Incorporating Word Count Information into Depression Risk Summary Generation: INF@UoS CLPsych 2024 Submission

Judita Preiss and **Zenan Chen**
University of Sheffield, Information School
The Wave, 2 Whitham Rd, Sheffield S10 2SJ, United Kingdom
judita.preiss@sheffield.ac.uk and zchen249@sheffield.ac.uk

Abstract

Large language model classifiers do not directly offer transparency: it is not clear why one class is chosen over another. In this work, summaries explaining the suicide risk level assigned using a fine-tuned `mental-roberta-base` model are generated from key phrases extracted using SHAP explainability using Mistral-7B. The training data for the classifier consists of all Reddit posts of a user in the University of Maryland Reddit Suicidality Dataset, Version 2, with their suicide risk labels along with selected features extracted from each post by the Linguistic Inquiry and Word Count (LIWC-22) tool. The resulting model is used to make predictions regarding risk on each post of the users in the evaluation set of the CLPsych 2024 shared task, with a SHAP explainer used to identify the phrases contributing to the top scoring, correct and severe risk categories. Some basic stoplisting is applied to the extracted phrases, along with length based filtering, and a locally run version of Mistral-7B-Instruct-v0.1 is used to create summaries from the highest value (based on SHAP) phrases.

1 Introduction

With the ability to use large language models (LLMs) to classify people's suicide risk level comes the need for transparency: artificial intelligence (AI) has been known to learn incorrect patterns and make incorrect generalizations (Narla et al., 2018). To this end, especially in a sensitive domain such as mental health, insight into the reasons for the prediction made is required to allow an expert to look through the output and correct it as needed. In this work, we employ SHAP values to extract phrases contributing to the LLM's decision regarding suicide level risk which we further summarize using locally run generative AI to offer an explanation for the suicide risk level assigned.

The CLPsych 2024 shared task (Chim et al., 2024) explores the use of LLMs in order to find evidence within text supporting an assigned suicide risk level. The University of Maryland Reddit Suicidality Dataset version 2 dataset, which was made available to participants, contains user-linked posts from Reddit annotated for level of suicide risk labelled on a four point scale (no risk, low, moderate, and severe risk) as described in (Shing et al., 2018) and (Zirikly et al., 2019). The evidence supporting the risk level could be supplied in one of two ways:

1. By highlighting the relevant portions of posts.
2. By summarizing the evidence into a short explanation.

For the first task, we fine-tune a pre-trained Reddit based mental health model for suicide risk level classification and extract SHAP value based phrases which represent the highest contributors to the decision. For the second task, a subset of the phrases extracted from the first task is used as part of a prompt to a generative AI algorithm which is instructed to produce a summary focusing on the aspects highlighted in the task definition, namely: emotions, cognitions, behaviour and motivation, interpersonal and social support, mental health related issues and additional risk factors.

2 Related work

The approach is composed of two distinct phases: (1) fine-tuning of a suicide risk classifier, and (2) generation of a summary. The work also explores the integration of additional psycholinguistic based information and transparency via explainability.

2.1 Detection of mental health state

Online social media is increasingly used by users to share a variety of user-generated or user-curated information, including publishing of personal status updates and engaging in topic-specific channels (Wongkoblap et al., 2017). Language use

Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pages 211–217
March 21, 2024 ©2024 Association for Computational Linguistics

has been shown to change depending on a person's mental health state (Coppersmith et al., 2015), fuelling the creation of classifiers based on social media posts with Reddit forming a frequently used resource due to the presence of topic-specific channels, subreddits, such as r/SuicideWatch, r/depression.

Increased prediction performance has been observed when language models (LMs) used targetted texts in training, for example PsychBERT, a specialized BERT model trained on PubMed papers in the domain of psychology, psychiatry, mental health, and behavioral health and social media conversations about mental health (Vajre et al., 2021), or MentalBERT and MentalRoBERTa, which trained BERT and RoBERTa models respectively based on data from social forums for mental health discussion (Ji et al., 2022). The models were fine-tuned for classification of a number of mental health conditions and evaluated on standard datasets, and therefore lend themselves to fine-tuning for suicide risk level classification.

2.2 Generative AI system

Generative AI is frequently used in chatbots, where an AI system is generating its own, new, responses to hold a conversation with a human participant. The knowledge they hold stems from the wide variety of training data used to create such models; non-open-source models, such as GPT-4 (OpenAI et al., 2023) or PaLM (Chowdhery et al., 2022), do not share exact details of their training data or their architectures, however open source models, such as LLaMA-2 (Touvron et al., 2023) or Mistral (Jiang et al., 2023) can be deployed in local environments, enabling customisation with particular datasets while preserving data privacy. Their suitability for the mental health domain can be observed, for example, in the number of mental-health chatbot apps (Haque and Rubya, 2023).

