Topic Modeling Based Multi-modal Depression Detection

Yuan Gong
University of Notre Dame
ygong1@nd.edu

Christian Poellabauer
University of Notre Dame
cpoellab@cse.nd.edu

ABSTRACT
Major depressive disorder is a common mental disorder that affects almost 7% of the adult U.S. population. The 2017 Audio/Visual Emotion Challenge (AVEC) asks participants to build a model to predict depression levels based on the audio, video, and text of an interview ranging between 7-33 minutes. Since averaging features over the entire interview will lose most temporal information, how to discover, capture, and preserve useful temporal details for such a long interview are significant challenges. Therefore, we propose a novel topic modeling based approach to perform context-aware analysis of the recording. Our experiments show that the proposed approach outperforms context-unaware methods and the challenge baselines for all metrics.

KEYWORDS
Topic modeling; depression detection; multi-modal; emotion recognition; natural language processing

1 INTRODUCTION

1.1 Background
Major depressive disorder (MDD), also usually called depression, is one of the most common mood disorders, which is characterized by a persistent low mood. The study in [6] showed that men have a risk of 10-20% and women have a risk of 5-12% to develop MDD in their lifetime. Early and accurate detection of MDD will ensure that appropriate treatment and intervention options can be considered. Therefore, there is a strong need for a simple method to detect depression. In the 2017 Audio/Visual Emotion Challenge (AVEC) [17], the depression sub-challenge task requires participants to predict the depression level (i.e., the PHQ-8 score [10]) using audio, video, and text analysis. The database used in this challenge is the distress analysis interview corpus (DAIC-WOZ) [7], [5], which includes data from 189 subjects. For each subject, the database includes the audio/video features as well as the transcript of the recording. Our experiments show that the proposed approach outperforms context-unaware methods and the challenge baselines for all metrics.

Figure 1: Illustration of the proposed topic modeling based multi-modal feature vector building scheme.

1.2 Challenges and Contribution
A big difference between the depression detection task and a traditional emotion detection task is the decision unit. Since human emotion can change rather quickly, traditional emotion detection typically requires second-level prediction. Therefore, popular emotion recognition databases usually provide labels for short-term recordings, e.g., the IEMOCAP database [3] provides labels for each utterance, while the SEWA database provides labels for each segment of 100ms. In contrast, depression is expressed through a persistently low mood, which is very different from short-term sadness. The study in [18] shows that the median duration of depression is three months and consequently, prediction of the depression level of an individual should be based on much longer observation periods. This difference between depression and emotion detection leads to two main challenges:

(1) Large decision unit. In the DAIC-WOZ database, each data sample is the audio and video recording of an interview of a specific subject, where the interview ranges from 7 to 33 minutes. Here, only one decision needs to be made for the entire interview. The length of the decision unit is much longer than for typical emotion recognition tasks,
e.g., the 2017 AVEC emotion sub-challenge requires making decisions for each 100ms segment. While a large data volume is typically beneficial for the accuracy, processing large amounts of data can be challenging. When analyzing very long audio/video data, applying statistical functions (e.g., max, min, mean, quartiles) to short-term features over the entire interview will lead to loss of detailed temporal information such as short-term sighs in despair, laughing, or anger. However, these short-term details within the interview can be useful when determining the depression level of the subject, especially when analyzed together with contextual information (e.g., sighing in despair when being asked about sleep quality, laughing when talking about a journey, and anger when remembering unhappy experiences). Therefore, it is important to map the whole interview to a feature vector such that short-term details and context are maintained.

(2) **Limited number of samples.** Since each subject has one sample (the entire audio/video recording), the number of samples is significantly lower than in the case when each recording consists of many small samples (e.g., each utterance being a sample). In the 2017 AVEC depression sub-challenge, the number of samples in the training set is only 107. In addition, the database is unevenly distributed, i.e., the number of depression samples in the training set is 30. With such a small sample size, the number of features should also be small to avoid the problems of dimensionality and overfitting. However, the dimensions of audio and video features are very large and therefore, generating and selecting an appropriate number of discriminative features is essential.

