Predicting Depression in Screening Interviews from Latent Categorization of Interview Prompts

Alex Rinaldi
Department of Computer Science
UC Santa Cruz
arinaldi@ucsc.edu

Jean E. Fox Tree
Department of Psychology
UC Santa Cruz
foxtree@ucsc.edu

Snigdha Chaturvedi
Department of Computer Science
University of North Carolina
at Chapel Hill
snigdha@cs.unc.edu

Abstract

Despite the pervasiveness of clinical depression in modern society, professional help remains highly stigmatized, inaccessible, and expensive. Accurately diagnosing depression is difficult– requiring time-intensive interviews, assessments, and analysis. Hence, automated methods that can assess linguistic patterns in these interviews could help psychiatric professionals make faster, more informed decisions about diagnosis. We propose *JLPC*, a method that analyzes interview transcripts to identify depression while jointly categorizing interview prompts into latent categories. This latent categorization allows the model to identify high-level conversational contexts that influence patterns of language in depressed individuals. We show that the proposed model not only outperforms competitive baselines, but that its latent prompt categories provide psycholinguistic insights about depression.

1 Introduction

Depression is a dangerous disease that effects many. A 2017 study by Weinberger et al. (2018) finds that one in five US adults experienced depression symptoms in their lifetime. Weinberger et al. also identify depression as a significant risk factor for suicidal behavior.

Unfortunately, professional help for depression is not only stigmatized, but also expensive, time-consuming and inaccessible to a large population. Lakhan et al. (2010) explain that there are no laboratory tests for diagnosing psychiatric disorders; instead these disorders must be identified through screening interviews of potential patients that require time-intensive analysis by medical experts. This has motivated developing automated depression detection systems that can provide confidential, inexpensive and timely preliminary triaging that can help individuals in seeking help from medical experts. Such systems can help psychiatric professionals by analyzing interviewees for predictive behavioral indicators that could serve as additional evidence (DeVault et al., 2014).

Language is a well-studied behavioral indicator for depression. Psycholinguistic studies by Segrin (1990), Rude et al. (2004), and Andreasen (1976) identify patterns of language in depressed individuals, such as focus on self and detachment from community.

To capitalize on this source of information, recent work has proposed deep learning models that leverage linguistic features to identify depressed individuals (Mallol-Ragolta et al., 2019). Such deep learning models achieve high performance by uncovering complex, unobservable patterns in data at the cost of transparency.

However, in the sensitive problem domain of diagnosing psychiatric disorders, a model should offer insight about its functionality in order for it to be useful as a clinical support tool. One way for a model to do this is utilizing the structure of the input (interview transcript) to identify patterns of conversational contexts that can help professionals in understanding how the model behaves in different contexts.

A typical interview is structured as pairs of prompts and responses such that participant *responses* follow interviewer *prompts* (such as “How have you been feeling lately?”). Intuitively, each interviewer prompt serves as a context that informs how its response should be analyzed. For example, a short response like “yeah” could communicate agreement in response to a question such as “Are you happy you did that?”, but the same response could signal taciturnity or withdrawal (indicators of depression) in response to an encouraging prompt like “Nice!”. To enable such context-dependent analysis, the model should be able to group prompts based on the types of conversa-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7–18
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

tional context they provide.

To accomplish this, we propose a neural *Joint Latent Prompt Categorization* (*JLPC*) model that infers latent prompt categories. Depending on a prompt's category, the model has the flexibility to focus on different signals for depression in the corresponding response. This prompt categorization is learned jointly with the end task of depression prediction.

Beyond improving prediction accuracy, the latent prompt categorization makes the proposed model more transparent and offers insight for expert analysis. To demonstrate this, we analyze learned prompt categories based on existing psycholinguistic research. We also test existing hypotheses about depressed language with respect to these prompt categories. This not only offers a window into the model's working, but also can be used to design better clinical support tools that analyze linguistic cues in light of the interviewer prompt context.

Our key contributions are:

- We propose an end-to-end, data-driven model for predicting depression from interview transcripts that leverages the contextual information provided by interviewer prompts
- Our model jointly learns latent categorizations of prompts to aid prediction
- We conduct robust experiments to show that our model outperforms competitive baselines
- We analyze the model's behavior against existing psycholinguistic theory surrounding depressed language to demonstrate the interpretability of our model

2 Joint Latent Prompt Categorization

We propose a *Joint Latent Prompt Categorization* (*JLPC*) model that jointly learns to predict depression from interview transcripts while grouping interview prompts into latent categories.[1].

