Multimodal Mental Health Analysis

Rojin Bakhti
University of Southern California
bakhti@usc.edu

Sukavanan Nanjundan
University of Southern California
snanjund@usc.edu

PROBLEM DEFINITION

Depression is a serious mental health disorder that is becoming more and more prevalent among both adults and children around the world. According to the World Health Organization, currently, almost 280 million people suffer from depression. Depression detection has been an active area of research in the Machine Learning and Deep Learning community. In recent times data from social networking sites such as Facebook, Instagram, etc., have been used to develop models for detecting depression among users [1] [2]. Inspired by the AVEC, 2019 *Detecting Depression Sub Challenge*, [3], we propose a method to detect depression among patients using audio and text modalities and perform an analysis of the performance of different modalities. First, we propose to develop unimodal models using the audio and text input features and then develop multimodal models using the best unimodal models by applying different fusion techniques. Later, we propose to compare the performance improvement, if any, going from unimodal to multimodal (in this case bimodal) models. For the purpose of this project, we will be using the Extended Distress Analysis Interview Corpus (E – DAIC) [4].

LITERATURE REVIEW

Our work builds on the paper from Audio/Visual Emotion Challenge and Workshop (AVEC 2019) which is a competition aimed at developing methods for automatic audiovisual health and emotion analysis. This challenge contains data from the DAIC – Wizard-of-Oz corpus (DAIC-WOZ) which is gathered from clinical sessions of patients with an AI agent [3]. There have been other papers looking at different deep learning models performance on the depression prediction problem. Particularly, we looked at a paper by Muhammad et al that have done a comparative analysis of several deep neural networks' architectures for multimodal depression recognition [5]. The researchers found the LSTM models to be surpassing the CNN based architectures on DAIC dataset. For learning about different fusion techniques in this problem, we looked at the Handbook of Artificial Intelligence in Healthcare book which explained a recurrent model architecture for detecting depression severity. The book suggests extracting LLD's and facial features separately and then concatenating them in a decision strategy.

DATA

The Extended Distress Analysis Interview Corpus (E-DAIC) is an extended version of the DAIC – Wizard-of-Oz corpus (DAIC-WOZ) [6]. The E-DAIC dataset contains semi-clinical interviews, designed to diagnose mental health disorders such as anxiety, depression and Post-Traumatic Stress Disorder (PTSD). These interviews were collected as part of a large effort to create a computer agent that interviews people and identifies verbal and nonverbal indicators of mental illnesses.

The dataset includes audio and video recordings of the interviews, which was automatically transcribed using Google Cloud's speech recognition service. The data also includes extensive questionnaire responses from the participants. The interviews were conducted by an animated virtual interview agent called Ellie. In the WoZ setting the interview was conducted by a virtual agent which was conducted by a human from another room. In the AI interviews, the virtual agent conducts the interview fully autonomous using automated perception and behavior generation modules.

For the purpose of the Challenge, the E-DAIC dataset was partitioned into training, development, and test sets while preserving the overall speaker diversity – in terms of age, gender distribution, and the eight-item Patient Health Questionnaire (PHQ-8) scores – within the partitions. Whereas the training and development sets include a mix of WoZ and AI scenarios, the test set is solely constituted from the data collected by the autonomous AI. Details regarding the speaker distribution over the partitions are given in Figure 1 which is taken from [3].

Partition	# Subjects	Duration [h:min:s]
Training	163	43:30:20
Development	56	14:47:31
Test	56	14:52:42
All	**275**	**73:10:33**

Figure 1: Data Split [3]

For the purpose of this project, we only make use of the PHQ – 8 scores and the PHQ – 8 binary scores as the labels of the datapoints.

METHOD

Unimodal Models

Audio Modality

Initial data exploration showed audio recordings to have varying lengths. MFCC and eGeMAPS data were extracted using