In order to overcome these two challenges, we propose a topic modeling based multi-modal feature vector building scheme as shown in Figure 1 to provide the basis for context-aware analysis. The interview is first segmented according to topics. Then, audio, video, and semantic features are generated for each topic segment separately and further placed into a separate slot of the topic in the feature vector. After the features for all topics have been placed, a two-step feature selection algorithm is executed to shrink the feature vector and only keep the most discriminative features. The proposed algorithm is inspired by the observation that all interviews contain not a fixed, but a limited range of topics. Further, we assume that each question by Ellie triggers a response on a new topic, which makes “topic tracking” feasible. We expect the following advantages from the proposed scheme:

(1) **Logically organize short-term details based on context.** When retaining the short-term details of the interview, we need to do it in a fashion that keeps the feature vector space relatively small and also makes it logical. Extracting details according to utterances is not context-organizable and will lead to a dimension explosion since each interview contains hundreds of utterances. For example, both subject 1 and subject 2 smile at utterance 10, but their smiling might convey different information since their 10th utterance is in different contexts. In contrast, the proposed scheme tracks the topic and places feature of utterances, no matter where it is in the interview, into the slot of the topic it belongs to in the feature vector. In addition, one topic can cover multiple utterances, which makes the feature dimension much smaller.

(2) **More flexible and precise discovery of useful features.** In traditional feature building schemes, one feature can only be kept or discarded as a whole. However, it is common that one feature is only useful in some specific contexts and useless in others. Further, same features in different contexts might convey different information and should be regarded as separate features. For example, smiling in the context of discussing family can be more discriminative than smiling in the context of greeting someone, because the latter might only be due to etiquette. Therefore, we would like to only keep the feature when it is in a useful context. The proposed feature building scheme provides any combination of features and contexts such as smiling (family) and smiling (greeting). Thus, the feature selection algorithm can perform a more flexible and accurate filtering. In summary, the proposed scheme allows us to perform a finer analysis of the subject’s reaction to a specific topic, such as a lower voice when discussing family, irritation when discussing an unhappy situation, and the expressions used when describing recent emotions. We believe that this finer-grained analysis can improve the performance of depression detection.

### 1.3 Related Work

In the 2016 AVEC depression classification sub-challenge [19], a few proposed techniques adopted text analysis for their model building. In [14] and [13], the text is analyzed on a subject level and audio/video features are separately extracted and then fused with semantic features, i.e., topic modeling is not used in these approaches. In [21], the authors conduct a question/answer extraction (which is similar to topic extraction) before text analysis. However, the question/answer extraction is only applied to text analysis. Audio and video analysis is still conducted separately. In [22], the authors also conduct topic extraction, but merely use the semantic features of very few topics (3 topics for women and 4 topics for men) to build a simple decision tree. This approach achieved the best performance in the 2016 AVEC, which demonstrates the effectiveness of simple model and key topic analysis. However, its performance on the test set is much worse than that on the development set. Its limited ability to generalize is probably due to the very small number of features. Further, audio and video features are not used in this work.

On the other hand, topic modeling, which is a technique to discover topics from documents, has been widely adopted in applications such as text mining [9] and recommendation systems [20]. It recently has also been used for depression and neuroticism assessment [16]. In [16], the authors demonstrate that taking automatically derived topics into account improves prediction performance. However, audio and video analysis are again not involved in this work.

In summary, the work in [21], [22], and [16] use topic modeling, but only for text analysis. We further extend the application
of topic modeling by using it for context-aware audio and video analysis. To the best of our knowledge, the proposed work is the first effort to combine topic modeling with multi-modal text, audio, and video analysis.

2 TOPIC MODELLING BASED MULTI-MODAL DEPRESSION DETECTION

2.1 Topic Modeling

Topic modeling typically requires a sophisticated algorithm such as latent Dirichlet allocation (LDA) [2] and network regularization [12]. However, for the transcription of clinical interviews (such as provided by the DAIC-WOZ database), topic modeling can be done much simpler for multiple reasons. First, in the interview, only Ellie determines the topic by asking a question and the subject does not initiate a topic proactively. Second, the number of topics in clinical interviews is limited. And third, Ellie is an animated interviewer controlled by human command and, therefore, has a relatively fixed way to start a topic. We observed that when starting a specific topic, Ellie chooses one sentence from the library, which typically consists of only 1-3 fixed sentences per topic.

