Measuring Depression Symptom Severity from Spoken Language and 3D Facial Expressions

Albert Haque[1] **Michelle Guo**[1] **Adam S Miner**[2,3] **Li Fei-Fei**[1]
[1]Department of Computer Science, Stanford University
[2]Department of Psychiatry and Behavioral Sciences, Stanford University
[3]Department of Health Research and Policy, Stanford University

Abstract

With more than 300 million people depressed worldwide, depression is a global problem. Due to access barriers such as social stigma, cost, and treatment availability, 60% of mentally-ill adults do not receive any mental health services. Effective and efficient diagnosis relies on detecting clinical symptoms of depression. Automatic detection of depressive symptoms would potentially improve diagnostic accuracy and availability, leading to faster intervention. In this work, we present a machine learning method for measuring the severity of depressive symptoms. Our multi-modal method uses 3D facial expressions and spoken language, commonly available from modern cell phones. It demonstrates an average error of 3.67 points (15.3% relative) on the clinically-validated Patient Health Questionnaire (PHQ) scale. For detecting major depressive disorder, our model demonstrates 83.3% sensitivity and 82.6% specificity. Overall, this paper shows how speech recognition, computer vision, and natural language processing can be combined to assist mental health patients and practitioners. This technology could be deployed to cell phones worldwide and facilitate low-cost universal access to mental health care.

1 Introduction

Worldwide, more than 300 million people are depressed [48]. In the worst case, depression can lead to suicide, with close to 800,000 people committing suicide every year. In general, patients with mental disorders are seen by a wide spectrum of health care providers, including primary care physicians [22]. However, compared to physical illnesses, mental disorders are more difficult to detect. The burden of mental health is exacerbated by barriers to care such as social stigma, financial cost, and a lack of accessible treatment options. To address entrenched barriers to care, scalable approaches for detecting mental health symptoms have been called for [19]. If successful, early detection may impact access for the 60% of mentally-ill adults who do not receive treatment [33].

In practice, clinicians identify depression in patients by first measuring the severity of *depressive symptoms*[1] during in-person clinical interviews. During these interviews, clinicians assess both verbal and non-verbal indicators of depressive symptoms including monotone pitch, reduced articulation rate, lower speaking volumes [16, 41], fewer gestures, and more downward gazes [46, 40, 37]. If such symptoms persist for two weeks [4], the patient is considered to have a *major depressive episode*. Structured questionnaires have been developed and validated in clinical populations to assess the severity of depressive symptoms. One of the most common questionnaires is the Patient Health Questionnaire (PHQ) [23]. This clinically-validated tool measures depression symptom severity across several personal dimensions [21]. Assessing symptom severity is time-intensive, and critical for both initial diagnosis and improvement across time. Thus, AI-based solutions to assessing symptom severity may address entrenched barriers to access and treatment.

[1] Depressive symptoms include feelings of worthlessness, loss of interest in hobbies, or thoughts of suicide.

Machine Learning for Health (ML4H) Workshop at NeurIPS 2018, Montréal, Canada.

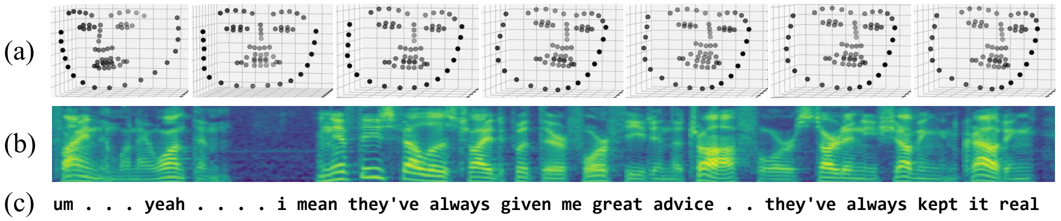
(a)
(b)
(c)
um . . . yeah . . . . i mean they've always given me great advice . . they've always kept it real

Figure 1: **Multi-modal data.** For each clinical interview, we use: (a) video of 3D facial scans, (b) audio recording, visualized as a log-mel spectrogram, and (c) text transcription of the patient's speech. Our model predicts the severity of depressive symptoms using all three modalities.

