Representation Learning for Identifying Depression Causes in Social Media

Priyanshul Govil
International Institute of Information Technology
Hyderabad, Telangana, India
priyanshul.govil@research.iiit.ac.in

Vamshi Krishna Bonagiri
International Institute of Information Technology
Hyderabad, Telangana, India
vamshi.b@research.iiit.ac.in

Muskan Garg
Mayo Clinic
Rochester, Minnesota, USA
garg.muskan@mayo.edu

Ponnurangam Kumaraguru
International Institute of Information Technology
Hyderabad, Telangana, India
pk.guru@iiit.ac.in

ABSTRACT

Social media provides a supportive and anonymous environment for discussing mental health issues, including depression. Existing research on identifying the cause of depression focuses primarily on improving classifier models, while neglecting the importance of learning better data representations. To address this gap, we introduce an architecture that enhances the identification of the cause of depression by learning improved data representations. Our work enables a deeper interpretation of the cause of depression in social media contexts, emphasizing the significance of effective representation learning for this task. Our work can act as a foundation for self-help applications in the field of mental health.

CCS CONCEPTS

• **Applied computing** → **Life and medical sciences**; • **Human-centered computing**; • **Social and professional topics**; • **Computing methodologies** → **Natural language processing**;

KEYWORDS

mental health, depression, representation learning, cause of depression, natural language processing

ACM Reference Format:
Priyanshul Govil, Vamshi Krishna Bonagiri, Muskan Garg, and Ponnurangam Kumaraguru. 2023. Representation Learning for Identifying Depression Causes in Social Media. In *M. Gaur, E. Tsamoura, S. Sreedharan, S. Mittal. Proceedings of the Third ACM SIGKDD Workshop on Knowledge-infused Learning (KDD KiL 2023). Long Beach, California, USA, August 6, 2023.* ACM, New York, NY, USA, 6 pages.

1 INTRODUCTION

With the rapid spread of awareness around mental health, the field is ever-growing. In 2021, 22.8% of U.S. adults and 16.5% of U.S. youth aged 6-17 suffered from a mental health illness [4]. Depression is the second highest diagnosed mental health disorder [3]. However, there still exists a stigma around discussing mental health and depression. Children, teenagers, and sometimes even adults fear admitting their mental health issues as they fear society's judgment.

Social media allows users to participate in discussions anonymously, thereby allowing for confidentiality in the discussions. Therefore, many people turn to social media platforms like Reddit and Twitter to express their mental discomforts and discuss solutions to their problems. The communities on these platforms are supportive and helpful [9], giving people suffering a positive environment for betterment.

Identifying the cause of depression is the first step in treating depression. Posts on social media can indicate depression [7] and can be used to identify the cause of depression [14]. Therefore, studying social media is crucial in the context of mental health and depression. We utilize the CAMS dataset [14] due to its unique inference annotation (Table 1). In this work, we provide the existing classification models scope to improve by building on top of our work. Our contributions include the following:

(1) An architecture to accurately learn representations of social media depression posts for the downstream task of classifying it into its causal category of depression.
(2) An evaluation metric to assess a model's quality for embedding generation for social media depression posts.

2 RELATED WORK

There has been a growing focus on the problem of monitoring mental health on social media, especially within the field of natural language processing (NLP) [16]. In the past, identification of mental health disorders [17] in social media has been approached using methods such as social network analysis [22], emotion fusion [27], and other deep learning methods [21] [15] [24]. More recently, mental health disorder detection has predominantly been approached with the help of language models pre-trained on social media data such as MentalBERT and MentalRoBERTa [20]. Other problems, such as suicidal intent detection, have been approached using supervised learning [19] and deep learning methods [8] [26].

Detecting depression in social media using NLP has been a popular topic of research [13] [11] [12]. Han et al. [18] proposed an explainable depression detection model using a hierarchical attention network, which was adopted by Zogan et al. [28] for the same task. While there are many methodologies and models being developed for the task of depression detection, there is not a lot of work focusing on learning better data representations for identifying the cause of depression, making our contributions unique in the intersecting field of mental health and NLP. Our work would enable existing studies to build safer classifiers due to increased explainability.

