Multi-Explainable TemporalNet: An Interpretable Multimodal Approach using Temporal Convolutional Network for User-level Depression Detection

Anas Zafar[1], Danyal Aftab[2], Rizwan Qureshi[3*], Yaofeng Wang[3], and Hong Yan[4]
[1]Fast School of Computing, National University of Computer and Emerging Sciences, Karachi, Pakistan
[2]Technological University Dublin, Dublin, Ireland
[3]Center for Regenerative Medicine and Health, Hong Kong Institute of Science and Innovation, Chinese Academy of Sciences, Hong Kong, SAR China
[4]Department of Electrical Engineering, City University of Hong Kong, Hong Kong
{Anaszafar98, danyalftab97}@gmail.com, {Rizwan, Yaofeng.Wang}@crmh-cas.edu.hk, h.yan@cityu.edy.hk

Abstract

Multimodal depression detection through internet-based data such as social media platforms has been an important problem in the research community, aiming to predict human mental states for ensuring wellbeing of the society. Recently, attention-based networks have gained significant popularity for depression detection. However, existing multimodal methods primarily rely on images and text assuming no correlation between temporal aspects such as relative time of different posts or tweets, which is a crucial factor in deriving depression related behavior patterns. Moreover, they lack model interpretability resulting in limited understanding of how different features are contributing to the model's final prediction. In this paper, we propose Multi-Explainable TemporalNet (METN), a Temporal Convolution Network (TCN) based multi-modal transformer network with relative timestamp embeddings. We leverage pretrained foundation models for text and image embeddings and attention maps for model interpretability. We perform extensive experiments and ablation studies to validate the performance of METN for user-level depression detection task. Our model shows state-of-the-art results on various benchmarks, such as ***0.945 F1 score*** *on multimodal Twitter dataset, and* ***0.913 F1 score*** *on multimodal Reddit dataset. We further demonstrate that our model enhances the accuracy of identifying depression in individuals who publicly post messages on social media platforms with enhanced interpretable compatibility. Code and models are available at Github.* [1]

1. Introduction

Depression is one of the most common and crucial psychiatric disorder in today's community. According to World Health Organization (WHO)[19], the severity of depression, known as Major Depressive Disorder (MDD), hinders daily activities for over 280 million people globally, leading to suicides[30], self-harm and mass shootings[32], causing the loss of innocent lives. The symptoms of depression[30] can vary drastically ranging from low self-esteem to extreme mood swings and cynical behavior. Due to its subjective nature, depression detection is a challenging problem in modern day community. Approximately 33% of patients with depression are not identified during clinical diagnostic procedures, and less than 40% receive appropriate treatment[13]. Therefore, a range of methods have been developed using assistive technologies for early diagnosis for depression detection[35]. Social media platforms such as twitter[10] and reddit[20] have been very popular in today's world, where users can share their thoughts and mental issues freely. Reddit contains different threads where individuals can discuss their problems with complete anonymity. Many methods[1, 2, 21, 48] leverage this social media user data for early depression detection using pretrained language models such as BERT[12] and LSTM[45] based models, leading to significantly better results when compared with traditional machine learning approaches. Recent works[17, 33, 36, 46] such as multimodal depression techniques use both visual and text data for detecting depression, but fail to leverage temporal cues such as relative time between different tweets and post leading to degraded performance. Moreover, current state-of-the-art models[31, 47, 50] lack model interpretability and explainability causing trust issues with the users, as they are not able to understand the driving features causing depression.