The general problem of classifying interview text is defined as follows: let X denote the set of N interview transcripts. Each interview X_i is a sequence of j conversational turns consisting of interviewer's prompts and participant's responses: $X_i = \{(P_{ij},\ R_{ij})$ for $j = \{1...M_i\}$, where M_i is the number of turns in X_i, P_{ij} is the j^{th} prompt in the i^{th} interview, and R_{ij} is the participant's response to that prompt. Together, $(P_{ij},\ R_{ij})$ form the j^{th} turn in i^{th} interview. Each interview X_i is labeled with a ground-truth class $Y_i \in \{1, ..C\}$, where C is the number of possible labels. In our case, there are two possible labels: *depressed* or *not depressed*. Our model, shown in Figure 1, takes as input an interview X_i and outputs the predicted label $\hat{Y}_i$.

Our approach assumes that prompts and responses are represented as embeddings $\mathbf{P}_{ij} \in \mathbb{R}^E$ and $\mathbf{R}_{ij} \in \mathbb{R}^E$ respectively. We hypothesize that prompts can be grouped into latent categories (K in number) such that corresponding responses will exhibit unique, useful patterns. To perform a soft assignment of prompts to categories, for each prompt, our model computes a category membership vector $\mathbf{h_{ij}} = [h^1_{ij}, \cdots, h^K_{ij}]$. It represents the probability distribution for the j^{th} prompt of the i^{th} interview over each of K latent categories. $\mathbf{h_{ij}}$ is computed as a function ϕ of $\mathbf{P_{ij}}$ and trainable parameters θ_{CI} (illustrated as the *Category Inference layer* in Figure 1):

$$\mathbf{h_{ij}} = \phi(\mathbf{P}_{ij}, \theta_{CI}) \tag{1}$$

Based on these category memberships for each prompt, the model then analyzes the corresponding responses so that unique patterns can be learned for each category. Specifically, we form K category-aware response aggregations. Each of these aggregations, $\bar{\mathbf{R}}^k_i \in \mathbb{R}^E$, is a category-aware representation of all responses of the i^{th} interview with respect to the k^{th} category.

$$\bar{\mathbf{R}}^k_i = \frac{1}{Z^k_i} \sum_{j=1}^{M_i} h^k_{ij} \times \mathbf{R}_{ij} \tag{2}$$

$$Z^k_i = \sum_{j=1}^{M_i} h^k_{ij} \tag{3}$$

where, h^k_{ij} is the k^{th} scalar component of the latent category distribution vector $\mathbf{h_{ij}}$ and Z^k_i is a normalizer added to prevent varying signal strength, which interferes with training.

We then compute the output class probability vector $\mathbf{y_i}$ as a function ψ of the response aggregations $[\bar{\mathbf{R}}^1_i, \cdots, \bar{\mathbf{R}}^K_i]$ and trainable parameters θ_D (illustrated as the *Decision Layer* in Figure 1).

$$\mathbf{y_i} = \psi(\bar{\mathbf{R}}^1_i, \cdots, \bar{\mathbf{R}}^K_i, \theta_D) \tag{4}$$

The predicted label $\hat{Y}_i$ is selected as the class with the highest probability based on $\mathbf{y_i}$.

[1] Code and instructions for reproducing our results are available at https://github.com/alexwgr/LatentPromptRelease

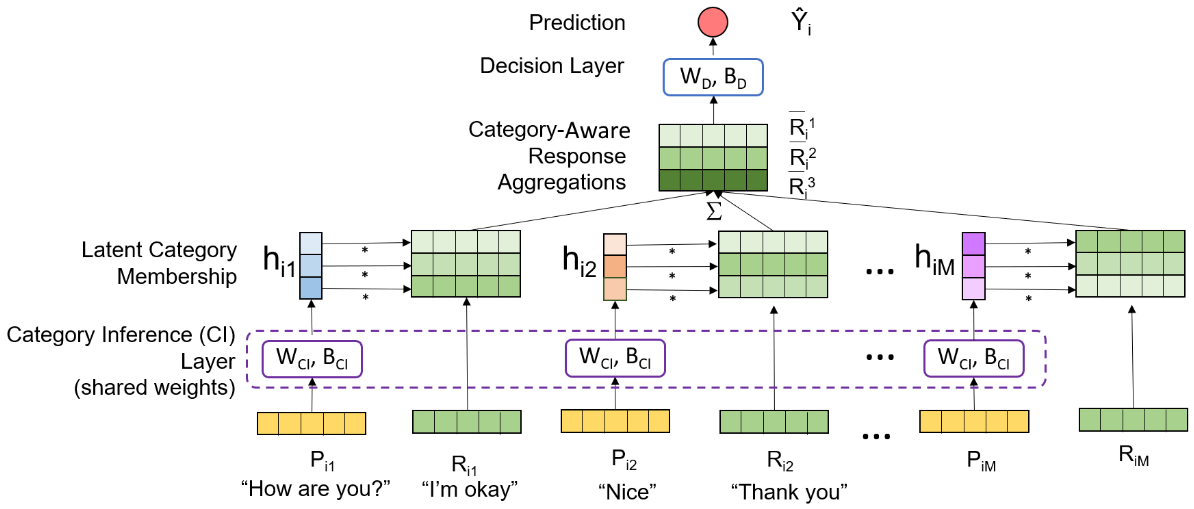

Figure 1: The architecture of our *JLPC* model with $K = 3$. For each prompt $\mathbf{P}_{ij}$ in interview i, the *Category Inference layer* computes a *latent category membership* vector, $\mathbf{h}_{ij}$. These are used as weights to form K separate *Category-Aware Response Aggregations*, which in turn are used by the *Decision Layer* to predict the output.