OpenSmiles and summarized over a sliding window of 4 seconds length and hop size of 1 second and were provided as part of the dataset. Each row in the MFCC and eGeMAPS dataframe represented 0.01 second of the audio. Total of ~23 acoustic features were extracted for the eGeMaps and ~39 features were extracted for the MFCC's for each 0.01 second of the audio. To create baseline models, we cut down the audio features to the maximum sequence length of 20 minutes and padded the ones that had less than 20 minutes with zeros. We built a baseline model with 1 layer 64-d GRU followed by 1 64-d fully connected layer. We chose 0.2 dropout and batch size of 15. We started with preprocessing and feeding the eGeMAPS and MFCC features into the model. The CCC score and accuracy for the baseline model was at the random level for eGeMAPS and MFCC's. Finally, the output was extracted from a single output node. We used the batch size of 2 for the data points to fit in the memory. For each patient, the audio was downsampled to the number of sentences uttered by each patient which was calculated from the transcript data for each data point. For each sentence uttered, downsampling was achieved by averaging the rows in the audio that their frame times fell between the start and end time of the sentence utterance. The batches were padded to the largest sequence in the batch. Later, the MFCC and eGeMAPS features were concatenated together to form a new dataset with 62 features.

```
ege_GRU_regression(
  (dropout): Dropout(p=0, inplace=False)
  (gru): GRU(23, 1024, num_layers=2, batch_first=True)
  (linear): Linear(in_features=1024, out_features=512, bias=True)
  (fc): Linear(in_features=512, out_features=1, bias=True)
)
```

Figure 2: eGeMAPS features based model architecture

```
mfcc_GRU_regression(
  (dropout): Dropout(p=0, inplace=False)
  (gru): GRU(39, 2048, num_layers=2, batch_first=True)
  (linear): Linear(in_features=2048, out_features=1024, bias=True)
  (linear_2): Linear(in_features=1024, out_features=512, bias=True)
  (fc): Linear(in_features=512, out_features=1, bias=True)
)
```

Figure 3: mfcc features based model architecture

```
audio_combined_GRU_regression(
  (dropout): Dropout(p=0, inplace=False)
  (gru): GRU(62, 2048, num_layers=2, batch_first=True)
  (linear): Linear(in_features=2048, out_features=2048, bias=True)
  (linear_2): Linear(in_features=2048, out_features=1024, bias=True)
  (fc): Linear(in_features=1024, out_features=1, bias=True)
)
```

Figure 4: Combined audio features based model architecture

Text Modality

Our previous method was to convert the words from the interview transcripts into word vectors using the GloVe embeddings [7]. And, since the lengths of the transcripts are, on an average, above 512 words, we decided against using BERT [8] to generate contextual word embeddings as the max sequence length is just 512 tokens. Dividing the transcript text into chunks of lengths lesser than or equal to 512 tokens defeats the purpose of contextualized word embeddings generated through BERT. The above method did not yield good results and the CCC score (0.011) indicated that the predictions were almost random. Then, we decided to use the turn-wise sentences uttered by the patients as the features. So, we used pre-trained sentence-transformer BERT [9] to encode the sentences uttered by the patients into vectors. Hence, each sentence spoken by the patient during the interview becomes one timestep. The sentence vectors are used to train a model of fully – connected layers built on top of a Gated Recurrent Unit (GRU) [10].

```
sbert_GRU_regression(
  (gru): GRU(768, 1024, batch_first=True)
  (dropout): Dropout(p=0, inplace=False)
  (linear_1): Linear(in_features=1024, out_features=512, bias=True)
  (output): Linear(in_features=512, out_features=1, bias=True)
)
```

Figure 5: Text Modality Model Architecture

Fusion Models (Audio and Text Modalities)

Early Fusion

We first identify the best audio feature from the results of the three unimodal audio models. Then we concatenate the inputs of the best performing audio modality dataset with that of the text modality dataset to form a new dataset for early fusion. After trying out different architectures, it was found that the model architecture used for the text modality model gave the best results with just a ReLU layer added as final layer and a few modifications to other hyperparameters such as batch size and learning rate.

```
early_fusion_GRU_regression(
  (gru): GRU(791, 1024, batch_first=True)
  (dropout): Dropout(p=0, inplace=False)
  (linear_1): Linear(in_features=1024, out_features=512, bias=True)
  (output): Linear(in_features=512, out_features=1, bias=True)
  (relu): ReLU()
)
```

Figure 6: Early Fusion Model Architecture

Late Fusion

Since the CCC scores of audio modality model and text modality models were vastly different, with the text modality model producing much better results, it was imprudent to give equal weightage to the results of both models and just output the average of the results of both the models. Therefore, we decided to do a weighted average of the results and the weights were based on the CCC scores of the respective models. The formula used to calculate the weights is given below,

$$model\ weight = \frac{ccc\ score\ of\ model}{sum\ of\ ccc\ scores\ of\ both\ models}$$

	Unimodal Models					Fusion Models		
	Audio Modality			Text Modality				
	eGeMAPS	MFCC	eGeMAPS + MFCC	GloVe Embeddings	Sentence BERT	Early Fusion	Late Fusion	Hybrid Fusion
RMSE	6.499	6.343	6.831	6.647	**5.048**	6.708	5.081	**5.214**
CCC	0.148	0.146	0.134	0.011	**0.582**	0.148	0.528	**0.608**