To address these limitations, we propose Multi-

[1]* Corresponding author

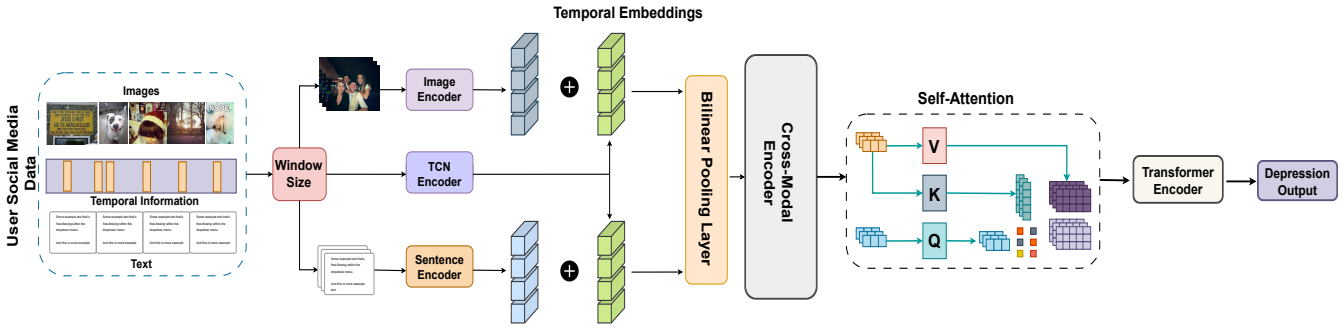

Figure 1. Overview of the MTEN architecture for user-level depression detection task. A window size of n posts are sampled from a user's profile containing images, texts and time stamps. The images and text are then encoded using pretrained encoders which are then fused with temporal embeddings obtained from the TCN encoder. These are further refined by a bilinear pooling layer and fed into a cross modal encoder. Self-attention and cross attention is then applied on these embeddings which are then later fed into a classification head for depression detection.

Explainable TemporalNet (MTEN), a Temporal Convolution Network (TCN)[25] based multimodal transformer with temporal embeddings for depression detection from social media datasets [14, 16, 42]. TCN is a framework that utilizes casual convolutions and dilations for sequential data addressing temporal data and large receptive fields. MTEN utilizes pretrained state-of-the-art models for generating text and image embeddings, and later fuses them with time stamp embeddings obtained through a TCN encoder. We leverage self-attention and cross-attention mechanisms that assigns weights to different parts of an image and sentences. The attention maps generated from these modules are used for model interpretability, in order to understand how different features in an image or text are contributing to the final model prediction. We used multimodal Twitter[16] and multimodal Reddit[42] datasets for training and evaluation purposes and our model achieved state-of-the-art results for user-level depression detection task.

Our main contributions in this work are:

- We propose a novel TCN based multi-modal transformer network (MTEN) which leverages temporal cues; relative time stamps, along with image and text embeddings for enhanced user-level depression detection and applied self and cross attention mechanism through which it dynamically adapts the fusion of weights for different modalities.
- Within the MTEN Block, we introduce an interpretable module through attention maps for better understanding and explainability of how different features in an image or text are contributing to the model's final prediction.
- We perform extensive experiments and ablation studies to validate the performance of METN. Our method achieves state-of-the-art results on multi-modal twitter dataset 0.945 F1 score and 0.913 F1 score on multi-modal reddit dataset.

The paper is organized as follow; Section 2 provides a comprehensive overview of related work in multi-modal depression detection task, and Section 3 presents our architecture in detail. Section 4 presents results and experiments, including an extensive ablation study, and qualitative and quantitative evaluations, comparing our approach with other multimodal depression detection baselines. In Section V, we discuss some of the limitations of the proposed model, and Section VI summarizes our research findings and potential future works.

2. Related Work

There have been several works on depression detection over the years. Text-based approaches using classical machine learning algorithms[6, 24, 38, 49] mostly used number of tweets and social interaction as network features. Different linguistic features such as profile bio, comments and status data were utilized from social media platforms for user-level depression detection. However, due to their limitations they failed to generalize in most of the unseen cases. In recent years, many deep learning algorithms[2, 34, 41, 43] were able to achieve state-of-the-art results across different data modalities. Rao *et al.*[34] proposed a Multi-Gated LeakyReLU CNN (MGL-CNN) for identifying depressed individuals in online forums using list of words and posts. Trotzek *et al.*[41] developed a convolutional neural network based on different word embeddings using ensemble learning. Uban *et al.*[43] developed a correlation between depressed users and images containing different animals.