2.3 Linguistic Inquiry Word Count

Linguistic Inquiry Word Count (LIWC) is a computing software used to extract features for mental health studies (Pennebaker and King, 1999). It has been used widely in research related to mental health condition identification. Chen et al. (2018) trained a log-linear classier, using LIWC as one of the feature sets to detect mental issues, while Sekulic et al. (2018) used LIWC features to predict bipolar disorder. In the social media domain, Coppersmith et al. (2015) extracted LIWC features from Twitter data to examine various mental health conditions.

2.4 Transparency of LMs

Surrogate models, such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), tweak a model's input slightly to explore the change in prediction. This enables them to highlight words / phrases which are particularly significant to a specific decision within a black box model such as a LM, with SHAP enabling straightforward extraction of important phrases from multi-class text based classifiers.

3 Method

The creation of short summaries describing the mental health state and depression risk of users is of a two step design: (1) a deep learning classifier is built for risk level prediction, which provides access to important phrases in each post, and (b) a subset of such phrases is summarized by a generative AI system.

3.1 Classification of suicide risk level

The provided data contains an expert assigned suicide risk level alongside numerous Reddit posts made by these users. The posts span a relatively short time frame, with the earliest posts in the data from 2015-09-01 and the latest 2016-01-29. While a person's mental health state may change over time, we make the assumption that over the time period covered by the data, their mental health state, and specifically their suicide risk level, has not changed. Therefore, each post made by the user is assigned the same risk level label. The data is balanced and a stratified 70 / 30 split is created to yield training and evaluation datasets: the data is only stratified by risk level, not user, as (a) individual posts of a user are not linked, and (b) the ultimate goal is not risk level prediction. Each post is converted to a single text, by concatenating the title and body as follows: *Title:* ... *Body:* ...

The classifier is built by fine-tuning `mental/mental-roberta-base`, a moderately-sized pre-trained language model which has been trained using mental health-related posts on top of RoBERTa-Base (cased_L-12_H-768_A-12) (Ji et al., 2022). Early stopping is applied, which allows (limited) exploration of hyperparameters, specifically the learning rate, as well as the (best portion of) input data.

LIWC	Description
affiliation	Desire for connection
allnone	Certainty
tone_neg	Negative tone
emotion	Emotion
emo_neg	Negative emotion
emo_sad	Sadness emotion
mental	Mental health behaviour
allure	Persuasiveness
feeling	Feeling

Table 1: Selected LIWC features with description

3.1.1 Inclusion of LIWC information

LIWC-22 (Boyd et al., 2022) is used to extract additional information from each post. The most informative of these features (see below) are integrated into the training phase of the classifier. LIWC uses word counting to determine the percentage of words indicative of specific psychological constructs or categories within a text. The words of interest (such as personal pronouns) are based on internal dictionaries, with LIWC-22's dictionary containing over over 12,000 words associated with the selected psychologically relevant categories, resulting in values for 119 different features output for each post.

Many of the features are relatively sparse for the current dataset, enabling feature reduction to be performed. Standard statistics of each feature were explored, as were correlations with risk categories. Statistical information was extracted from data constrained to specific risk categories: i.e. all posts of a user were assigned the user's risk category, and the mean value of each LIWC feature was computed. Features with a monotonically increasing mean across risk categories were included in the final selection shown in Table 1; the description was used to construct a phrase which was prepended to the post information. Since LMs do not interpret numbers well, values below a feature's mean were converted to *low* and above the mean were considered *high*. Thus a post with the title *"I feel sad"* and body *"It's that time of year"* with an associated LIWC score of 0.3 for the *emo_sad* feature (which has a mean of 0.12) becomes: *High emotional sadness. Title: I feel sad. Body: It's that time of year.* This augmented input is used to train a suicide risk classifier as described in Section 3.1.

3.1.2 Extraction of important phrases

SHAP values are a game theory based approach to gaining insights into the predictions made by machine learning by producing an explanation based on feature contributions towards the final decision (Lundberg and Lee, 2017). The approach is model agnostic and can be applied to all machine learning models including neural networks. For a multiclass problem, such as suicide risk classification, the partition explainer can be used to compute the SHAP values for each text. These values explain the impact of unmasking each word to the final prediction (see official SHAP example in Figure 1 from `https://shap.readthedocs.io`).

For a given user, phrases highlighted by SHAP as contributing to the highest suicide risk prediction were extracted from each post in the r/SuicideWatch subreddit.[1] Words between selected phrases were added if their contribution was low to other classes, increasing the quantity of continuous text. I.e. for the example shown in Figure 1, *feeling* and *hopeless* would be extracted initially and *so* would be added to produce the highlighted phrase *feeling so hopeless*. Any phrases consisting of at most a single content word alongside 0 or more (nltk) stoplist words are removed.