We envision an AI-based solution where depressed individuals can receive evidence-based mental health services while avoiding existing barriers to access. Such a solution could leverage multi-modal sensors or text messages, as is common on modern smartphones, to increase timely and cost-effective symptom screening [3]. Conversational AIs are another potential solution [31, 32]. Our hope is that automated feedback will (i) provide actionable feedback to individuals who may be depressed, and (ii) improve automated depression screening tools for clinicians, by including visual, audio, and linguistic signals.

Contributions. We propose a machine learning method for measuring depressive symptom severity from de-identified multi-modal data. The input to our model is audio, 3D video of facial keypoints, and a text transcription of a patient speaking during a clinical interview. The output of our model is either a PHQ score or classification label indicating major depressive disorder. Our method leverages a causal convolutional network (C-CNN) to "summarize" sentences into a single embedding. This embedding is then used to predict depressive symptom severity. In our experiments, we show how our sentence-based model performs in relation to word-level embeddings and prior work.

2 Dataset

We use the DAIC-WOZ dataset [15] containing audio and 3D facial scans of depressed and non-depressed patients. For each patient, we are provided with the PHQ-8 score. This corpus is created from semi-structured clinical interviews where a patient speaks to a remote-controlled digital avatar. The clinician, through the digital avatar, asks a series of questions specifically aimed at identifying depressive symptoms. The agent prompts each patient with queries that included questions (e.g. "How often do you visit your hometown?") and conversational feedback (e.g. "Cool."). A total of 50 hours of data was collected from 189 clinical interviews from a total of 142 patients. Following prior work [2], results in our paper are from the validation set. More details can be found in Appendix A.

Privacy. Data used in this work does not contain protected health information (PHI). Mentions of personal names, specific dates, and locations were removed from the audio recording and transcription by the dataset curators [15]. The 3D facial scans are low-resolution (68 pixels) and do not contain enough information to identify the individual but contain just enough to measure facial motions such as eye, lip, and head movements. While the dataset is publicly available, future researchers who apply our method to other datasets may encounter PHI and should design their experiments appropriately.

3 Model

Our model consists of two technical pieces: (i) a sentence-level "summary" embedding and (ii) a causal convolutional network (C-CNN). An overview is shown in Figure 2.

Sentence-Level Embeddings. For decades, word and phoneme-level embeddings[2] have been the go-to feature for encoding text and speech [10, 8, 36, 43]. While these embeddings work well for some tasks [42, 20, 38], they are limited in their sentence-level modeling ability. This is because word and phoneme embeddings capture a narrow temporal context, often a few hundred milliseconds at most [45, 24]. In this work, we propose a novel multi-modal sentence-level embedding. This allows us to capture long-term acoustic, visual, and linguistic elements.

[2] The goal of an *embedding* is "summarize" a variable-length sequence as a fixed-size vector of numbers.

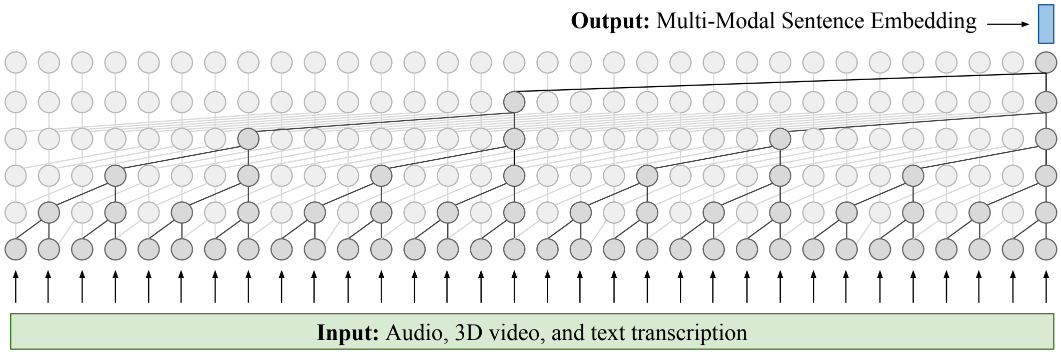
Output: Multi-Modal Sentence Embedding
Input: Audio, 3D video, and text transcription

Figure 2: **Our method: Learning a multi-modal sentence embedding.** Overall, our model is a causal CNN [5]. The input to our model is: audio, 3D facial scans, and text. The multi-modal sentence embedding is fed to a depression classifier and PHQ regression model (not shown above).