Recently, multi-modal architectures and vision language models [18, 26, 29, 36, 44, 46] have become extremely popular, leveraging both text and image features and are able to achieve superior results compared to prior baseline approaches. Lin *et al.*[12]proposed a CNN-based architecture

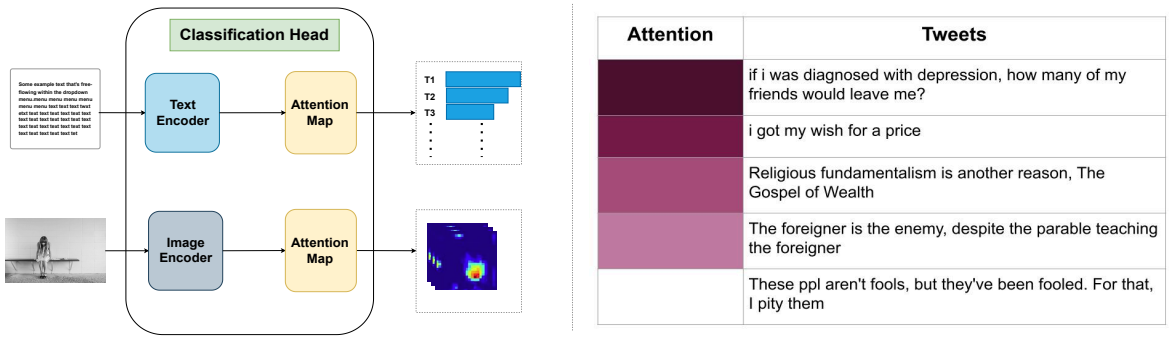

Figure 2. Attention maps from each encoder are used for model interpretabilty

using multi-modal features by utilizing Bidirectional Encoder Representations from Transformers (BERT) for textual features and CNN for visual features. Gui *et al.*[17] introduced a novel cooperative multi-agent model based on text and image features using a reinforcement learning based approach. Fang *et al.*[9] used a Bi-LSTM with attention mechanism to learn visual feature and rich text feature, respectively. However, these approaches ignore temporal cues such as posting time or assume that the posts were posted in a synchronous manner. Stankevich *et al.*[39] and Choudhury *et al.*[11] demonstrate that depression symptoms tends to higher at late night interval, highlighting the importance of temporal information such as posting time for depression detection output.

Although, there has been some recent works incorporating time component such as Sawhney *et al.*[37] introduced a STATENet, a time-aware transformer incorporating historical contexts of tweets as well. Bucur *et al.*[5] proposed a time-enriched multimodal transformer architecture for detecting depression from social media posts using time2vec positional embeddings. However, these approaches do not effectively handle temporal dependencies. Time-enriched multimodal transformer handles sequential processing of user posts by treating the user timeline as a "bag-of-posts" which does not accurately capture temporal information. Similarly, these approaches lack model explainability thus limiting user understanding of how different features are contributing to final prediction of depression detection.

To address the above limitations we introduce Multi-Explainable TemporalNet (MTEN). A TCN based multi-modal transformer which leverages temporal cues such as relative time stamps, along with image and text embeddings for enhanced user-level depression detection. Through self and cross attention mechanism it allows to dynamically adapt the fusion of weights for different modalities. We utilize attention maps for each modality for model interpretability, helping users to understand key features leading to depression.

3. Method

3.1. Framework

The proposed MTEN is a TCN based multi-modal transformer incorporating temporal cues such as posting time stamps along with image and textual features along with an interpretable module for better understanding of different features, as shown in Figure 1. Many existing approaches struggle with long-range temporal information and modality fusion. In contrast MTEN, effectively captures long-range dependencies and temporal information over long sequences. It efficiently fuses different modalities while preserving rich features leading to more accurate predictions. It consists of the following properties; (i) the ability to capture long-range temporal dependencies (ii) self and cross attention mechanism for preserving features and attention maps for interpretability (iii) effective modality fusion.