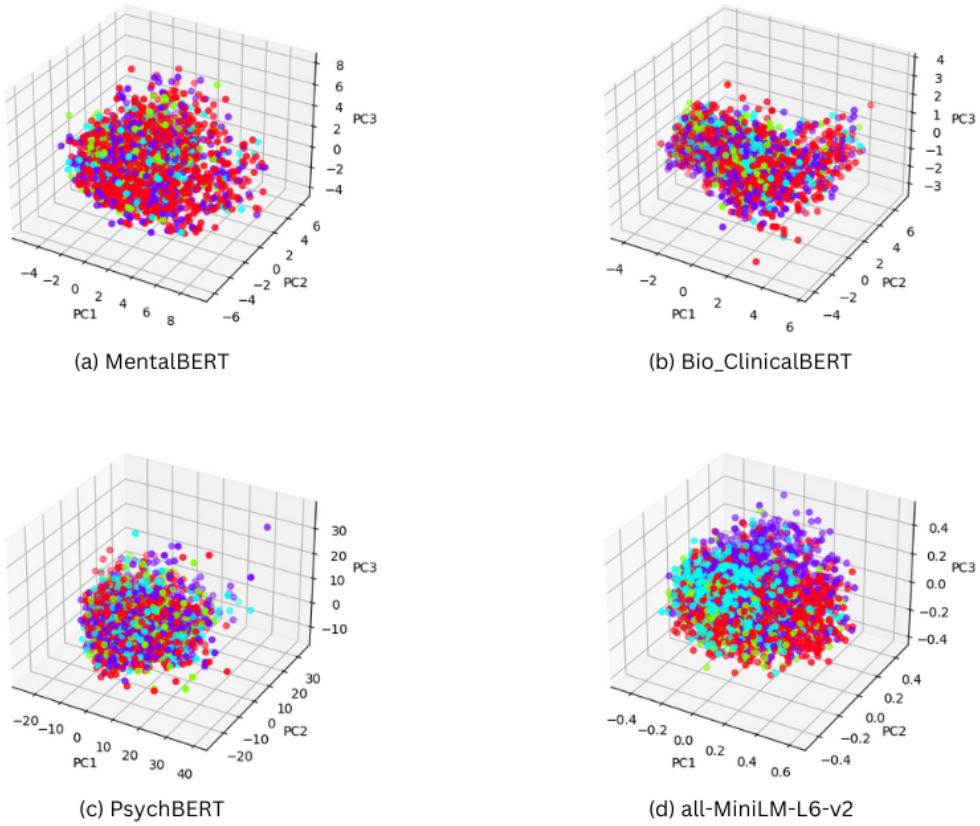

Figure 1: Embeddings of CAMS data from models pre-trained on mental health and medical data – (a) MentalBERT [20], (b) Bio_ClinicalBERT [2], (c) PsychBERT [6] reveal no significant structure for the downstream task of depression-cause identification. The sentence transformer model (d) all-MiniLM-L6-v2 [1] shows the most significant structure for the downstream task.

3 METHODOLOGY

3.1 Dataset

We use the CAMS dataset [14] for our work. The dataset contains 5051 instances of Reddit posts sampled from r/depression and r/suicidewatch. An example datapoint is shown in Figure 1. Each datapoint has been annotated for the underlying cause of depression, which is one of – bias/abuse, jobs/career, medication, relationships, alienation, or 'no cause' if the post is not about depression or does not have an identifiable cause of depression.

Along with the cause of depression, each data point has also been annotated with an inference entry, which is part of the post indicative of depression that the annotators used to classify it into one of the five causal categories. This gives novelty and uniqueness to the CAMS dataset, and makes it suitable for our work.

Table 1: An example entry from the CAMS dataset. The cause number indicates the cause of depression of the given social media post. The inference text indicates the depression.

Attribute	Value
Post	I guess I just started to realize that all of this "new year new me" is bullshit and I'm going to stay the same loser in 2018 so I'm not getting my hopes up
Cause	5
Inference	bullshit,loser,not getting my hopes up

3.2 Evaluation Metric

Existing traditional metrics like accuracy and F-scores focus on the quantitative aspect but not the qualitative aspect. In the domain of mental health, it is imperative that any model developed is not only accurate, but also safe and explainable [10]. We utilize the inference annotations from the CAMS dataset to define a custom evaluation metric that focuses on the qualitative aspect of a model.