2.1 The Category Inference Layer

We compute the latent category membership for all prompts in interview i using a feed-forward layer with K outputs and softmax activation:

$$\phi(\mathbf{P}_{ij}, \theta_{CI}) = \sigma(\text{row}_j(\mathbf{P}_i \mathbf{W}_{CI} + \mathbf{B}_{CI})) \quad (5)$$

As shown in Equation 1, $\phi(\mathbf{P}_{ij}, \theta_{CI})$ produces the desired category membership vector $\mathbf{h_{ij}}$ over latent categories for the j^{th} prompt of the i^{th} interview. $\mathbf{P}_i \in \mathbb{R}^{M \times E}$ is defined as $[\mathbf{P}_{i1}, \cdots, \mathbf{P}_{iM}]^T$, where M is the maximum conversation length in X_i and $\mathbf{P}_{im} = \mathbf{0}^E$ for all $M_i < m \leq M$. $\mathbf{P}_i \mathbf{W}_{CI} + \mathbf{B}_{CI}$ computes a matrix where row j is a vector of energies for the latent category distribution for prompt j, and σ denotes the softmax function. $\mathbf{W}_{CI} \in \mathbb{R}^{E \times K}$ and $\mathbf{B}_{CI} \in \mathbb{R}^K$ are the trainable parameters for this layer: $\theta_{CI} = \{\mathbf{W}_{CI}, \mathbf{B}_{CI}\}$.

2.2 The Decision Layer

The Decision Layer models the probabilities for each output class (*depressed* and *not-depressed*) using a feed-forward layer over the concatenation $\bar{\mathbf{R}}_i$ of response aggregations $[\bar{\mathbf{R}}_i^1, \cdots, \bar{\mathbf{R}}_i^K]$. This allows each response aggregation $\bar{\mathbf{R}}_i^k$ to contribute to the final classification through a separate set of trainable parameters.

$$\psi(\bar{\mathbf{R}}_i^1, \cdots, \bar{\mathbf{R}}_i^K, \theta_D) = \sigma(\bar{\mathbf{R}}_i^T \mathbf{W}_D + \mathbf{B}_D) \quad (6)$$

As shown in Equation 4, $\psi(\bar{\mathbf{R}}_i^1, \cdots, \bar{\mathbf{R}}_i^K, \theta_D)$ produces the output class probability vector $\mathbf{y_i}$. $\mathbf{W}_D \in \mathbb{R}^{(E*K) \times C}$ and $\mathbf{B}_D \in \mathbb{R}^C$ are the trainable parameters for the decision layer: $\theta_D = \{\mathbf{W}_D, \mathbf{B}_D\}$.

We then compute the cross entropy loss $L(Y, \hat{Y})$ between ground truth labels and $\mathbf{y_i}$.

2.3 Entropy regularization

The model's learning goal as described above only allows the output prediction error to guide the separation of prompts into useful categories. However, in order to encourage the model to learn *distinct* categories, we employ entropy regularization (Grandvalet and Bengio, 2005) by penalizing overlap in the latent category distributions for prompts. That is, we compute the following entropy term using components of the category membership vector $\mathbf{h}_{ij}$ from Equation 1:

$$E(X_i) = \frac{1}{u_i} \sum_{i=1}^{N} \sum_{j=1}^{M_i} E_j(X_i) \quad (7)$$

where,

$$E_j(X_i) = -\sum_{k=1}^{K} h_{ij}^k \ln h_{ij}^k \quad (8)$$

$$u_i = \sum_{i=1}^{N} M_i \quad (9)$$

Finally, the model's overall learning goal minimizes entropy regularized cross entropy loss:

$$\arg\min_{\theta} L(Y, \hat{Y}) + \lambda E(X_i)$$

where, λ is a hyper-parameter that controls the strength of the entropy regularization term.

2.4 Leveraging Prompt Representations in the Decision Layer

While prompt representations are used to compute latent category assignments, the model described so far (*JLPC*) cannot directly leverage prompt features in the final classification. To provide this capability, we define two additional model variants with pre-aggregation and post-aggregation prompt injection: *JLPCPre* and *JLPCPost*, respectively.