3.1.1 Model Pipeline

There is a User U, who has multiple user posts U_p such that each post has a posting time *J*, an image *I* and text *T*. Using a post sequence P_x is defined as S sampled posts from user $U_p : P_x = \{(T^y, I^y, \Delta^y) \sim U_p \mid y \in (1 \dots S)\}$ where y is a specific post within the user's post-sequence and Δ is the posting time. Different window sizes of user posts are sampled from the user's profile which is then used as an input batch for training. The images and text are encoded using pretrained foundation models; CLIP[33] for images and EmoBERTa[22] for textual data due to their ability to accurately capture these features. Contrastive Language-Image Pre-training (CLIP) is a a vision transformer that is trained to match images with corresponding textual descriptions, enabling it to perform well on various visual tasks without task-specific training. EmoBERTa utilizes both textual and emotional data from users' posts. Instead of using same encoders, we use two different encoders, as foundation models pretrained on ImageNet tend to output general purpose em-

Method	Modality	F1	Prec	Recall	Acc
T-LSTM [4]	T	0.848±8e-3	0.896±2e-2	0.804±1e-2	0.855±5e-3
EmoBERTa Transformer	T	0.864±1e-2	0.843±1e-2	0.887±3e-2	0.861±1e-2
LSTM + RL [15]	T	0.871	0.872	0.870	0.870
CNN + RL [15]	T	0.871	0.871	0.871	0.871
MTAL [3]	T+I	0.842	0.842	0.842	0.842
GRU + VGG-Net + COMMA [17]	T+I	0.900	0.900	0.901	0.900
MTAN [8]	T+I	0.908	0.885	0.931	-
Vanilla Transformer [5]	T+I	0.886±1e-2	0.868±2e-2	0.905±2e-2	0.883±5e-3
SetTransformer [5]	T+I	0.927±8e-3	0.921±1e-2	0.934±2e-2	0.926±8e-3
Time2VecTransformer [5]	T+I	0.931±4e-3†	0.931±2e-2†	0.931±1e-2†	0.931±4e-3†
MTEN (Ours)	**T+I**	**0.945±1e-2**	**0.945±2e-2**	**0.945±2e-2**	**0.945±5e-3**

Table 1. Quantitative results of multimodal depression detection on twitter dataset[17]

Method	Modality	F1	Prec.	Recall	Acc
T-LSTM [4]	T	0.831 ± 0.01	0.825 ± 0.008	0.837 ± 0.01	0.872±7e-3
EmoBERTa Transformer	T	0.843 ± 0.006	0.828 ± 0.003	0.858 ± 0.01	0.879±4e-3
Uban et al. [43]	T+I	-	-	-	0.663
Vanilla Transformer [5]	T+I	0.837 ± 0.008	0.827 ± 0.01	0.848 ± 0.01	0.876±6e-3
Set Transformer [5]	T+I	0.902 ± 0.007†	0.878 ± 0.006†	$\mathbf{0.928 \pm 0.01}$†	$0.924 \pm 5e-3$†
Time2Vec Transformer [5]	T+I	0.869 ± 0.007	0.869 ± 0.007	0.869 ± 0.008	$0.901 \pm 5e-3$
MTEN (Ours)	**T+I**	**0.913±0.05**	**0.913±0.05**	0.913±0.05	**0.926±0.002**

Table 2. Quantitative results of multimodal depression detection on the Reddit dataset [42].

beddings. Given the diversified nature of our dataset these general purpose embeddings are not suitable for our use-case. The posting time is encoded through a TCN encoder which is able to capture long-range dependencies and handles very long input sequences efficiently as it computes all outputs in parallel during training. The TCN encoder contains causal convolution layers which preserves the temporal information of the user's post is able to capture long range dependences due to dilated convolution present in its architecture.