Based on the above-mentioned observations, we perform simple topic modeling on the text of interviews. First, we build a preliminary sentence dictionary by traversing all of Ellie’s speech and record all non-redundant sentences. Then, we perform manual cleaning of the preliminary dictionary, where sentences that do not start new topics (e.g., “that’s good”) are discarded. After that, we perform clustering of the dictionary, where the sentences that start the same topic are grouped together. This is done in two steps. First, very similar sentences with up to 3 characters difference are clustered automatically. Second, further manual clustering and checks are performed. Then, we review each sentence cluster, link each cluster to the corresponding topic, and put it into the topic dictionary. Therefore, the topic dictionary is formatted as [topic name, corresponding Ellie sentences]. The complete list with 83 extracted topics is shown in Table 1.

Note that only a few topics are discussed in most interviews, e.g., only 14 topics cover over 80% of the interviews. In other words, topics are sparsely distributed in the interviews. The histogram of the topic cover rate is shown in Figure 2.

2.2 Features

In this work, we use audio, video, and semantic features to build a multi-modal model. Audio and video features are provided by the 2017 AVEC organizers while semantic features are extracted by ourselves. All features are computed in a topic-wise fashion.

2.2.1 Audio Features. We use the audio features extracted by the COVAREP toolkit [4] and formant features. The COVAREP toolkit generates a 74-dimensional feature vector that includes common features such as fundamental frequency and peak slope. Formant features contain the first 5 formants, i.e., the vocal tract resonance frequencies. Both COVAREP and formant features are extracted every 10ms. For each topic, we further apply three statistic functions (mean, max, and min) to each feature over time to reduce the dimension. That is, for each topic, 

\[(74 + 5) \times 3 = 237\]

audio features are used.

2.2.2 Video Features. We use the action units (AUs) features extracted by the OpenFace toolkit [1], which includes the information of 20 key AUs. For each topic, we further apply three statistic functions (mean, max, and min) over time to each feature to reduce the dimension. Thus, for each topic, 

\[20 \times 3 = 60\]

video features are used.

2.2.3 Semantic Features. For each of the 83 topics, we use the Linguistic Inquiry and Word Count (LIWC) [15] software to count the frequency of word presence of the subject’s speech of the topic in 93 categories such as anger, negative emotion, and positive emotion. That is, the LIWC software takes the speech of a subject and generates a 93-dimensional feature vector. Further, inspired by [22], which demonstrates that some key topics such as sleep quality (topic index: 78) and PTSD diagnose history (topic index: 82) have a high correlation with depression level, we further extract additional semantic features for 8 topics (topic index: 76-83, marked with an asterisk in Table 1) that we believe might be most discriminative. We use a dictionary based method to classify each topic into 2 or 3 categories according to the content. For example, for the topic easy (sleep (topic index: 78), the speech of each subject is classified into three categories: easy (when phrases such as ‘no problem’ are present), fair (when words such as ‘it depends’ are present), and hard (when words such as ‘difficult’ are present). The dictionary is built manually for each key topic.

![Figure 2: Histogram of topic cover rate.](image)

Figure 3: Illustration of the structure of feature vector.

2.3 Topic-wise Feature Mapping

In order to conduct context-aware analysis, the feature vector needs to record the features of each topic separately. Therefore, in the feature vector, each topic has a separate slot.