JLPCPre is similar to the *JLPC* model, except that it aggregates both prompt and response representations based on prompt categories. In other words, the aggregated response representation, $\bar{\mathbf{R}}_i^k$ in Equation 2, is computed as:

$$\bar{\mathbf{R}}_i^k = \frac{1}{Z_i^k} \sum_{j=1}^{M_i} h_{ij}^k [\, \mathbf{P}_{ij}, \mathbf{R}_{ij} \,]$$

JLPCPost is also similar to *JLPC* except that it includes the average of prompt representations as additional input to the decision layer. That is, Equation 6 is modified to the following:

$$\psi(\bar{\mathbf{R}}_i^1, \cdots, \bar{\mathbf{R}}_i^K, \theta_D) = \sigma([\bar{\mathbf{P}}_i, \bar{\mathbf{R}}_i]^T \mathbf{W_D} + \mathbf{B}_D) \tag{10}$$

$\bar{\mathbf{P}}_i$ is the uniformly-weighted average of prompt representations in X_i.

3 Dataset

We evaluate our model on the Distress Analysis Interview Corpus (DAIC) (Gratch et al., 2014). DAIC consists of text transcripts of interviews designed to emulate a clinical assessment for depression. The interviews are conducted between human participants and a human-controlled digital avatar. Each interview is labeled with a binary depression rating based on a score threshold for the 9th revision of the *Patient Health Questionnaire* (PHQ-9). In total, there are 170 interviews, with 49 participants identified as depressed.

To achieve stable and robust results given the small size of the DAIC dataset, we report performance over 10 separate splits of the dataset into training, validation, and test sets. For each split, 70% is used as training data, and 20% of the training data is set aside as validation data.

3.1 Preprocessing and Representation

DAIC interview transcripts are split into utterances based on pauses in speech and speaker change, so we concatenate adjacent utterances by the same speaker to achieve a prompt-response structure. We experiment with two types of continuous representations for prompts and responses: averaged word embeddings from the pretrained GloVe model (Pennington et al., 2014), and sentence embeddings from the pretrained BERT model (Devlin et al., 2019). Further details are given in Appendix A.1. Reported results use GloVe embeddings because they led to better validation scores.

3.2 Exclusion of Predictive Prompts

Our preliminary experiments showed that it is possible to achieve better-than-random performance on the depression identification task using only the set of prompts (excluding the responses). This is possibly because the interviewer identified some individuals as potentially depressed *during* the interview, resulting in predictive follow-up prompts (for example, “How long ago were you diagnosed?”). To address this, we iteratively remove predictive prompts until the development performance using prompts alone is not significantly better than random (see Appendix A.3). This is to ensure our experiments evaluate the content of prompts and responses rather than fitting to any bias in question selection by the DAIC corpus interviewers, and so are generalizable to other interview scenarios, including future fully-automated ones.

4 Experiments

We now describe our experiments and analysis.

4.1 Baselines

Our experiments use the following baselines:

- The *RO* baseline only has access to responses. It applies a dense layer to the average of response representations for an interview.
- The *PO* baseline only has access to prompts, following the same architecture as *RO*.
- The *PR* baseline has access to both prompts and responses. It applies a dense layer to the average of prompt and response concatenations.

Model	F1 depressed	F1 not depr.
Random	0.303 (0.081)	0.690 (0.044)
PO	0.246 (0.080)	0.784 (0.032)
RO	0.309 (0.121)	0.798 (0.031)
PR	0.324 (0.121)	0.787 (0.030)
BERT	0.362 (0.080)	0.780 (0.062)
JLPC	0.416 (0.110)	0.761 (0.057)
JLPCPre	0.358 (0.121)	0.776 (0.037)
JLPCPost	**0.440 (0.080)**	0.768 (0.078)

Table 1: Mean F1 scores for the positive (depressed) and negative (not depressed) across the 10 test sets. Standard deviation is reported in parentheses. Two of the proposed models, *JLPC* and *JLPCPost*, improve over baselines including the BERT fine-tuned model (Devlin et al., 2019), with the *JLPCPost* achieving a statistically significant improvement ($p < 0.05$).

- BERT refers to the BERT model (Devlin et al., 2019) fine-tuned on our dataset (see Appendix A.2).

4.2 Training details

All models are trained using the Adam optimizer. We use mean validation performance to select hyper-parameter values: number of epochs $= 1300$, learning rate $= 5 \times 10^{-4}$, number of prompt categories $K = 11$ and entropy regularization strength $\lambda = 0.1$.

4.3 Quantitative Results

We computed the F1 scores of the positive (depressed) and negative (not-depressed) classes averaged over the 10 test sets. Given the class imbalance in the DAIC dataset, we compare models using F1 score for the depressed class.

As an additional baseline, we also implemented methods from Mallol-Ragolta et al. (2019) but do not report their performance since their model performs very poorly (close to random) when we consider averaged performance over 10 test sets. This is likely because of the large number of parameters required by the hierarchical attention model.