We use Self-DIstillation with NO Labels (DINO)[7] as regularizer for CLIP generated embeddings for enriched feature representations and making it more robust to noise and variations in the input data. DINO is a self-supervised vision transformer that directly predicts the output of a teacher network using cross entropy loss. During this process CLIP embeddings are fed into DINO, which refines these embedding by making it more consistent with DINO trained network's output. The image and the text embeddings are then projected to a fixed size through a learnable linear feed-forward layer. The attention maps obtained from the text and image encoder are used for model interpretability and explainability, of how different features are contributing the output as show in Figure 2. To make it time-aware, we fuse these image and text embeddings with temporal embeddings generated by the TCN encoder followed by bi-linear pooling layer to capture all pairwise interactions between features from these modalities, hence improving model's robustness to noise and variations. Although, EmoBERTa and CLIP have distinct latent spaces, in our approach we leverage ViLBERT (Vision-and-Language BERT)[28], a cross modal encoder for the alignment between these two modality embeddings. ViLBERT supports cross-modal interactions through its co-attention transformer layers. It allows joint representation learning by bridging the gap between the two generated embeddings from CLIP and EmoBERTa. Through this it ensures that the features from both the modalities are effectively fused and are aligned in the latent space hence obtaining richer representation for joint learning representations.

We also use multimodal dropout in which we randomly drop modalities during training to make our model more robust to missing data. The embeddings are then processed through cross and self-attention across different user posts, after which they are finally fed into a classification head for depression detection. We use binary cross entropy loss for training this model. For model interpretabilty we leverage attention maps from text and image encoder as shown in Figure 2. The attention maps demonstrate how different features are contributing to the model prediction.This figure shows METN assigns higher weights to depressed tweets compared to less relevant tweets highlighting which tweets are more responsible to the final model output.

Dataset	Class	T	(T+I)
Reddit	Depr	6,6M	46.9K
	Non-Depr	8,1M	73.4K
Twitter	Depr	213K	19.3K
	Non-Depr	828K	50.6K

Table 3. Data distribution of different modalities using benchmark Reddit and Twitter dataset, T represents number of posts having only text and T+I represents number of posts having both images and text

4. Experiments and Results

4.1. Dataset & Evaluation Metrics

We trained MTEN using benchmark multimodal twitter dataset[16]and multimodal Reddit dataset[42] for user-level depression detection. The multimodal twitter dataset consists of 1,402 depressed users and 1,402 control users. Similarly, the reddit dataset has 1,419 depressed users and 2,344 control users. Reddit has a more detailed textual information containing upto 40,000 characters, whereas twitter dataset only contains 280 characters. Please refer to Table 3 for detailed data distribution for the two multimodal datasets. For training, we followed the same setting as Gui *et al.*[17] and Bucur *et al.*[5]. To evaluate the results, we calculate the Precision, Recall, F1-score and accuracy metrics.

4.2. Implementation Details

We trained the model using Adam[23] optimizer with (1=0.9, 2=0.9, weight decay 0) for 10 epochs. The initial learning rate is set to 1e4, which gradually reduced to 1e6 with the cosine annealing scheduler[27]. The model consists of consists of four cross-encoder layers, each with eight heads, and an embedding size of 128. For evaluation purposes 10 samples per user is used for each classification label. We use the PyTorch framework to implement our architecture, and trained our model using NVIDIA DGX server with a NVIDIA A-100 GPU, having 40-GB RAM.

4.3. Results and Comparison With Other Methods

We evaluate our model on both multi-modal benchmark twitter and reddit dataset. Quantitative results are shown in Table 1 and Table 2 using precision, recall, F1 score and accuracy metrics. We compare our model against state-of-the-art multimodal and text-based baselines. MTEN was able to achieve state-of-the-art results on both the multimodal datasets with an f1 score of 0.945 F1 score on multimodal twitter dataset, and 0.913 F1 score multimodal reddit dataset respectively. Our model outperforms all the time-enriched models Time2VecTransformer[5], SetTransformer[5], T-LSTM[4] and MTAN[8] on both the datasets demonstrating that our model is able to accurately capture long-range temporal information. We also obtain attention maps for each modality for model interpretability demonstrating how different features are contributing to the models predictions and can seen in Figure 2.