3.2 Generation of summary

Locally run generative AI was explored for the purpose of building a summary based on the important phrases extracted in Section 3.1.2. For each user, the phrases were ordered by decreasing length and the longest phrases were retained until a pre-specified length limit was reached. A number of prompts was explored with the `meta-llama/Llama-2-13b-chat-hf` model (Touvron et al., 2023), however, the model was found to be hard to (a) restrict to a specific maximum length, and (b) stop from deteriorating into a more social media style. The `mistralai/Mistral-7B-Instruct-v0.1` model (Jiang et al., 2023), which uses sliding-window attention, did not suffer from the same problems. The instruction given to the model was

> Summarize the (1) emotions, (2) cognitions, (3) social support, (4) mental health issues and (5) conceptual risk factors (one average length sentence for

[1] Note that posts were uniformly shortened for the explainer until post length matched the explainer's expectations – for the majority of posts, this corresponded to 512 tokens.

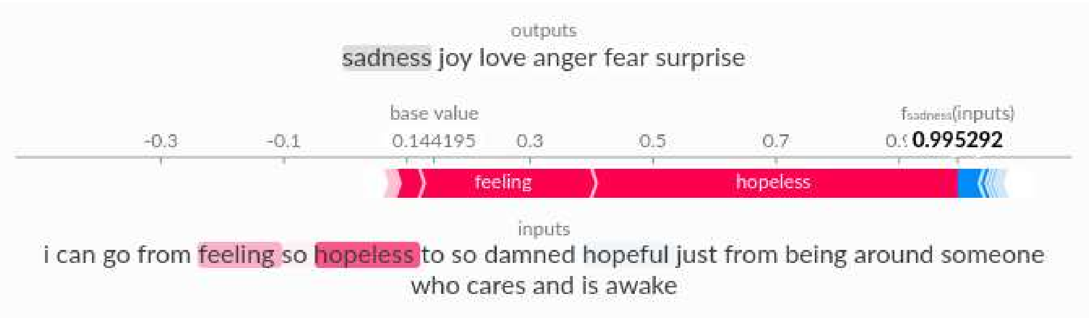

Figure 1: Example showing contributions of words towards final prediction of emotion (example from https://shap.readthedocs.io, not CLPsych dataset)

> each of the five factors) indicating depressed or suicidal thoughts in following phrases:

followed by the subset of phrases identified above. The prompt may be considered relatively complex, however prompts such as *Generate a 250 word summary based on the following excerpts explaining why the following phrases may indicate depressed or suicidal thought:* frequently failed to address one or more of the requested five aspects. Since no data was available for optimization, the model was used with its default parameter values.

	Recall	HM	Mean consistency
Value	0.850	0.896	0.934

Table 2: Results of the INF@UoS system

4 Results and discussion

Optimization, with early stopping, was performed over learning rate, the quantity of data used in training and inclusion or exclusion of LIWC information. The training data was balanced and the best performing model, at 51% (over a balanced evaluation dataset which included all 4 classes), was found to be using expert data only with LIWC information included with a learning rate of $2 - e6$.

The pipeline, starting from risk level classification, through extraction of important phrases and ending with summary generation, was run for all 125 users in the evaluation set. While important phrases were extracted from all posts, only the highlights that were used in summary generation were submitted, alongside summaries, resulting in some submissions having empty highlights for specific posts (but having a non empty summary, as this was generated from posts which were deemed more informative). Fourteen users were therefore submitted with empty highlights for at least one post (21 posts, out of 166 posts, in total): this affects the overall metrics for the system shown in Table 2. When computed only over the 111 users with a submitted set of complete highlights, the recall increases to 0.958.[2] Interestingly, mean consistency is identical (to 2 d.p.) for users where posts other than those in the test set were used to summarize evidence (i.e. users with empty highlights). To reiterate, empty highlights forming part of the submission do not mean that SHAP failed to extract important phrases from the appropriate post, only that other posts by the same author were selected for summary generation – extracted SHAP phrases were not submitted if they were not used for evidence generation.

After the competition, manual analysis was performed of the summarized evidence and the extracted highlights. Note that the official summaries and highlights were not released, so the results presented are only our judgements. The evaluation set contained 125 users: 39% of the submitted summaries were complete sentences summarizing the requested aspects of its inputs, 39% were also good summaries, but rather than sentences, they consisted of lists (such as "Emotions: hopelessness, loneliness"). 4% answered each point with an exact quote from the SHAP phrases and 8% were a mix of quotes from posts and generated text. Also relevant were 3% of summaries which in addition contained information which wasn't linked to the required points – such as a basic sentence containing only the person's age (e.g. "They are 30 years old."). The remaining summaries were either partial (2%), or probably too general, appearing to outline importance of the various aspects for suicide risk evaluation. Only one summary was nonsensi-

[2] Note that this is not comparable to the overall results for other teams.

cal and all summaries were within the permitted length, with a mean of 85 words.