Table 1 summarizes our results. The below-random performance of the *PO* baseline is expected, since the prompts indicative of depression were removed as described in Section 3.2. This indicates the remaining prompts, by themselves, are not sufficient to accurately classify interviews. The *RO* model performs better, indicating the response information is more useful. The *PR* baseline improves over the *RO* baseline indicating that the combination of prompt and response information is informative. The *BERT* model, which also has access to prompts and responses, shows a reasonable improvement over all baselines.

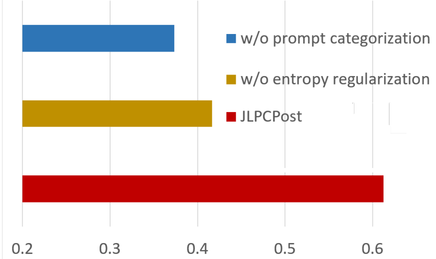

Figure 2: Ablation study on validation set demonstrating the importance of prompt categorization and entropy regularization for our model.

JLPC and *JLPCPost* outperform the baselines, with *JLPCPost* achieving a statistically significant improvement over both the *PR* and *BERT* baselines ($p < 0.05$).[2] This indicates the utility of our prompt-category aware analysis of the interviews.

4.4 Ablation study

We analyzed how the prompt categorization and entropy regularization contribute to our model's validation performance. The contributions of each component are visualized in Figure 2. Our analysis shows that while both components are important, latent prompt categorization yields the highest contribution to the model's performance.

4.5 Analyzing Prompt Categories

Beyond improving classification performance, the latent categorization of prompts yields insight about conversational contexts relevant for analyzing language patterns in depressed individuals.

To explore the learned categories, we isolate interviews from the complete corpus that are correctly labeled by our best-performing model. We say that the model "assigns" an interview prompt to a given category if the prompt's membership for that category (Equation 1) is stronger than for other categories. We now describe the various prompts assigned to different categories.[3]

Firstly, all prompts that are questions like "Tell me more about that", "When was the last time you had an argument?", etc. are grouped together into

[2] Statistical significance is calculated from the test prediction using two-sided T-test for independent samples of scores

[3] To verify consistency of prompt categorization, we rerun the model with multiple initialization and they all yielded the same general trends as described in the paper.

a single category, which we refer to as the *Starters* category. Previous work has identified usefulness of such questions as conversation starters since they assist in creating a sense of closeness (Mcallister et al., 2004; Heritage and Robinson, 2006).

Secondly, there are several categories reserved exclusively for certain *backchannels*. Backchannels are short utterances that punctuate longer turns by another conversational participant (Yngve, 1970; Goodwin, 1986; Bavelas et al., 2000). Specifically, the model assigns the backchannels “mhm,” “mm,” “nice,” and “awesome” each to separate categories. Research shows that it is indeed useful to consider the effects different types of backchannels separately. For example, Bavelas et al. (2000) propose a distinction between *specific backchannels* (such as “nice” and “awesome”) and *generic backchannels* (such as “mm” and “mhm”), and Tolins and Fox Tree (2014) demonstrated that each backchannel type serves a different purpose in conversation.

Thirdly, apart from starters and backchannels, the model isolates one specific prompt - “Have you been diagnosed with depression?”[4] into a separate category. Clearly, this is an important prompt and it is encouraging to see that the model isolates it as useful. Interestingly, the model assigns the backchannel “aw” to the same category as “Have you been diagnosed with depression?” suggesting that responses to both prompts yield similar signals for depression.

Lastly, the remaining five categories are empty - no prompt in the corpus has maximum salience with any of them. A likely explanation for this observation stems from the choice of normalizing factor Z_i^k in Equation 3: it causes $\bar{\mathbf{R}}_i^k$ to regress to the unweighted average of response embeddings when all prompts in an interview have low salience with category k. Repeated empty categories then function as an “ensemble model” for the average response embeddings, potentially improving predictive performance.

4.6 Category-based Analysis of Responses

The prompt categories inferred by our *JLPCPost* model enable us to take a data-driven approach to investigating the following category-specific psycholinguistic hypotheses about depression:

	Starters		Backchannels	
	D	ND	D	ND
RL	$\underline{23.2}$	$\underline{27.2}$	19.9	15.1
DMF ($\times 10^{-2}$)	$\underline{6.55}$	$\underline{7.31}$	7.98	8.55

Table 2: Indicators for social skills: mean response length (RL) and discourse marker/filler rates (DMF) for responses to prompts in *starters* and *backchannel* (collectively representing “mhm”, “mm”, “nice”, and “awesome”) categories, for depressed (D) and not-depressed (ND) participants. Statistically significant differences are underlined ($p < 0.05$). Both measures are significantly lower for the depressed class for responses to *starters*, but not to *backchannels*.