4.4. Ablation Study

We performed an in depth ablation study to explain the role of different components in our proposed MTEN architecture. We evaluated our model on both LXMERT[40] and ViLBERT (Vision- and-Language BERT[28] for the cross modal encoder as shown in Table 4. MTEN(ViLBERT) can be observed to achieve a performance boost (0.94 F1 score) when as compared to LXMERT for user-level depression detection. Similarly, we also compare DINO vs CLIP for image encoder and we can observe an improvement in the performance when we regularize CLIP with DINO embeddings.

To decide the best window size we also carried an ablation study on different window sizes $W = \{16, 32, 64, 128, 256, 512\}$ as show in Table 5 using CLIP, EmoBERTa and ViLBERT for text and image embeddings. For twitter dataset, window size 256 is able to the highest accuracy of 0.945 F1 score. For reddit dataset, window size 128 is best suitable accurate predictions.

5. Limitations

Both twitter and reddit datasets are limited datasets with a bias towards USA demographics group and may not be a true reflection of all the features related to depression detection. As depression symptoms may drastically vary and can be highly subjective, this work aims to only highlight depression related features and should not be primary source for depression diagnosis. The performance of MTEN is bottlenecked by the quality of the training dataset. Due to privacy concerns depression detection datasets are sparse, we plan to use synthetic data in the future to overcome this gap by learning data distribution and depression features from real data and at the same time maintaining user anonymity. Future iterations could include training the model with limited data or exploring zero-shot learning for better generalizabilty and enhanced model capabilities.

6. Conclusion

In this paper we introduce Multi-Explainable Temporal-Net (METN), a novel TCN-based multimodal transformer network incorporating temporal embeddings and attention mechanisms for enhanced user-level depression detection on social media datasets. The model utilizes pre-trained foundation models for text and image embeddings; EmoBERTA and CLIP and attention maps for interpretabil-

	Twitter	Reddit
Text+Image Enc	**F1**	**F1**
EmoBERTa+CLIP	0.930	0.891
EmoBERTa+DINO	0.916	0.873
EmoBERTa+CLIP+LXMERT	0.934	0.897
EmoBERTa+DINO+LXMERT	0.932	0.894
EmoBERTa+CLIP+LXMERT (Dino Regularizer)	0.936	0.899
EmoBERTa+CLIP+ViLBERT	0.943	0.911
EmoBERTa+DINO+ViLBERT	0.941	0.908
EmoBERTa+CLIP+ViLBERT (Dino Regularizer)	**0.945**	**0.913**

Table 4. Ablation study results on MTEN using different image encoders and cross modal encoders on Twitter and Reddit dataset

Dataset	Window Size	F1-Score
Twitter	16	0.933
	32	0.938
	64	0.941
	128	0.943
	256	0.945
	512	0.942
Reddit	16	0.895
	32	0.905
	64	0.910
	128	0.913
	256	0.912
	512	0.910

Table 5. Ablation study results on different window sizes on both twitter and reddit dataset

ity. We leverage TCN encoder which is able to capture long-range dependencies and input sequences efficiently as it computes all outputs in parallel during training. Extensive experiments demonstrate METN's state-of-the-art performance on multimodal Twitter and Reddit datasets, with F1 scores of 0.945 and 0.913, respectively. The model's interpretable compatibility and attention mechanisms provide insights into the contribution of different features to the final prediction, offering a promising approach for user-level depression detection with improved accuracy and understanding. This model can be potentially be used as an assistive technology for early depression detection among users on social media platforms.

7. Acknowledgement

This work is supported by the Research Funds from Health@InnoHK Program launched by Innovation Technology Commission of the Hong Kong SAR, China.