For every data point in our test set, we consider the set of the top five tokens by attention

$$T = \{t_1, t_2, t_3, t_4, t_5\}$$

and the set of words in the inference column

$$I = \{i_1, i_2, ..., i_n\}$$

where n is the number of words in the inference column of the data point. We calculate the number of tokens the model gives high attention to, but are not present in the inference of the data point, and take its inverse. This gives us a score for the data point

$$s = \min\left(\frac{1}{|T \setminus I|}, 1\right)$$

We define a model to have a better quality if the tokens to which the model gives higher attention are present in the inference entry of the data point. Therefore, a higher s for a data point is better. We calculate the mean of s for all data points in the test set, which gives us a score for the model

$$S = \frac{1}{m}\sum_{i=1}^{m} s_i$$

where m is the number of data points in the test set. A model $M1$ is qualitatively better than a model $M2$ if

$$S_{M1} > S_{M2}$$

The complexity of accurately giving attention to t tokens varies non-linearly on t. Accurately giving attention to four tokens is more than twice as hard as accurately giving attention to two tokens. Therefore, we developed our evaluation metric to be non-linear to account for this disparity. Our proposed evaluation metric can be generalized to any t number of tokens.

3.3 Representation Learning

Figure 1 visualizes embeddings of CAMS data from various different models pre-trained on mental health and medical data. It is evident that there is a need to improve the quality of these embeddings for the downstream task of identifying the cause of depression.

A prevalent challenge is that posts categorized under the same causal category of depression may not necessarily be semantically similar. This is illustrated with the following example of two posts:

(1) My girlfriend broke up with me because she found out I cheated on her. This makes me feel so bad, and I want to end my life.
(2) I wanted to go on a trip, but my parents refused. I don't like my parents, they are such annoying people. It makes my brain go nuts.

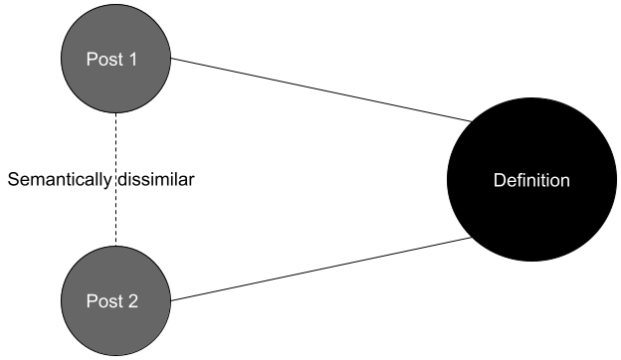

Figure 2: The annotation guideline definition of a causal category of depression can anchor two semantically dissimilar posts of the same causal category (*Post 1* and *Post 2*) together in the vector space.

The two posts provided above demonstrate that despite being categorized under the same causal category of 'relationships' as the cause of depression, they lack semantic similarity. To overcome this, we utilize the causal category definitions from the annotation guidelines of the CAMS dataset as anchors in the vector space.

We hypothesize that there is underlying similarity between the definitions of a causal category and the posts belonging to the category, if the annotators used those definitions to perform the classification. The use of such an anchor (Figure 2) is able to bring semantically dissimilar posts under the same causal category of depression closer together in the vector space.

We fine-tune pre-trained models using contrastive loss [23] – pairs of definitions and posts of the same causal category are treated as positive samples, whereas those of different causal categories are treated as negative samples.

3.4 Models

On manual evaluation of the embedding visualizations of various pre-trained models, sentence transformers [25] showed the best structure of the data for the classification task. We used the all-MiniLM-L6-v2 as our base model due to its relatively smaller size and comparatively higher performance [5]. We fine-tuned all-MiniLM-L6-v2 using AdamW as the optimizer for 15 epochs with the hyper-parameters – lr=2e-05, and bs=32. We used our custom evaluation metric (section 3.2) to assess the quality of our model at the end of each epoch. The scores on the test set are detailed in Table 2.