		Classification: Major Depressive Disorder				Regression: PHQ Score
#	Method	Modalities	F1 Score	Precision	Recall	Average Error
1	SVM [44]	A	46.2	31.6	85.7	6.74
2	CNN+LSTM [27]	A	52.0	35.0	100.0	—
3	SVM [44]	V	50.0	60.0	42.8	7.13
4	Williamson et al. [47]	V	53.0	—	—	5.33
5	Williamson et al. [47]	L	84.0	—	—	3.34
6	Alhanai et al. [2]	AL	77.0	71.0	83.0	5.10
7	SVM [44]	AV	50.0	60.0	42.8	6.62
8	Gong et al. [14]	AVL	70.0	—	—	2.77
9	Williamson et al. [47]	AVL	81.0	—	—	4.18
10	C-CNN [5]	AVL	76.9	71.4	83.3	3.67

Table 1: **Comparison of machine learning methods for detecting depression.** Two tasks were evaluated: (i) binary classification of major depressive disorder and (ii) PHQ score regression. Modalities: A: audio, V: visual, L: linguistic (text), AVL: combination. For prior work, numbers are reported from the original publications. Dashes indicate the metric was not reported.

Causal Convolutional Networks. During clinical interviews, patients may stutter and frequently pause between words. This causes audio-video recordings to be longer than non-depressed patients. Recently, causal convolutional networks (C-CNNs) have been shown to outperform recurrent neural networks (RNNs) on long sequences [5]. In [30], the authors even show that RNNs can be approximated by fully feed-forward networks (i.e., CNNs). Combined with dilated convolutions [34], C-CNNs are well-poised to model the long sequences for depression screening interviews. For a more thorough comparison of C-CNNs vs RNNs, we refer the reader to Bai et al. [5].

4 Experiments

Our experiments consist of two parts. First, we compare our method to existing works for measuring the severity of depressive symptoms (Table 1). We predict both the PHQ score and output a binary classification as to whether the patient has *major depressive disorder*, typically with a PHQ score greater than or equal to ten [28]. Second, we perform ablation studies on our model to better understand the effect of multiple modalities and sentence-level embeddings (Table 2). Data formats, neural network architectures, and key hyperparameters can be found in Appendix A.

4.1 Automatically Measuring the Severity of Depressive Symptoms

In Table 1, we compare our method to prior work on measuring depressive symptom severity. One difference between our method and prior work is that our method does not rely on interview context. Prior work depends heavily on interview context such as the type of question asked [2, 14], whereas our method accepts a sentence without such metadata. While additional context typically helps the

		Classification: Major Depressive Disorder								Regression: PHQ Score			
		Sensitivity (TPR)				Specificity (TNR)				Average Error			
#	Method	A	V	L	AVL	A	V	L	AVL	A	V	L	AVL
1	Log-Mel	70.5	—	—	—	64.5	—	—	—	6.40	—	—	—
2	MFCC	74.3	—	—	—	67.7	—	—	—	5.96	—	—	—
3	3D Face [6]	—	71.4	—	—	—	69.4	—	—	—	5.82	—	—
4	W2V [29]	—	—	65.3	—	—	—	64.6	—	—	—	6.16	—
5	D2V [25]	—	—	67.7	—	—	—	71.4	—	—	—	5.81	—
6	USE [12]	—	—	63.4	—	—	—	62.5	—	—	—	6.27	—
7	LSTM [18]	53.9	61.0	57.5	74.2	54.6	59.7	60.3	76.1	7.15	6.12	6.57	5.18
8	C-CNN [5]	71.1	73.7	67.7	83.3	66.7	71.2	65.5	82.6	5.78	5.01	6.14	3.67

Table 2: **Ablation study.** Rows 1-2 are hand-crafted embeddings. Rows 3-6 are pre-trained embeddings. Rows 7-8 denote our learned sentence-level embeddings. Modalities: A: audio, V: visual, L: linguistic (text), AVL: combination. TPR and TNR denote true positive and negative rate, respectively. The input to 7-8 were sequences of log-mel spectrograms, 3D faces, and Word2Vecs.

model, it can introduce technical challenges such as having too few training examples per contextual class. Another difference is that our method uses *raw* input modalities: audio, visual, and text. Prior work uses engineered features such as min/max vocal pitch and word frequencies.