References

[1] Manex Agirrezabal and Janek Amann. Kucst@ lt-edi-acl2022: Detecting signs of depression from social media text. *arXiv preprint arXiv:2204.04481*, 2022. 1

[2] Hassan Alhuzali, Tianlin Zhang, and Sophia Ananiadou. Predicting sign of depression via using frozen pre-trained models and random forest classifier. In *CLEF (Working Notes)*, pages 888–896, 2021. 1, 2

[3] Minghui An, Jingjing Wang, Shoushan Li, and Guodong Zhou. Multimodal topic-enriched auxiliary learning for depression detection. In *proceedings of the 28th international conference on computational linguistics*, pages 1078–1089, 2020. 4

[4] Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 65–74, 2017. 4, 5

[5] Ana-Maria Bucur, Adrian Cosma, Paolo Rosso, and Liviu P Dinu. It's just a matter of time: Detecting depression with time-enriched multimodal transformers. In *European Conference on Information Retrieval*, pages 200–215. Springer, 2023. 3, 4, 5

[6] Sergio G Burdisso, Marcelo Errecalde, and Manuel Montes-y Gómez. A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, 133:182–197, 2019. 2

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 4

[8] Ju Chun Cheng and Arbee LP Chen. Multimodal time-aware attention networks for depression detection. *Journal of Intelligent Information Systems*, 59(2):319–339, 2022. 4, 5

[9] Qing Cong, Zhiyong Feng, Fang Li, Yang Xiang, Guozheng Rao, and Cui Tao. Xa-bilstm: a deep learning approach for depression detection in imbalanced data. In *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 1624–1627. IEEE, 2018. 3

[10] Aron Culotta, Nirmal Kumar, and Jennifer Cutler. Predicting the demographics of twitter users from website traffic data. In *Proceedings of the AAAI conference on artificial intelligence*, 2015. 1

[11] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, pages 128–137, 2013. 3

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional

transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2

[13] Leonard E Egede. Failure to recognize depression in primary care: issues and challenges. *Journal of General Internal Medicine*, 22:701–703, 2007. 1

[14] Muskan Garg. Mental health analysis in social media posts: a survey. *Archives of Computational Methods in Engineering*, 30(3):1819–1842, 2023. 2

[15] Tao Gui, Qi Zhang, Liang Zhu, Xu Zhou, Minlong Peng, and Xuanjing Huang. Depression detection on social media with reinforcement learning. In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 613–624. Springer, 2019. 4

[16] Tao Gui, Liang Zhu, Qi Zhang, Minlong Peng, Xu Zhou, Keyu Ding, and Zhigang Chen. Cooperative multimodal approach to depression detection in twitter. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):110–117, 2019. 2, 5

[17] Tao Gui, Liang Zhu, Qi Zhang, Minlong Peng, Xu Zhou, Keyu Ding, and Zhigang Chen. Cooperative multimodal approach to depression detection in twitter. In *Proceedings of the AAAI conference on artificial intelligence*, pages 110–117, 2019. 1, 3, 4, 5

[18] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*, 2023. 2

[19] Yueqin Huang, YU Wang, Hong Wang, Zhaorui Liu, Xin Yu, Jie Yan, Yaqin Yu, Changgui Kou, Xiufeng Xu, Jin Lu, et al. Prevalence of mental disorders in china: a cross-sectional epidemiological study. *The Lancet Psychiatry*, 6(3):211–224, 2019. 1

[20] Matthew R Jamnik and David J Lane. The use of reddit as an inexpensive source for high-quality data. *Practical Assessment, Research, and Evaluation*, 22(1):5, 2019. 1

[21] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, page 2, 2019. 1

[22] Taewoon Kim and Piek Vossen. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*, 2021. 3

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[24] Akshi Kumar, Aditi Sharma, and Anshika Arora. Anxious depression prediction in real-time social data. *arXiv preprint arXiv:1903.10222*, 2019. 2

[25] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017. 2

[26] Chenhao Lin, Pengwei Hu, Hui Su, Shaochun Li, Jing Mei, Jie Zhou, and Henry Leung. Sensemood: depression detection on social media. In *Proceedings of the 2020 international conference on multimedia retrieval*, pages 407–411, 2020. 2