Using the embeddings from our fine-tuned model, we employed several machine learning models from the scikit-learn library for classifying posts into their causal category of depression. The Logistic Regression model was trained using L2 regression with C=10. The Random Forest model utilized 100 decision trees. For the SVM model, a polynomial kernel ('poly') was employed. The MLP model consisted of a single hidden layer with 170 neurons, and the training process was limited to 100 iterations. The GaussianNB model remained with its default settings, while the KNN model was configured with n_neighbors=80. These customized parameter choices

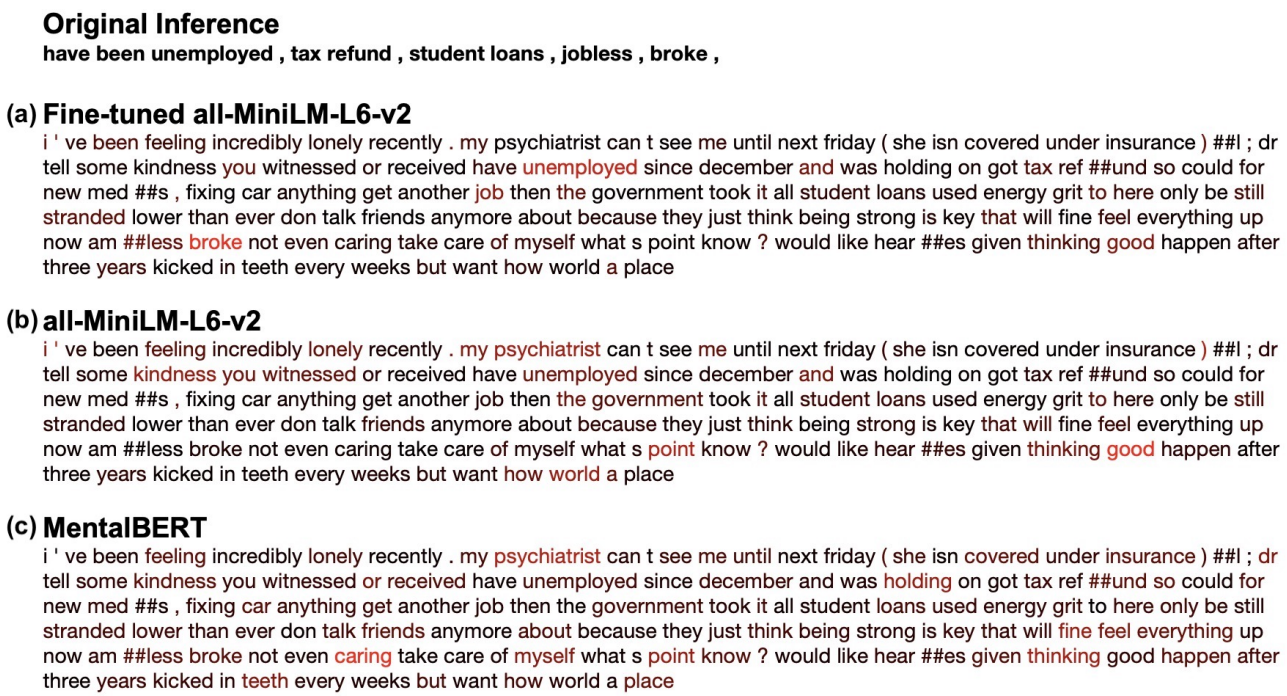

Figure 3: An example of the quality of attention to tokens in various models. The attention scores range in color from black (low) to bright red (high). The post's original inference has been provided. Our (a) fine-tuned model is able to focus on words from the inference, such as 'unemployed' and 'broke', whereas (b, c) non-fine-tuned models focus on unimportant words, such as 'psychiatrist', 'good', and 'caring'.

were made to optimize the performance of the respective models. All other parameters were left as default of the respective model classes.

4 RESULTS

Table 2: Evaluation Scores on the test set show that the fine-tuned model outperforms the models pre-trained on medical data.

