'Llama-2-7b-chat-hf'` model card. We experiment using a zero-shot learning approach with different prompts.

3.3 Evidence 1: Summary

Prompts are questions or statements that are provided to the model to initiate and guide a conversation or specific task or to generate desired text.

We experimented with Llama 2 7b to find the summarized text evidence in two rounds with different prompts. In the first round, we are seeking an explanation of the state of the user who wrote the post. A set of the prompts used in extracting evidence 1 (the summary) round 1 is illustrated in Table 1. After receiving the response, we then prompt the model again aiming at summarizing the paragraphs received from the first round. The prompt used for the second round is: `Rewrite this text as a descriptive paragraph of the person who wrote it in less than 300 words starting with This person is at [suicide level] risk because Text:...`

Table 1: A set of the prompts used finding evidence 1 (the summary)

Explain the suicide risk level of the person who wrote this text
Explain why the user who wrote this text has [suicide level] suicide risk level.
Explain why the user who wrote this text has depressive episodes.
Why do you think who wrote this text has [suicide level] suicide risk level?
A psychologist identifies the person who wrote the following text as having a [suicide level] risk of suicide, can you explain why?
Write a paragraph on why this text might contain [suicide level] suicide risk.
Can you let me know in a paragraph why this text is considered low mood?

3.4 Evidence 2: Highlights

Llama Prompts. We experimented with Llama 2 7b to extract the highlights evidence from the posts using different prompts, a set of the prompts used in extracting highlights evidence is illustrated in Table 2.

Lexical Extraction. Previous studies indicate that enhancing prediction outcomes can be achieved by incorporating lexical features in conjunction with machine learning models (AlHamed and AlGwaiz, 2020; Carvalho and Plastino, 2021). Thus, we inspected the posts of the three classes

[1] https://platform.openai.com/docs/libraries/python-library

[2] https://bard.google.com/chat

Table 2: A set of the prompts used finding evidence 2 (the highlights)

Can you identify pieces of text that indicate low mood in the following text and answer with a list of texts?
A psychologist identifies the person who wrote the following text as having a [suicide level] risk of suicide, can you identify pieces of text that indicate that
Identify all quotes of low mood in this text
Identify all quotes about suicide risk in this text
Can you identify all text spans of depressive symptoms in this text?

Table 3: List of suicidal words for Task B

Suicidal Words	
kill	die
knife	survive
dead	end my life
I'm gone	live anymore
I'm done	taking my life
killing	overdose
jump	suicide
wrist	hang
burn	self-harm
self harm	pesticide
death	take my life
call for help	
Depressive Words	
depression	depressive
depressed	sad
mood	cry

and found that they contain many words related to suicide attempts and depressive thoughts. Thus, in addition to the highlights extracted by Llama, we used the list of suicidal words created by (Alhamed et al., 2022) and we added other words of depression inferred from manual posts' inspection. The word list is shown in Table 3. For each word from the text found in a post, we retrieved the sentence as two words before and 2 words after the word found in the lexicon (5 words window size). This sentence was added to the highlights list.

4 Evaluation

As per task organizers (Chim et al., 2024), submissions are evaluated against a test set annotated by two domain experts. Each test set example comprises (i) the risk level label of an individual, (ii) a list of posts written by the individual, (iii) text spans highlighted by annotators from the posts with evidence that support the risk level label, and (iv) a human-written summary that aggregates the highlighted evidence and observations into a single piece of text. Evaluation metrics are as follows:

Summarized Evidence

- **Consistency** is the lack of contradiction. At a user-level, score each sentence in the submitted evidence summary by running a natural language inference (NLI) model on it and every gold summary sentence, using it as hypothesis and the gold sentence as a premise, to obtain the probability of it contradicting the gold sentence. The sentence-level consistency score is thus `1 - (the probability of the "contradiction" prediction)`. Then, take the average consistency score across all sentences for the user. Overall submission-level score is the mean consistency score across all users.

- **Contradiction** Penalizes information that contradicts the gold evidence summary. Lower scores are better. Note that some contradictions are expected, since the same text can describe both risk and protective factors. At a user-level, the organaizer score each sentence in the submitted evidence summary by running an NLI model on it (hypothesis) and every gold summary sentence (premise), taking the maximum contradiction probability. Then, average across all submitted sentences. Overall submission-level score is the mean contradiction score across all users.

Highlights

- **Recall** Measures how much relevant supporting evidence information was predicted. For a given user, for every gold highlight, find the candidate highlight with the highest semantic similarity (based on BERTScore (Zhang et al., 2019)). Take the average similarity across all gold highlights. Overall submission-level score is the mean recall across all test users.

- **Precision** Measures the quality of predicted supporting evidence. For a given user, for every candidate highlight, find the gold highlight with the highest semantic similarity (based on BERTScore). Take the average similarity across all candidate highlights. Overall submission-level score is the mean precision across all test users.

- **Weighted Recall** A length-sensitive version of recall. Measures how much relevant supporting evidence information was predicted and whether the evidence lengths are similar to human-highlighted ones. At a user-level, sum the length (token count) of gold highlights and of submitted highlights. If the number of submitted highlight tokens exceeds the number of gold highlight tokens, weigh the user-level recall score by the ratio of gold:candidate tokens. Overall submission-level score is the mean weighted recall across all test users.

- **Harmonic Mean** Balances between precision and recall when evaluating how well the submission identified supporting evidence. The user-level harmonic mean between unweighted recall and precision is mean-averaged across all test users.