[27] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 4, 5

[29] Paulo Mann, Aline Paes, and Elton H Matsushima. See and read: detecting depression symptoms in higher education students using multimodal social media data. In *Proceedings of the International AAAI Conference on Web and social media*, pages 440–451, 2020. 2

[30] Douglas M Maurer. Screening for depression. *American family physician*, 85(2):139–144, 2012. 1

[31] Roger S McIntyre, Mohammad Alsuwaidan, Bernhard T Baune, Michael Berk, Koen Demyttenaere, Joseph F Goldberg, Philip Gorwood, Roger Ho, Siegfried Kasper, Sidney H Kennedy, et al. Treatment-resistant depression: definition, prevalence, detection, management, and investigational interventions. *World Psychiatry*, 22(3):394–412, 2023. 1

[32] Jonathan M Metzl and Kenneth T MacLeish. Mental illness, mass shootings, and the politics of american firearms. *American journal of public health*, 105(2):240–249, 2015. 1

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3

[34] Guozheng Rao, Yue Zhang, Li Zhang, Qing Cong, and Zhiyong Feng. Mgl-cnn: a hierarchical posts representations model for identifying depressed individuals in online forums. *IEEE Access*, 8:32395–32403, 2020. 2

[35] Charles F Reynolds III, Pim Cuijpers, Vikram Patel, Alex Cohen, Amit Dias, Neerja Chowdhary, Olivia I Okereke, Mary Amanda Dew, Stewart J Anderson, Sati Mazumdar, et al. Early intervention to reduce the global health and economic burden of major depression in older adults. *Annual review of public health*, 33:123–135, 2012. 1

[36] Ramin Safa, Peyman Bayat, and Leila Moghtader. Automatic detection of depression symptoms in twitter using multimodal analysis. *The Journal of Supercomputing*, 78(4): 4709–4744, 2022. 1, 2

[37] Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Shah. A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7685–7697, 2020. 3

[38] Tiancheng Shen, Jia Jia, Guangyao Shen, Fuli Feng, Xiangnan He, Huanbo Luan, Jie Tang, Thanassis Tiropanis, Tat Seng Chua, and Wendy Hall. Cross-domain depression detection via harvesting social media. International Joint Conferences on Artificial Intelligence, 2018. 2

[39] Maxim Stankevich, Vadim Isakov, Dmitry Devyatkin, and Ivan V Smirnov. Feature engineering for depression detection in social media. In *ICPRAM*, pages 426–431, 2018. 3

[40] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 5

[41] Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 32(3): 588–601, 2018. 2

[42] Ana-Sabina Uban, Borja Chulvi, and Paolo Rosso. Explainability of depression detection on social media: From deep learning models to psychological interpretations and multimodality. pages 289–320, 2022. 2, 4, 5

[43] Ana-Sabina Uban, Berta Chulvi, and Paolo Rosso. Explainability of depression detection on social media: From deep learning models to psychological interpretations and multimodality. In *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the eRisk Project*, pages 289–320. Springer, 2022. 2, 4

[44] Nikhita Vedula and Srinivasan Parthasarathy. Emotional and linguistic cues of depression from social media. In *Proceedings of the 2017 International Conference on Digital Health*, pages 127–136, 2017. 2

[45] Shuohang Wang and Jing Jiang. Learning natural language inference with lstm. *arXiv preprint arXiv:1512.08849*, 2015. 1

[46] Zhentao Xu, Verónica Pérez-Rosas, and Rada Mihalcea. Inferring social media users' mental health status from multimodal information. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6292–6299, 2020. 1, 2

[47] Uma Yadav and Ashish K Sharma. A novel automated depression detection technique using text transcript. *International Journal of Imaging Systems and Technology*, 33(1): 108–122, 2023. 1

[48] Andrew Yates, Arman Cohan, and Nazli Goharian. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*, 2017. 1

[49] Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pages 1191–1198, 2017. 2

[50] Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. Depressionnet: learning multi-modalities with user post summarization for depression detection on social media. In *proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 133–142, 2021. 1