4.2 Ablation Study

In Table 2, rows 1-6 denote hand-crafted or pre-trained sentence-level embeddings. That is, the entire input sentence (audio, 3D facial scans, and transcript) is summarized into a single vector [29, 25, 12]. However, we propose to learn a sentence-level embedding from the input. These are shown in rows 7 and 8. It is important to note that our method *does* use hand-crafted and pre-trained *word*-level embeddings as input. However, internally, oure model learns a *sentence*-level embedding. Following prior work on sentence-level embeddings, rows 1-6 were computed via simple average [12]. To learn sentence-level embeddings, we evaluate: (i) long short-term memory [18] and (ii) causal convolutional networks [34, 5].

5 Discussion

Before adapting our work to future research, there are some points to consider. First, although a human was controlling the digital avatar, the data was collected from human-to-*computer* interviews and not human-to-*human*. Compared to a human interviewer, research has shown that patients report lower fear of disclosure and display more emotional intensity when conversing with an avatar [26]. Additionally, people experience psychological benefits from disclosing emotional experiences to chatbots [17]. Second, although it is commonly used in treatment settings and clinical trials, the symptom severity score (PHQ) is not the same as a formal diagnosis of depression. Our work is meant to augment existing clinical methods and not issue a formal diagnosis. Finally, while pre-existing embeddings are easy to use, recent research suggests these vectors may contain bias due to the underlying training data [11, 9, 13]. Mitigating bias is outside the scope of our work, but is crucial to providing culturally sensitive diagnosis and treatment.

Future work could better utilize longitudinal and temporal information such as depression scores across interview sessions that are weeks or months apart. Understanding *why* the model made certain predictions could also be valuable. Visualizations such as confidence maps over the 3D face and “usefulness" scores for audio segments could shed new insights.

In conclusion, we presented a multi-modal machine learning method which combines techniques from speech recognition, computer vision, and natural language processing. We hope this work will inspire others to build AI-based tools for understanding mental health disorders beyond depression.

Acknowledgements. This work was supported by a National Institutes of Health, National Center for Advancing Translational Science, Clinical and Translational Science Award (KL2TR001083 and UL1TR001085). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, 2016.

[2] T. Al Hanai, M. Ghassemi, and J. Glass. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, 2018.

[3] T. Althoff, K. Clark, and J. Leskovec. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 2016.

[4] A. P. Association et al. Diagnostic and statistical manual of mental disorders (dsm-5®), 2013.

[5] S. Bai, J. Z. Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv*, 2018.

[6] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *WACV*, 2016.

[7] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *Winter Conference on Applications of Computer Vision*, 2016.

[8] S. Bengio and G. Heigold. Word embeddings for speech recognition. In *Interspeech*, 2014.

[9] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*, 2016.

[10] H. Bourlard and N. Morgan. A continuous speech recognition system embedding mlp into hmm. In *NIPS*, 1990.

[11] A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 2017.

[12] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al. Universal sentence encoder. *arXiv*, 2018.

[13] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 2018.

[14] Y. Gong and C. Poellabauer. Topic modeling based multi-modal depression detection. In *Annual Workshop on Audio/Visual Emotion Challenge*, 2017.

[15] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, et al. The distress analysis interview corpus of human and computer interviews. In *LREC*. Citeseer, 2014.

[16] J. A. Hall, J. A. Harrigan, and R. Rosenthal. Nonverbal behavior in clinician—patient interaction. *Applied and Preventive Psychology*, 1995.

[17] A. Ho, J. Hancock, and A. S. Miner. Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *Journal of Communication*, 2018.

[18] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

[19] A. E. Kazdin and S. L. Blase. Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspectives on psychological science*, 2011.

[20] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush. Character-aware neural language models. In *AAAI*, 2016.

[21] K. Kroenke and R. L. Spitzer. The phq-9: a new depression diagnostic and severity measure. *Psychiatric Annals*, 2002.

[22] K. Kroenke, R. L. Spitzer, and J. B. Williams. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 2001.

[23] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 2009.

[24] W. Labov and M. Baranowski. 50 msec. *Language variation and change*, 2006.

[25] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, 2014.

[26] G. M. Lucas, J. Gratch, A. King, and L.-P. Morency. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 2014.

[27] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang. Depaudionet: An efficient deep model for audio based depression classification. In *International Workshop on Audio/Visual Emotion Challenge*, 2016.

[28] L. Manea, S. Gilbody, and D. McMillan. Optimal cut-off score for diagnosing depression with the patient health questionnaire (phq-9): a meta-analysis. *CMAJ*, 2012.

[29] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[30] J. Miller and M. Hardt. When recurrent models don't need to be recurrent. *arXiv*, 2018.