Since the summarized evidence is generated from SHAP extracted phrases, 25% of these (136 highlights) were also manually explored. While 88% appeared OK (in this we include highlights which were not clearly supporting suicide risk judgement alone, but they complemented other selected highlights), a large portion contained fragments within the highlight: such as portions of a previous or following sentence, or ending at a point where is was clear how the fragment continued but with the end missing (e.g. "...one way or"). Some fragment highlights were also not entire sentences from the original post, but they were a self contained sentence. The remaining sources of error were either fragments that were too short to carry enough meaning (5%), fragments that - due to their selection - were inconsistent (1%), and highlights which didn't appear pertinent to the assessment of someone's risk of suicide (6%).

5 Conclusion and future work

We have shown the utility of SHAP explainability for the extraction of important phrases from text for the purpose of transparency within a text based suicide risk level classifier. Mistral 7B performs well for summary generation in this domain, retaining text integrity and producing minimal hallucinations.

Further investigations are required as to the contribution of the LIWC tool to the changes of SHAP extracted phrases, alongside a comparison with a one step (rather than the two step, SHAP + Mistral) summary generation process. Ablation tests evaluating changes to each step of the pipeline would also bring more insights.

In future work, the quality of highlights selected by SHAP could be improved by ensuring complete sentences surrounding the highlight are extracted.

6 Limitations

Some assumptions are made in this work, resulting in a number of limitations. We assume that the user's mental health state has not changed over the period of time the posts are from. While the period from which the posts were gathered was deemed short, this may not always hold. In addition, all posts of a user were included in training, including posts from subreddits other than r/SuicideWatch. It is unclear whether the variability in length as well as topic and emphasis may not be affecting the performance of the resulting classifier. Currently, the length of posts is limited to the max length of the model; recent models (such as MentalXLNet and MentalLongformer (Ji et al., 2023)), which allow longer contexts should be explored.

The integration of LIWC data is not optimized: large language models are not designed for interpreting numeric content, and the integration of an approach capable of understanding numeric values may result in better classifier results. The SHAP values produced highlight correlations between features (words) and the classification category. However, individual behaviours may be different, and a feature which is indicative of a low risk with other person may not be so with another.

Lastly, using generative AI in a sensitive domain is risky due to its ability to hallucinate.

7 Ethics

Secure access to the shared task dataset was provided with the task's approval under University of Maryland, College Park and approval by the University of Sheffield Information School Ethics Committee (ethical application reference 058377).

Acknowledgements

We would like to acknowledge the assistance of the American Association of Suicidology in making the dataset available.

References

Ryan L. Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W. Pennebaker. 2022. The development and psychometric properties of LIWC-22. Technical report, Austin, TX: University of Texas at Austin.

Xuetong Chen, Martin Sykora, Thomas Jackson, Suzanne Elayan, and Fehmidah Munir. 2018. Tweeting your mental health: An exploration of different classifiers and features with emotional signals in identifying mental health conditions. In *Proceedings of the 51st Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences.

Jenny Chim, Adam Tsakalidis, Dimitris Gkoumas, Dana Atzil-Slonim, Yaakov Ophir, Ayah Zirikly, Philip Resnik, and Maria Liakata. 2024. Overview of the clpsych 2024 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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling language modeling with pathways.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Stroudsburg, PA, USA. Association for Computational Linguistics.

M D Romael Haque and Sabirat Rubya. 2023. An overview of chatbot-based mobile mental health apps: Insights from app description and user reviews. *JMIR Mhealth Uhealth*, 11:e44838.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of LREC*.

Shaoxiong Ji, Tianlin Zhang, Kailai Yang, Sophia Ananiadou, Erik Cambria, and Jörg Tiedemann. 2023. Domain-specific continued pretraining of language models for capturing long context in mental health.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Akhila Narla, Brett Kuprel, Kavita Sarin, Roberto Novoa, and Justin Ko. 2018. Automated classification of skin lesions: From pixels to practice. *Journal of Investigative Dermatology*, 138(10):2108–2110.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C J Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. GPT-4 technical report.

J W Pennebaker and L A King. 1999. Linguistic styles: language use as an individual difference. *J. Pers. Soc. Psychol.*, 77(6):1296–1312.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Ivan Sekulic, Matej Gjurković, and Jan Šnajder. 2018. Not just depressed: Bipolar disorder prediction on reddit. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Stroudsburg, PA, USA. Association for Computational Linguistics.

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, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Vedant Vajre, Mitch Naylor, Uday Kamath, and Amarda Shehu. 2021. PsychBERT: A mental health language model for social media mental health behavioral analysis. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE.

Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin. 2017. Researching mental health disorders in the era of social media: Systematic review. *J. Med. Internet Res.*, 19(6):e228.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 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.