H1 Depression correlates with social skill deficits (Segrin, 1990)

H2 Depressed language is vague and qualified (Andreasen, 1976)

H3 Depressed language is self-focused and detached from community (Rude et al., 2004)

For hypothesis H1, we evaluate measures of social skill in responses to different categories of prompts. While research in psychology uses several visual, linguistic and paralinguistic indicators of social skills, in this paper we focus on two indicators that are measurable in our data: average response length in tokens and the rate of spoken-language *fillers* and *discourse markers* usage.[5] The first measure - response length - can be seen as a basic measure of taciturnity. The second measure - usage of *fillers* and *discourse markers* - can be used as proxy for conversational skills, since speakers use these terms to manage conversations (Fox Tree, 2010). Christenfeld (1995) and Lake et al. (2011) also find that discourse marker usage correlates with social skill. Following is the list of *fillers* and *discourse markers*: “um”, “uh”, “you know”, “well”, “oh”, “so”, “I mean”, and “like”.

Table 2 shows the values of these measures for social skill for responses to *backchannels* and *starters* categories. We found that both measures were significantly lower for responses to starters-category prompts for depressed participants as opposed to not-depressed participants ($p < 0.05$). However, the measures showed no significant difference between depressed and not-depressed individuals for responses to categories

[4] Note that this prompt was not removed in Section 3.2 since by itself, the prompt’s presence is not predictive of depression (without considering the response).

[5] We compute this measure as the ratio of discourse marker and filler occurrences to number of tokens, averaged over responses.

representing backchannels (“mhm,” “mm,” “awesome,” and “nice”). Note that a conversation usually begins with prompts from the starters category and thereafter backchannels are used to encourage the speaker to continue speaking (Goodwin, 1986). Given this, our results suggest that depressed individuals in the given population indeed initially demonstrate poorer social skills than not-depressed individuals, but the effect *levels off* as the interviewer encourages them to keep speaking using backchannels. Given this, our results suggest that depressed individuals in the given population indeed initially demonstrate poorer social skills than not depressed individuals, but the effect stops being visible as the conversation continues, either because the depressed individuals become more comfortable talking or because the interviewers’ encouragement through backchannels elicits more contributions.

Hypotheses H2 and H3 - regarding qualified language and self-focus, respectively - involve semantic qualities of depressed language. To explore these hypotheses, we use a reverse engineering approach to determine salient words for depression in responses to each prompt category.

We describe this reverse engineering approach as follows: since the aggregated representation of an individual’s responses in a category ($\bar{\mathbf{R}}_i^k$ computed in Equation 2) resides in the same vector space as individual word embeddings, we can identify words in our corpus that produce the strongest (positive) signal for depression in various categories. [6] We refer to these as *signal words*. Signal words are ranked not by their frequency in the dataset, but by their predictive potential - the strength of association between the word’s semantic representation and a given category. We evaluate hypotheses H2 and H3 by observing semantic similarities between these signal words and the language themes identified by the hypotheses. Selections from the top 10 signal words for depression associated with categories corresponding to starters, specific backchannels, and generic backchannels are shown in Figure 3.

Figure 3 shows hypothesis H2 is supported by signal words in responses to *generic backchannels*; words such as “theoretical” and “plausible” constitute qualified language, and in the context of generic backchannels, the proposed model identifies them as predictive of depression. Similarly, hypothesis H3 is also supported in responses to *generic backchannels*. The model identifies words related to community (“kids,” “neighborhood,” “we”) as strong negative signals for depression, supporting that depressed language reflects detachment from community.

	Starters	Generic Backchannels	Specific Backchannels
	“When was the last time you…?” “How have you been feeling lately?”	“mhm” “mm”	“nice” “awesome”
↑ Correlation with depression	nah	theoretical	mission
	ah	mechanical	invaluable
	eh	plausible	wished
	huh	indirect	secure
	wow	variable	accomplished
	sold	moms	California
	drinks	we	bars
	explorer	do	restaurants
	communications	neighborhood	mild
	veterinary	kids	hair

Figure 3: Signal words associated with language in depressed individuals. Columns represent various types of prompts (*Starters*, *Generic Backchannels* and *Specific Backchannels*). The bottom half shows ranked lists of signal words from the responses. Blue words are strongly indicative and red words are least indicative of depression.

However, the model only focuses on these semantic themes in responses to *generic backchannel* categories. As we found in our evaluation of hypothesis H1, the model localizes cues for depression to specific contexts. Signal words for depression in responses to the *starters* category are more reflective of our findings for hypothesis H1: the model focuses on short, low-semantic-content words that could indicate social skill deficit. For example, Figure 3 shows we identified “wow” as a signal word for the *starters* category. In one example from the corpus, a depressed participant uses “wow” to express uncomfortability with an emotional question: the interviewer asks, “Tell me about the last time you were really happy,” and the interviewee responds, “wow (laughter) um.”

For responses to *specific backchannels*, strong signal words reflect themes of goals and desires

[6] A word’s *signal strength* is computed for a given category k by taking the dot product of the word’s embedding with the weights in the decision layer corresponding to category k. Large positive numbers correspond to positive predictions and vice versa. Since the Decision Layer is a dot product with all response aggregations, it is intuitive to compute prediction strength for a group of categories by adding together prediction strengths from individual groups.