Table 1: Results

Hybrid Fusion

In hybrid fusion, we concatenate the last hidden states from the GRUs of the best performing models, in this case eGeMAPS model for audio modality and the sentence BERT model for text modality, and use that as the input to a fully connected network which will then result in the PHQ-Score. This is a more sophisticated way of performing weighted late fusion.

```
hybrid_fusion_regression(
  (linear_1): Linear(in_features=2048, out_features=4096, bias=True)
  (output): Linear(in_features=4096, out_features=1, bias=True)
  (relu): ReLU()
)
```

Figure 7: Hybrid Fusion model architecture

RESULTS

All the models were trained on an AMD Ryzen 9 5900 HS machine with an RTX 3070 (8GB memory) graphics card.

The results mentioned in Table 1 are obtained on the test set and are rounded to the 3rd decimal point.

As shown in Table 1, among the unimodal models, the text-based sentence BERT model seems to be outperforming all other models by a huge margin, both in terms of RMSE score and CCC score. The RMSE score of the sentence BERT model is also the best overall. Among the audio models, the model trained with eGeMAPS features gives the best results when it comes to CCC score and the MFCC features based model gives the best RMSE score. But, when both the features are combined together, the model trained on that set of features performs worse than the other two models. The results all the audio modality models, though not random, are not significant enough. On the other hand, as mentioned earlier, the GloVe embeddings-based text modality model, provides results that are almost random.

In the case of fusion models, the hybrid fusion model produces the best results. The CCC score of the hybrid fusion model is the best among all the models (both unimodal and multimodal). The performance of early fusion model is comparable to that of the eGeMAPS feature based model. This might be due to the fact that the audio modality is acting as a bottleneck and reducing the overall performance of the combined audio and text features. Late fusion seems to be performing much better than the early fusion model in terms of CCC score but the performance is still bottlenecked by the audio modality as we can see that the CCC score is still less than that of the sentence BERT text modality-based model.

CONCLUSION

Between audio and text modalities, the text modality model, specifically sentence BERT model, performs the best indicating that the text features are better than the audio features when modeling the depression scores of the patients of the DAIC-WoZ dataset [6].

Among the audio features, the eGeMAPS feature produces the best results and when the audio features (eGeMAPS and MFCC) are combined together, the results are worse than when the features are used alone.

Combining the two modalities to build a multimodal fusion model definitely helps in producing better results in all but one case, the early fusion model. Both in the early fusion and late fusion (weighted late fusion) models, the audio modality acts as a bottleneck for the text modality and brings down the performance of the model, which is much more clearly visible in the case of early fusion model.

Hybrid fusion model, a more complex version of the weighted late fusion model, seems to overcome this bottleneck and produces the best results overall.

TEAM MEMBER CONTRIBUTIONS

Both team members were involved in data extraction and exploration. Rojin Bakhti specifically worked on the unimodal audio models and late fusion model. Sukavanan Nanjundan worked on the unimodal text models and on the early and hybrid fusion models. The reports and the presentations were done by both the team members.

REFERENCES

[1] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang and A. Ulhaq, "Depression detection from social network data using machine learning techniques," *Health Information Science and Systems*, 2018.

[2] A. G. Reece and C. M. Danforth, "Instagram photos reveal predictive markers of depression," *EPJ Data Science*, no. 6.1, 2017.

[3] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani and M. Pantic, "AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition," in *9th International on Audio/visual Emotion Challenge and Workshop*, 2019.

[4] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo and L.-P. Morency, "The Distress Analysis Interview Corpus of human and computer interviews," in *Ninth International Conference on Language Resources and Evaluation*, 2014.

[5] M. Muzammela, H. Salamb and A. Othmani, "End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis," *Computer Methods and Programs in Biomedicine*, vol. 211, 2021.

[6] R. Artstein, D. DeVault, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, A. S. Giota Stratou, D. Traum, R. Wood, Y. Xu, A. Rizzo and . L.-P. Morency, "SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support," in *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems (AAMAS '14)*, 2014.

[7] J. Pennington, R. Socher and C. D. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[8] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

[9] R. Nils and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," 2019. [Online]. Available: https://arxiv.org/abs/1908.10084.

[10] K. Cho, B. v. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," 2014. [Online]. Available: https://arxiv.org/abs/1406.1078.

[11] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," [Online]. Available: arXiv preprint arXiv:1908.10084 (2019).

[12] M. Muzammela, H. Salamb and A. Othmani, "End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis," *Computer Methods and Programs in Biomedicine*, vol. 211, 2021.