We first find the topics discussed in each interview. For each interview, speech sentences of Ellie are traversed and when the sentence is found in the topic dictionary, the corresponding topic
Table 1: The list of topics extracted from the DAIC-WOZ (potential key topics are marked with an asterisk).

| Ind. Topic Abbr. | Sample Ellie Question | Ind. Topic Abbr. | Sample Ellie Question |
|-----------------|-----------------------|------------------|----------------------|
| 1 more           | can you tell me about that | 43 best_parent  | what’s the best thing about being a parent |
| 2 why            | why                   | 44 are_you_okay  | are you okay with this |
| 3 last_happy_time| tell me about the last time you felt really happy | 45 mad           | what are some things that make you really mad |
| 4 origin         | where are you originally | 46 they_triggered | are they triggered by something |
| 5 argue          | when was the last time you argued with someone | 47 easy_parent   | do you find it easy to be a parent |
| 6 advice_ago     | what advice would you give to yourself ten or twenty years ago | 48 happy_did_that | are you happy you did that |
| 7 control_anger  | how are you at controlling your temper | 49 therapist_affected | how has seeing a therapist affected you |
| 8 things_like_La | what are some things you really like about La | 50 job           | what are you |
| 9 proud          | what are you most proud of in your life | 51 symptoms      | what were your symptoms |
| 10 positive_influence | who’s someone that’s been a positive influence in your life | 52 ideal_weekend | tell me how you spend your ideal weekend |
| 11 best_friend_describe | how would your best friend describe you | 53 avoid         | could you have done anything to avoid it |
| 12 things_dont_like_La | what are some things you don’t really like about La | 54 do_annoyed    | what do you do when you are annoyed |
| 13 major | what did you study at school | 55 got_in_trouble | has that gotten you in trouble |
| 14 regret | is there anything you regret | 56 your_kid      | tell me about your kids |
| 15 dream_job     | what’s your dream job | 57 someone_made_bad | tell me about a time when someone made you feel really badly about yourself |
| 16 enjoy_travel  | what do you enjoy about traveling | 58 different_parent | what are some ways that you’re different as a parent than your parents |
| 17 how_hard      | how hard is that | 59 today_kid     | what do you think of today’s kids |
| 18 do_sleep_not_well | what are you like when you don’t sleep well | 60 down         | do you feel down |
| 19 experiences   | what’s one of your most memorable experiences | 61 how_know_them | how do you know them |
| 20 hardest_decision | tell me about the hardest decision you’ve ever had to make | 62 feel_fallen   | do you feel that way often |
| 21 fun_relax      | what are some things you like to do for fun | 63 problem_before | did you think you had a problem before you found out |
| 22 handle_differently | tell me about a situation that you wish you had handled differently | 64 living_situation | how do you like your living situation |
| 23 what_decide    | what made you decide to do that | 65 why_stop      | why did you stop |
| 24 still_work     | are you still doing that | 66 how_you_do    | how are you doing today |
| 25 erase_memory   | tell me about an event or something that you wish you could erase from your memory | 67 roommate     | do you have roommates |
| 26 why_move_to_La | why did you move to La | 68 hard_on_yourself | do you think that maybe you’re being a little hard on yourself |
| 27 change_self    | what are some things you wish you could change about yourself | 69 like_living_with | what’s it like for you living with them |
| 28 best_quality  | what would you say are some of your best qualities | 70 disturb_thought | do you have disturbing thoughts |
| 29 often_back    | how often do you go back to your home town | 71 where_live    | where do you live |
| 30 how_long_diagnose | how long ago were you diagnosed | 72 after_military | what did you do after the military |
| 31 guilty        | what’s something you feel guilty about | 73 combat        | did you ever see combat |
| 32 when_move_to_La | when did you move to La | 74 talk_later    | why don’t we talk about that later |
| 33 easy_used_to_La | how easy was it for you to get used to living in La | 75 military_change | how did serving in the military change you |
| 34 seek_help      | what got you to seek help | 76 change_behavior | have you noticed any changes in your behavior or thoughts lately |
| 35 when_last_happy | when was the last time you felt really happy | 77* depression  | have you been diagnosed with depression |
| 36 cope          | how do you cope with them | 78* easy_sleep  | how easy is it for you to get a good night sleep |
| 37 compare_to_La  | how does it compare to La | 79* family_close | how close are you to your family |
| 38 hard_parent   | what’s the hardest thing about being a parent | 80* feeling_lately | how have you been feeling lately |
| 39 still_therapy | do you still go to therapy now | 81* shy_outgoing | do you consider yourself an introvert |
| 40 travel_La     | do you travel a lot | 82* ptsd        | have you ever been diagnosed with PTSD |
| 41 ever_served_military | have you ever served in the military | 83* therapy_useful | do you feel like therapy is useful |
Table 2: Dimension of each feature category.