Model	Evaluation Score
Finetuned all-MiniLM-L6-v2	**0.2269**
MentalBERT	0.2058
Bio_ClinicalBERT	0.2115
PsychBERT	0.2163
all-MiniLM-L6-v2	0.2163

Our fine-tuned model's evaluation scores see improvement over other pre-trained models. An evaluation score of 0.20 implies that on average, no tokens that were given attention to were present in the inference of the post; whereas, a score of 0.25 implies that on average, one token which was given attention was present in the inference of the post. We see that existing pre-trained models such as MentalBERT perform poorly (evaluation score = 0.2058) for the downstream task. On manual analysis, we find that the task of focusing attention accurately is extremely complex. Therefore, we believe that our evaluation score (0.2269) is a significant improvement.

We see that our model is able to focus attention much better than the existing pre-trained models (Figure 3). We also see improved classification scores (Table 3) with machine learning models. Since the training task is intended only for depression posts with a causal category, our classifiers are unable to perform well on class 0 ('no cause'). We observe that a simple K-nearest neighbors (KNN) model is able to perform extremely well due to the presence of clusters of posts of the same causal category in the vector space.

5 CONCLUSION AND FUTURE WORK

Cases of mental health disorders, specifically depression, are ever-growing; the use of social media and technology to improve treatment is vital. We show that although existing models perform well on the classification task, they are not explainable. We introduce a representation learning architecture using the annotation guidelines as anchors in the vector space. We define an evaluation metric to measure the explainability of a model, and show that our model is able to learn accurate post representations for the downstream task

Table 3: Classification Scores on models trained on our embeddings along with scores of the highest performing models in CAMS for comparison. SVM classifier improves overall performance, whereas KNN classifier improves performance for causal classes.

Model	F1: C0	F1: C1	F1: C2	F1: C3	F1: C4	F1: C5	Accuracy
CAMS LR	**0.63**	0.28	0.54	0.46	0.46	0.53	0.50
CAMS CNN+LSTM	0.54	0.22	0.54	0.47	**0.54**	0.47	0.48
Logistic Regression	0.54	0.41	0.65	0.49	0.53	0.58	0.54
Random Forest	0.59	0.36	0.60	0.40	0.52	0.55	0.51
SVM	0.58	0.43	0.65	0.48	0.50	0.60	**0.55**
MLP	0.57	0.44	**0.67**	0.47	0.52	0.58	0.54
GaussianNB	0.47	**0.46**	0.65	0.46	0.49	0.59	0.53
KNN	0.32	0.43	0.63	**0.50**	**0.54**	**0.61**	0.53

of identifying the cause of depression, as well as focus attention on the words that indicate the cause of depression.

We see an improvement in the classification scores. However, these scores are held back by the 'no cause' cause of depression (C0). Given that the choice of the evaluation metric is subjective and intuition-oriented, future work could comprise of a comparison study between different such evaluation scores. We also see that the models accurately focus on only one or two words that identify the cause of depression. We suggest that future work focuses on optimizing the attention such that more depression-indicative words are identified. This would enable improved recognition of the underlying cause of depression, and serve to build better self-help systems.

ACKNOWLEDGMENTS

We are grateful to the Knowledge Infused AI and Inference Lab at the University of Maryland Baltimore County for their continued support throughout the duration of this research.