("wished," "mission," "accomplished"). Psychologists have observed a correlation between depression and goal commitment and pursuit (Vergara and Roberts, 2011; Klossek, 2015), and our finding indicates that depressed individuals discuss goal-related themes as response to specific backchannels.

Overall, our model's design not only helps in reducing its opacity but also informs psycholinguistic analysis, making it more useful as part of an informed decision-making process. Our analysis indicates that even though research has shown strong correlation between depression and various interpersonal factors such as social skills, self-focus and usage of qualified language, clinical support tools should focus on these factors in light of conversational cues.

4.7 Sources of Error

In this section, we analyze major sources of error. We apply a similar reverse engineering method as in Section 4.6. For prompts in each category, we consider corresponding responses that result in strong *incorrect* signals (false positive or false negative) based on the category's weights in the decision layer. We focus on the categories with the most significance presence in the dataset: the categories corresponding to starters, the "mhm" backchannel, and the prompt "Have you been diagnosed with depression?".

For the starters category, false positive-signal responses tend to contain a high presence of fillers and discourse markers ("uh," "huh," "post mm traumatic stress uh no uh uh," "hmm"). It is possible that because the model learned to focus on short, low-semantic-content responses, it incorrectly correlates presence of fillers and discourse markers with depression. For the "mhm" category, we identified several false negatives, in which the responses included concrete words like "uh nice environment", "I love the landscape", and "I love the waters". Since the "mhm" category focuses on vague, qualified language to predict depression (see Figure 3), the presence of concrete words in these responses could have misled the model. For the "Have you been diagnosed with depression?" category, the misclassified interviews contained short responses to this prompt like "so," "never," "yes," "yeah," and "no," as well as statements containing the word "depression." For this category, the model seems to incorrectly correlate short responses and direct mentions of depression with the depressed class.

5 Related Work

Much work exists at the intersection of natural language processing (NLP), psycholinguistics, and clinical psychology. For example, exploring correlations between counselor-patient interaction dynamics and counseling outcomes (Althoff et al., 2016); studying linguistic development of mental healthcare counsellors (Zhang et al., 2019); identifying differences in how people disclose mental illnesses across gender and culture (De Choudhury et al., 2017); predicting a variety of mental health conditions from social media posts (Sekulic and Strube, 2019; De Choudhury et al., 2013a; Guntuku et al., 2019; Coppersmith et al., 2014); and analyzing well-being (Smith et al., 2016) and distress (Buechel et al., 2018).

Specifically, many researchers have used NLP methods for identifying depression (Morales et al., 2017). They focus on for predicting depression from Twitter posts (Resnik et al., 2015; De Choudhury et al., 2013b; Jamil et al., 2017), Facebook updates (Schwartz et al., 2014), student essays (Resnik et al., 2013), etc.

Previous works have also focused on predicting depression severity from screening interview data (Yang et al., 2016; Sun et al., 2017; Pampouchidou et al., 2016). Unlike ours, these approaches rely on audio, visual, and text input.

More recent approaches are based on deep learning. Yang et al. (2017) propose a CNN-based model leveraging jointly trained paragraph vectorizations, Al Hanai et al. (2018) propose an LSTM-based model fusing audio features with Doc2Vec representations of response text, Makiuchi et al. (2019) combine LSTM and CNN components, and Mallol-Ragolta et al. (2019) propose a model that uses a hierarchical attention mechanism. However, these approaches are more opaque and difficult to interpret.

Other approaches are similar to ours in the sense that they utilize the structure provided by interview prompts. Al Hanai et al. (2018) and Gong and Poellabauer (2017) propose models that extract separate sets of features for responses to each unique prompt in their corpus. However, these approaches require manually identifying unique prompts. Our model can instead automatically learn new, task-specific categorization of prompts.

Lubis et al. (2018) perform a K-means clustering of prompt to assign prompts to latent dialogue act categories. These are used as features in a neural dialogue system. Our approach expands upon this idea of incorporating a separate unsupervised clustering step by allowing the learning goal to influence the clustering. Our approach is also related to that of Chaturvedi et al. (2014) in that it automatically categorizes various parts of the conversation. However, they use domain-specific handcrafted features and discrete latent variables for this categorization. Our approach instead can leverage the neural architecture to automatically identify features useful for this categorization.

To the best of our knowledge, our approach is the first deep learning approach that jointly categorizes prompts to learn context-dependent patterns in responses.

6 Conclusion

This paper addressed the problem of identifying depression from interview transcripts. The proposed model analyzes the participant's responses in light of various categories of prompts provided by the interviewer. The model jointly learns these prompt categories while identifying depression. We show that the model outperforms competitive baselines and we use the prompt categorization to investigate various psycholinguistic hypotheses.