| Feature Name | Dimension |
|--------------|-----------|
| Gender       | 1         |
| Topic Presence | 83       |
| Key Topic    | 8         |
| LIWC         | 7719      |
| Formant      | 1245      |
| COVAREP      | 18426     |
| AUs          | 4980      |
| Sum          | 32462     |

and the subject’s speech, together with its timestamps are recorded. The subject’s speech is used to generate semantic features while the timestamps are used to synchronize audio and video features. Then, all features are placed into separate slots of the corresponding topic in the feature vector. As described in Section 2.2, each topic contains 237 audio features, 60 video features, and 93 LIWC features, and there are 83 topics in total, which leads to a $83 \times (237 + 60 + 93) = 32,370$ dimensional feature vector. Further, we add the presence of each topic to the feature vector, because each interview only covers a few topics and the topic presence might be correlated to the subject’s status. Finally, gender is also attached to the feature vector similar to the work in [22] and [14], where the authors report that gender information can greatly improve the classification performance. Figure 3 illustrates the structure of the feature vector and Table 2 shows the dimension of each feature category in the feature vector. Due to the sparsity of topics, the feature vector is also sparse, i.e., the features of topics that are not discussed in an interview are missing. However, the slots for all topics are preserved in our approach, i.e., the slot of a topic that is not discussed in the interview is padded with -1.

2.4 Feature Selection

In Section 2.3, a 32,462-dimensional feature vector is built, which maintains audio, video, and text information of each topic. However, only a small amount of features are actually useful and we expect the number of features to be small enough to avoid potential overfitting. Therefore, feature selection is an essential step of the proposed scheme.

We conduct feature selection in two steps. First, we conduct a quick model-independent feature selection on all the features. The algorithm we use in this step is correlation-based feature subset selection (CFS) [8], which evaluates the value of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. After this step, a subset of features is selected. Then, we conduct a fine model-dependent feature selection to find the optimized feature number. In this step, we first rank the features according to their F-value to the corresponding label. Then we run the regression algorithm using a various number of high-rank features and finally select the best feature set.

This unique two-step feature selection algorithm is designed based on the following consideration. In our feature generation scheme, we observe that more features are correlated to each other than with the context-unaware feature generation scheme, because features that belong to the same topic are likely to have high correlations. Thus, if we only conduct feature selection according to the individual feature score, we might get a set of features with high scores, but that are also closely correlated to each other. In other words, many features are redundant and provide little extra information in this case. To avoid that, we first conduct a CFS to select a feature subset, where features have a high correlation with the label, but low correlation with each other. Since CFS is a model-independent approach, which cannot tell us the overfitting risk for our specific model and dataset, we further conduct a model-based selection on our dataset to find the appropriate number of features for our task.

2.5 Regression Model Building

2.5.1 Data Balancing. It has been widely reported that imbalanced classes of data will greatly affect the performance of machine learning algorithms [11]. Unfortunately, most healthcare related databases, including DAIC-WOZ, are imbalanced. In the training set of the DAIC-WOZ database, only 30 subjects are depressed of a total of 107 subjects, which means that there are more subjects with low PHQ-8 scores than those with high PHQ-8 scores. Therefore, we perform random-oversampling to make the number of samples for each PHQ-8 score is roughly the same by simply duplicating samples before running the machine learning algorithm.