REFERENCES

[1] [n. d.]. all-MiniLM-L6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Accessed: 2023-06-04.
[2] [n. d.]. ClinicalBERT - Bio + Clinical BERT Model. https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT?doi=true. Accessed: 2023-06-04.
[3] [n. d.]. Mental disorders. https://www.who.int/news-room/fact-sheets/detail/mental-disorders. Accessed: 2023-06-04.
[4] [n. d.]. Mental Health By the Numbers. https://www.nami.org/mhstats. Accessed: 2023-06-04.
[5] [n. d.]. Pretrained Models. https://www.sbert.net/docs/pretrained_models.html. Accessed: 2023-06-04.
[6] [n. d.]. PsychBERT. https://huggingface.co/mnaylor/psychbert-cased. Accessed: 2023-06-04.
[7] Nafiz Al Asad, Md Appel Mahmud Pranto, Sadia Afreen, and Md Maynul Islam. 2019. Depression detection by analyzing social media posts of user. In *2019 IEEE international conference on signal processing, information, communication & systems (SPICSCON)*. IEEE, 13–17.
[8] Theyazn H. H. Aldhyani, Saleh Nagi Alsubari, Ali Saleh Alshebami, Hasan Alkahtani, and Zeyad A. T. Ahmed. 2022. Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models. *International Journal of Environmental Research and Public Health* 19 (2022).
[9] Johnna Blair and Saeed Abdullah. 2018. Supporting constructive mental health discourse in social media. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 299–303.
[10] Giovanni Cinà, Tabea Röber, Rob Goedhart, and Ilker Birbil. 2022. Why we do need explainable ai for healthcare. *arXiv preprint arXiv:2206.15363* (2022).
[11] Mandar Deshpande and Vignesh Rao. 2017. Depression detection using emotion artificial intelligence. *2017 International Conference on Intelligent Sustainable Systems (ICISS)* (2017), 858–862.
[12] Cangnai Fang, Gracia Dianatobing, Talia Atara, Ivan Sebastian Edbert, and Derwin Suhartono. 2022. Feature Extraction Methods for Depression Detection Through Social Media Text. *2022 6th International Conference on Informatics and Computational Sciences (ICICoS)* (2022), 117–121.
[13] Martín Di Felice, Parag Chatterjee, and María Florencia Pollo Cattaneo. 2022. Depression Diagnosis using Text-based AI Methods - A Systematic Review. In *ICAI Workshops*.
[14] Muskan Garg, Chandni Saxena, Sriparna Saha, Veena Krishnan, Ruchi Joshi, and Vijay Mago. 2022. CAMS: An Annotated Corpus for Causal Analysis of Mental Health Issues in Social Media Posts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 6387–6396. https://aclanthology.org/2022.lrec-1.686
[15] Smita Ghosh. 2022. Depression Detection using Machine and Deep Learning Models to Assess Mental Health of Social Media Users. *Machine Learning Techniques and Data Science Trends* (2022).
[16] Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C. Ryan, Jonathan Marsh, Jordan Devylder, Michel Walter, Sofian Berrouiguet, and Christophe Lemey. 2021. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. *Journal of Medical Internet Research* 23 (2021).
[17] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* 18 (2017), 43–49.
[18] Sooji Han, Rui Mao, and Erik Cambria. 2022. Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings. *arXiv preprint arXiv:2209.07494* (2022).
[19] Shaoxiong Ji, Celina Ping Yu, Sai fu Fung, Shirui Pan, and Guodong Long. 2018. Supervised Learning for Suicidal Ideation Detection in Online User Content. *Complex*. 2018 (2018), 6157249:1–6157249:10.
[20] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621* (2021).
[21] Jina Kim, Jieon Lee, Eunil Park, and Jinyoung Han. 2020. A deep learning model for detecting mental illness from user content on social media. *Scientific reports* 10, 1 (2020), 1–6.
[22] Madan Krishnamurthy, Khalid Mahmood, and Pawel Marcinek. 2016. A hybrid statistical and semantic model for identification of mental health and behavioral disorders using social network analysis. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (2016), 1019–1026.
[23] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
[24] Kuldeep Kumar Patel, Anikesh Pal, Kumar Saurav, and Pooja Jain. 2022. Mental Health Detection Using Transformer BERT. *Handbook of Research on Lifestyle Sustainability and Management Solutions Using AI, Big Data Analytics, and Visualization* (2022).
[25] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).
[26] Eldar Yeskuatov, Sook-Ling Chua, and Lee Kien Foo. 2022. Leveraging Reddit for Suicidal Ideation Detection: A Review of Machine Learning and Natural Language Processing Techniques. *International Journal of Environmental Research and Public Health* 19 (2022).
[27] Tianlin Zhang, Kailai Yang, Shaoxiong Ji, and Sophia Ananiadou. 2022. Emotion fusion for mental illness detection from social media: A survey. *ArXiv* abs/2304.09493 (2022).

[28] Hamad Zogan, Imran Razzak, Xianzhi Wang, Shoaib Jameel, and Guandong Xu. 2022. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web* 25, 1 (2022), 281–304.