[31] A. S. Miner, A. Milstein, and J. T. Hancock. Talking to machines about personal mental health problems. *JAMA*, 2017.

[32] A. S. Miner, A. Milstein, S. Schueller, R. Hegde, C. Mangurian, and E. Linos. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Internal Medicine*, 2016.

[33] National Alliance on Mental Illness. Mental health facts infographics.

[34] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv*, 2016.

[35] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch, 2017.

[36] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[37] J. E. Perez and R. E. Riggio. Nonverbal social skills and psychopathology. *Nonverbal Behavior in Clinical Settings*, 2003.

[38] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *NAACL*, 2018.

[39] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[40] J. T. M. Schelde. Major depression: Behavioral markers of depression and recovery. *The Journal of Nervous and Mental Disease*, 1998.

[41] C. Sobin and H. A. Sackeim. Psychomotor symptoms of depression. *American Journal of Psychiatry*, 1997.

[42] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio. Char2wav: End-to-end speech synthesis. *ICLR*, 2017.

[43] J. Turian, L. Ratinov, and Y. Bengio. Word representations: a simple and general method for semi-supervised learning. In *ACL*, 2010.

[44] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *International Workshop on Audio/Visual Emotion Challenge*, 2016.

[45] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang. Phoneme recognition using time-delay neural networks. *Readings in speech recognition*, 1990.

[46] P. Waxer. Nonverbal cues for depression. *Journal of Abnormal Psychology*, 1974.

[47] J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T. F. Quatieri. Detecting depression using vocal, facial and semantic communication cues. In *International Workshop on Audio/Visual Emotion Challenge*, 2016.

[48] World Health Organization. Depression key facts, 2018.

A Appendix

A.1 Data Format

Full data details can be found on the original dataset website [7]. Audio was recorded with a head-mounted microphone at 16 kHz. Video was recorded at 30 frames per second with a Microsoft Kinect. A total of 68 three-dimensional facial keypoints were extracted using OpenFace [7]. Audio was transcribed by the dataset curators and segmented into sentences and phrases with millisecond-level timestamps [15]. We use the dataset's train-val split: train (107 patients), validation (35 patients). Note that while a test set exists, the labels are not public.

We canonicalized slang words present in the transcription. For example, *bout* was translated to *about*, *till* was translated to *until*, and *lookin* was translated to *looking*. All text was forced to lower case. Numbers were canonicalized as well (e.g., 24 was represented as *twenty four*).

A.2 Implementation Details

A.2.1 Experiment 1: Automatically Measuring the Severity of Depressive Symptoms

Input to “our method", i.e. Causal CNN are as follows:

- Audio: Log-mel spectrograms with 80 mel filters.
- Visual: 68 3D facial keypoints.
- Linguistic: Word2Vec embeddings [29].

The network architecture is a 10-layer causal convolutional network [5] with kernel size of 5 with 128 hidden nodes per layer. Dropout was applied to all non-linear layers with a 0.5 probability of being zeroed. The loss objectives were binary cross entropy for classification and mean squared error for regression. The model was optimized with the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ with L2 weight decay of 1e-4. The initial learning rate was 1e-3 for classification and 1e-5 for regression. A batch size of 16 was used. The model was trained on a single Nvidia V100 GPU for 100 epochs. Our model was implemented with Pytorch [35].

A.2.2 Experiment 2: Ablation Studies

For Table 2, the details for each row are as follows:

1. Log-mel spectrograms were computed with 80 mel filters.
2. Mel-frequency cepstral coefficients were computed with 13 resulting values.
3. A total of 68 three-dimensional facial keypoints were provided by the dataset [15]. They were extracted using OpenFace [6].
4. Word2Vec vectors were computed using the publicly available Word2Vec model from Google and the Gensim python library [39]. Each vector is of length 300.
5. Doc2Vec vectors were also computed using Gensim [39]. Each vector is of length 300.
6. Universal sentence embeddings were computed using Tensorflow [1] from the publicly available release. Each vector is of length 512.
7. The LSTM consists of 10 layers with 128 hidden units and is also optimized with the same batch size, optimizer, etc. as stated in Appendix A.2.1.
8. Our causal CNN model is the same as the one outlined in Appendix A.2.1.

Public code implementations[3] were used for the core network architecture components for both the LSTM and causal CNN.

[3] https://github.com/locuslab/TCN