2.5.2 Regressors. In this work, we perform a grid search for the following regression models: random forest regression (number of trees: 1, 10, 20, 30, 40, 50, 100, and 200), stochastic gradient descent (SGD) regression, and support vector regression (SVR) (kernel: linear, polynomial, and radial basis function (RBF)).
We compare the proposed method with the following baseline methods:

3.3 Baseline
is depressed when the PHQ-8 score is greater than or equal to 10.

3.2 Metrics
We report four metrics for each test strategy mentioned above: 1) Root mean square error (RMSE) is the challenge target; therefore, all optimizations, including model selection and feature selection, are performed according to this metric. 2) Mean absolute error (MAE) is another metric reported by the official baseline [17] and we use it together with RMSE to analyze the difference between ground truth and prediction. 3) Pearson correlation coefficient (CC) is an important metric to evaluate the regression performance, which can reflect the linear correlation between ground truth and prediction. 4) F1-score measures the performance of binary depression classification, i.e., a subject is depressed when the PHQ-8 score is greater than or equal to 10 and non-depressed otherwise.

3.3 Baseline
We compare the proposed method with the following baseline methods:

1) Basic Baseline, where the model constantly predicts the mean PHQ-8 score of the training set. This is a very basic baseline that any workable regression algorithm should outperform.

2) Challenge Baseline, which is the official baseline provided in [17]. This baseline uses a random forest regressor (number of trees = 10) on the audio and video features extracted by the COVAREP toolkit [4] and OpenFace toolkit [1]. Regression is performed on a frame-wise basis and the temporal fusion over the interview is compressed by averaging the outputs over the entire interview. Fusion of audio and video modalities is performed by averaging the regression outputs of the unimodal result. In [17], the authors present the results of audio unimodal, video unimodal, and audio/video multimodal solutions using this baseline approach, where video unimodal has the best performance. Therefore, we use the results of the video unimodal solution for comparison.

3) Context-unaware Baseline Since the proposed method and the official challenge baseline method have a lot of differences in terms of features, regression model, and class balancing, it is hard to judge which factor cause any performance gap. Therefore, we use this baseline to check the effectiveness of context-aware analysis. This baseline method is exactly the same as the proposed method (i.e., the same audio, video, and LIWC features are extracted, the same feature selection algorithms are used, and the regression model is selected from the same grid), except topic modeling is not used. The differences are that features are extracted and averaged over the entire interview (instead of a topic-wise manner) and that topic related features (topic presence and key topic features) are not included.

(4) Proposed Method, as described in Section 2.

4 EVALUATION AND DISCUSSION

4.1 Overall Performance
Through a grid search in the CV test, we selected the best regression model (SGD regressor) and the best number of features (46). We then use these settings on the Dev and Test experiments. The results of these experiments are shown in Table 3. We observe that the proposed method achieves the best performance for all metrics and test strategies. Further, we find that the proposed method performs significantly better than the context-unaware baseline, which demonstrates the effectiveness of context-aware analysis. In addition, we observe that the performance of the proposed method on the test set is worse than that on the development set and cross validation. This is because the model-independent CFS feature selection is conducted not in a cross-validation manner, but instead on the entire training and development set, since it is meaningless to conduct CFS in a cross-validation manner. However, the performance of the proposed method is still much higher than the challenge baseline on the test data.

4.2 Selected Feature Analysis

Figure 4: Distribution of topics corresponding to the selected features (count in parentheses). Left: proposed feature selection algorithm. Right: baseline feature selection algorithm.

It is very interesting to see which features are actually selected and useful in depression detection. In our feature building scheme, each feature corresponds to one topic and one feature category. As shown in Figure 4 (left), from the perspective of the topics involved, we observe that 31 topics out of the total 83 topics are involved, in which the most frequent topics corresponding to the
Table 3: Result of the depression regression experiment

|                      | RMSE       | MAE        | CC         | F1-Score   |
|----------------------|------------|------------|------------|------------|
|                      | CV | Dev | Test | CV | Dev | Test | CV | Dev | Test | CV | Dev | Test |
| Basic baseline       | 5.84 | 6.57 | /    | 4.81 | 5.50 | /    | -0.35 | 0.00 | / | 0.00 | 0.00 | / |
| Challenge baseline   | /   | 7.13 | 6.97 | /   | 5.88 | 6.12 | /    | /    | / | /    | /    | / |
| Context-unaware baseline | 5.55   | 5.02 | /    | 4.56 | 4.42 | /    | 0.45 | 0.69 | / | 0.58 | 0.67 | / |
| Proposed method      | 3.68 | 3.54 | 4.99 | 2.94 | 2.77 | 3.96 | 0.78 | 0.87 | / | 0.80 | 0.70 | 0.60 |