5 Results

We applied prompts to Llama, and it responded to the majority of them with an explanation for the given post. Although Large Language Models (LLMs) are known for their robust language processing capabilities, they have encountered difficulties in addressing specific aspects of mental health. In some cases, Llama refuses to answer and responds with "`It is important to note that this is just one text message, and it is not possible to make a definitive assessment of the person's suicide risk level based on this one message`" or "`It is important to note that these are just a few potential indicators of suicide risk, and that each person's situation is unique. However, if you are concerned about someone's safety, it is important to take their concerns seriously and offer support and resources..`" The prompt that provides us with the best matching summary evidence for all of the posts was "`Can you let me know in a paragraph why this text is considered low mood?`" as it scored 0.964 consistency with the gold standard, where the prompts "`Explain why the user who wrote this text has [suicide level] suicide risk level`" and "`Write a paragraph on why this text might contain [suicide level] suicide risk`" scored 0.873 and 0.878, respectively.

The prompt that provides the best matching highlights evidence was "`Can you identify pieces of text that indicate low mood in the following text and answer with a list of texts?`"

Table 4: Results

Summarized Evidence	Mean Consistency	0.964
	Max Contradiction	0.060
Highlights	Precision	0.899
	Harmonic Mean	0.888

In this shared task, it was imperative to extract direct quotes, or highlights, from the text of posts. Our model adeptly performed this task while also rectifying any spelling mistakes within these quotes. However, during the subsequent validity check phase before submission, this spell-checking process inadvertently led to errors, resulting in some posts being submitted with empty quotes due to the approaching deadline. Consequently, this issue contributed to a lower overall recall (0.577). Nevertheless, it's worth noting that our model maintained a high level of recall (0.887) for all posts that did not contain empty quotes.

The incorporation of a lexicon in our work enhances the results by expanding the list of highlights, identified during the extraction process. This integration contributes to an increased recall rate as the lexicon serves as a valuable reference, allowing the model to recognize and include additional relevant quotes that align with predefined criteria. By leveraging the lexicon, our approach not only captures a broader spectrum of highlights but also augments the comprehensiveness of the extracted information in the summary, thereby improving the overall performance of the system.

The results obtained using our proposed method are illustrated in Table 4

6 Conclusions and Future Work

In conclusion, researchers have extensively utilized social media data over the years to gain valuable insights into users' mental health. However, the predominant focus of many investigations has been on categorizing users into those with depression and those deemed healthy, or on detecting suicidal ideation. In this study, our objective was to extract evidence corresponding to pre-assigned gold labels. To achieve this, we utilized a suicidality dataset comprising Reddit posts labeled with suicide risk levels. Our task involved employing Large Lan-

guage Models (LLMs) to extract evidence from the posts justifying the assigned labels. Through the utilization of Meta Llama 7b and lexicons, we attained commendable results, achieving a precision rate of 0.96. These findings underscore the efficacy of utilizing advanced language models and lexicon-based approaches in extracting evidence pertinent to the assigned suicide risk levels from social media posts. In the future, we aim to try other LLM models such as OpenAI GPT and Google Bard. We also aim to expand the suicidal words list and extract additional features from the text that could enhance the obtained results.

7 Limitations

Llama2 7b is limited to handling a maximum of 4096 tokens, which resulted in trimmed posts, leading to potential information loss and truncation of longer posts that possibly caused incomplete understanding and biased sampling. In addition, the llama2 model used in this paper is trained using extensive datasets collected from the internet. Although it can produce human-like text, it is worth noting that training it on domain-specific data (mental health data) could improve its performance.

8 Ethical Consideration

The collected dataset contains only publicly available posts from Reddit, and we have signed a data user agreement not to share or distribute any data outside the team. We are also committed to following ethical practices to protect users' privacy and anonymity. This includes not using commercial LLMs to protect user privacy, not submitting all or part of the data to any platform that may use the data for training, and data is only stored on password protected servers and computers.

Acknowledgements

The authors are particularly grateful to the anonymous users of Reddit whose data feature in this year's shared task dataset, to the clinical experts from BarIlan University who annotated the data. We extend our acknowledgements to the American Association of Suicidology for making the dataset available.

References

Falwah AlHamed and Aljohara AlGwaiz. 2020. A Hybrid Social Mining Approach for Companies Current Reputation Analysis. In *Recent Advances on Soft Computing and Data Mining*, pages 429–438, Cham. Springer International Publishing.

Falwah Alhamed, Julia Ive, and Lucia Specia. 2022. Predicting moments of mood changes overtime from imbalanced social media data. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 239–244, Seattle, USA. Association for Computational Linguistics.

Sravani Boinepelli, Tathagata Raha, Harika Abburi, Pulkit Parikh, Niyati Chhaya, and Vasudeva Varma. 2022. Leveraging mental health forums for user-level depression detection on social media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5418–5427.

Jonnathan Carvalho and Alexandre Plastino. 2021. On the evaluation and combination of state-of-the-art features in Twitter sentiment analysis. *Artificial Intelligence Review*, 54(3):1887–1936.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine*, 3(1):43.

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.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. *Conference on Human Factors in Computing Systems - Proceedings*, pages 2098–2110.

Jue Gong, Gregory E. Simon, and Shan Liu. 2019. Machine learning discovery of longitudinal patterns of depression and suicidal ideation. *PLoS ONE*, 14(9):1–15.

Matthew Matero, Akash Idnani, Youngseo Son, Sal Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H. Andrew Schwartz. 2019. Suicide Risk Assessment with Multi-level Dual-Context Language and. pages 39–44.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2020. A Time-Aware Transformer Based Model for Suicide Ideation Detection on Social Media. pages 7685–7697.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. 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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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, Özlem 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, Minneapolis, Minnesota. Association for Computational Linguistics.