Selected features are topic 30: how_long_diagnose, 31: guilty, 34: seek_help, and 77: depression. Further, we observe that our approach uses a variety of topics that seem not closely related to depression from humans’ perspective such as topic 6: advice Ago and 16: enjoy_travel. In addition, to check the effectiveness of the proposed two-step feature selection algorithm, we compare it with a baseline feature selection algorithm that only consists of step 2 of the proposed method, which only considers the score of each feature individually. As shown in Figure 4 (right), the feature vector selected by the baseline feature selection algorithm only includes three topics: 30: how_long_diagnose, 34: seek_help, and 39: still_therapy. We conduct an experiment using this feature vector and find that the result (RMSE: 5.60) is much worse than the result of the proposed approach (RMSE: 4.99) on the test set. This demonstrates that the proposed two-step feature selection algorithm is able to discover independent features and to improve the result. While it is possible that topics 30, 34, and 39 are the closest related topics of depression, taking more topics into consideration can lead to a more precise prediction. We believe that it is also an advantage of the proposed method over a clinician’s analysis, because for a clinician it is very hard to observe and model such a large volume of factors in the interview.

From the perspective of feature categories involved, we observe that the selected feature set involves LIWC features, key topic semantic features, COVAREP audio features, and AUs video features. The two key topics involved are 78: easy_sleep and 80: feeling_lately.

The gender feature, topic presence feature, and the formant features are not involved. A complete pie chart of the distribution of feature categories corresponding to the selected feature set is shown in Figure 5.

4.3 Regression Model and Feature Number Analysis

![Figure 6: The relationship between RMSE and number of features and regression models.](image)

![Figure 7: The relationship between CC and number of features and regression models.](image)

Due to the limited number of test attempts allowed in the 2017 AVEC, we are not able to provide the results on the test set for the baseline approaches. The challenge baseline paper [17] does not include test results of CC and F1-score and does also not include results tested in CV manner.

1
Two important hyperparameters in the proposed method are the number of features and the regression model. Thus, we performed a grid search in cross validation manner for the following regression models: random forest regression (RF) (number of trees: 1, 10, 20, 30, 40, 50, 100, 200), SGD regression, and support vector regression (SVR) (kernel: linear, polynomial, and RBF), and the following feature numbers: 1-46 (the total number of features in the subset selected by the first round CFS feature selection is 46). The relationship between regression performance and hyperparameters is shown in Figures 6 and 7. For clarity, we only plot the top 3 regression models with the best performance.

We observe that when the feature number is small, the random forest regressor (tree number = 40), SGD regressor, and SVR (RBF) regressor perform similarly. However, with the increase of feature numbers, SGD and SVR models continually improve their performance while the random forest model stops improving much earlier. The SGD and SVR regressor have close performance, while the SGD regressor has a little bit lower RMSE than SVR. Though the lowest RMSE is achieved when the feature number is 41, we believe it is more likely to be a fluctuation in the CV test and therefore we choose the feature number of 46, because we prefer to use more features to build a more discriminative model. The experiment shows that using 46 features (RMSE: 4.99) yields a better performance than using 41 features (RMSE: 5.22) on the test set.

5 CONCLUSIONS

Major depressive disorder is a widespread mental disorder and accurate detection will be essential for targeted intervention and treatment. In this challenge, participants are asked to build a model predicting the depression levels based on the audio, video, and text of an interview ranging between 7-33 minutes. Since averaging features over the entire interview will lose most temporal details, how to discover, capture, and preserve important temporal details for such long interviews are significant challenges. Therefore, we propose a novel topic modeling based approach to perform context-aware analysis. Our experiments show that the proposed approach performs significantly better than context-unaware method and the challenge baseline for all metrics. In addition, by analyzing the features selected by the machine learning algorithm, we found that our approach has the ability to discover a variety of temporal features that have underlying relationship with depression and further to build model on them, which is a task that is difficult to perform by humans.

REFERENCES

[1] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE, 1–10.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. Journal of machine learning research 3, Jan (2003), 993–1022.

[3] Carlos Bussu, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungjok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation 42, 4 (2008), 335.

[4] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raatto, and Stefan Scherer. 2014. COVAREF: A collaborative voice analysis repository for speech technologies. In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 960–964.

[5] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. Sim Sensei Kiosk: A virtual human interviewer for healthcare decision support.

In Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems. International Foundation for Autonomous Agents and Multiagent Systems, 1061–1068.

[6] Maurizio Fava and Kenneth S Kendler. 2000. Major depressive disorder. Neuron 28, 2 (2000), 335–341.

[7] Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The Distress Analysis Interview Corpus of human and computer interviews. In LREC. 3123–3128.

[8] MA Hall. 1998. Correlation-based feature subset selection for machine learning. Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato (1998).

[9] Liangjie Hong and Brian D Davison. 2010. Empirical study of topic modeling with network regularization. In Proceedings of the 17th international conference on World Wide Web. ACM, 80–88.

[10] Kurt Kroenke, Tara W Strine, Robert L Williams, Joyce T Berry, and Ali H Mokdad. 2009. The PHQ-8 as a measure of current depression in the general population. Journal of affective disorders 114, 1 (2009), 163–173.

[11] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. 2009. Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39, 2 (2009), 539–550.

[12] Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. 2008. Topic modeling with network regularization. In Proceedings of the 17th international conference on World Wide Web. ACM, 101–110.

[13] Md Nasir, Arindam Jati, Prashanth Gurunath Shivakumar, Sandeep Nallan Chakravarthula, and Panayiotis Georgiou. 2016. Multimodal and multiresolution depression detection from speech and facial landmark features. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, ACM, 43–50.

[14] Anastasia Pampoukhoudou, Olympia Simaniotaki, Amir Fazaeli, Matthew Pedditakis, Dimitris Manousos, Alexandros Bonitiotis, Georgios Giannakakis, Fabrice Meriaudeau, Panayiotis Simos, Kostas Marias, et al. 2016. Depression Assessment by Fusing High and Low Level Features from Audio, Video, and Text. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, ACM, 27–34.

[15] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LWCW/2015. Technical Report.

[16] Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of necroticism and depression. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language. Association for Computational Linguistics, 1348–1353.

[17] Fabien Ringeval, Bjorn Schuller, Michael Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummmins, Maximilian Schmitt, and Maja Pantic. 2017. AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge. In Proceedings of the 7th International Workshop on Audio/Visual Emotion Challenge. ACM, 1–8.

[18] JAN Spijker, Ron De Graaf, Rob V Bijl, Aartjan TF Beekman, Johan Ormel, and Willem A Nolen. 2002. Duration of major depressive episodes in the general population: results from The Netherlands Mental Health Survey and Incidence Study (NEMESIS). The British journal of psychiatry 181, 3 (2002), 208–213.

[19] Michel Valstar, Jonathan Gratch, Bjorn Schuller, Fabien Ringeval, Dennis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. ACM, 3–10.

[20] Chong Wang and David M Blei. 2011. Collaborative topic modeling for recommending scientific articles. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 448–456.

[21] James R Williamson, Elizabeth Godoy, Miriam Cha, Adrianne Schwarzenbru- ber, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie Dagli, and Thomas F Quatieri. 2016. Detecting Depression using Vocal, Facial and Semantic Communication Cues. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. ACM, 11–18.

[22] Le Yang, Dongmei Jiang, Lang He, Ercheng Pei, Masha Czrdic Ove neke, and Hichem Sahli. 2016. Decision Tree Based Depression Classification from Audio Video and Language Information. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. ACM, 89–96.