Depression prediction is a difficult task which requires especially trained experts to conduct interviews and do their detailed analysis (Lakhan et al., 2010). While the absolute performance of our model is low for immediate practical deployment, it improves upon existing methods and at the same time, unlike modern methods, provides insight about the model's workflow. For example, our findings show how language of depressed individuals changes when interviewers use backchannels to encourage continued speech. We hope that this combination will encourage the research community to make more progress in this direction. Future work can further investigate temporal patterns in how language used by depressed people evolves over the course of an interaction.

References

Tuka Al Hanai, Mohammad Ghassemi, and James Glass. 2018. Detecting Depression with Audio/Text Sequence Modeling of Interviews. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 1716–1720.

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Nancy J. C. Andreasen. 1976. Linguistic Analysis of Speech in Affective Disorders. *Archives of General Psychiatry*, 33(11):1361.

Janet B. Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling Empathy and Distress in Reaction to News Stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4758–4765.

Snigdha Chaturvedi, Dan Goldwasser, and Hal Daumé III. 2014. Predicting instructor's intervention in MOOC forums. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1501–1511, Baltimore, Maryland. Association for Computational Linguistics.

Nicholas Christenfeld. 1995. Does it hurt to say um? *Journal of Nonverbal Behavior*, 19:171–186.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013a. Predicting postpartum changes in emotion and behavior via social media. In *2013 ACM SIGCHI Conference on Human Factors in Computing Systems, CHI '13, Paris, France, April 27 - May 2, 2013*, pages 3267–3276.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013b. Predicting Depression via Social Media. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013.*

Munmun De Choudhury, Sanket S. Sharma, Tomaz Logar, Wouter Eekhout, and René Clausen Nielsen. 2017. Gender and Cross-Cultural Differences in Social Media Disclosures of Mental Illness. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, pages 353–369, Portland, Oregon, USA. ACM Press.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Edward Fast, Alesia Gainer, Kallirroi Georgila, Jonathan Gratch, Arno Hartholt, Margaux Lhommet, Gale M. Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David R. Traum, Rachel Wood, Yuyu Xu, Albert A. Rizzo, and Louis-Philippe Morency. 2014. SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support. In *International conference on Autonomous Agents and Multi-Agent Systems, AAMAS '14, Paris, France, May 5-9, 2014*, pages 1061–1068.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Jean E. Fox Tree. 2010. Discourse Markers across Speakers and Settings. *Language and Linguistics Compass*, 4(5):269–281.

Yuan Gong and Christian Poellabauer. 2017. Topic Modeling Based Multi-modal Depression Detection. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge - AVEC '17*, pages 69–76, Mountain View, California, USA. ACM Press.

Charles Goodwin. 1986. Between and within: Alternative sequential treatments of continuers and assessments. *Human Studies*, 9(2-3):205–217.

Yves Grandvalet and Yoshua Bengio. 2005. Semi-supervised Learning by Entropy Minimization. page 8.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The Distress Analysis Interview Corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 3123–3128.

Sharath Chandra Guntuku, Daniel Preotiuc-Pietro, Johannes C. Eichstaedt, and Lyle H. Ungar. 2019. What Twitter Profile and Posted Images Reveal about Depression and Anxiety. In *Proceedings of the Thirteenth International Conference on Web and Social Media, ICWSM 2019, Munich, Germany, June 11-14, 2019*, pages 236–246.

John Heritage and Jeffrey Robinson. 2006. The Structure of Patients' Presenting Concerns: Physicians' Opening Questions. *Health communication*, 19:89–102.

Zunaira Jamil, Diana Inkpen, Prasadith Buddhitha, and Kenton White. 2017. Monitoring Tweets for Depression to Detect At-risk Users. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 32–40, Vancouver, BC. Association for Computational Linguistics.

Ulrike Klossek. 2015. *The Role of Goals and Goal Orientation as Predisposing Factors for Depression*. Ph.D. thesis, University of Exeter.

Johanna K. Lake, Karin R. Humphreys, and Shannon Cardy. 2011. Listener vs. speaker-oriented aspects of speech: Studying the disfluencies of individuals with autism spectrum disorders. *Psychonomic Bulletin & Review*, 18(1):135–140.

Shaheen E Lakhan, Karen Vieira, and Elissa Hamlat. 2010. Biomarkers in psychiatry: drawbacks and potential for misuse. *International Archives of Medicine*, 3(1):1.

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2018. Unsupervised Counselor Dialogue Clustering for Positive Emotion Elicitation in Neural Dialogue System. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 161–170, Melbourne, Australia. Association for Computational Linguistics.

Mariana Rodrigues Makiuchi, Tifani Warnita, Kuniaki Uto, and Koichi Shinoda. 2019. Multimodal Fusion of BERT-CNN and Gated CNN Representations for Depression Detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop - AVEC '19*, pages 55–63, Nice, France. ACM Press.

Adria Mallol-Ragolta, Ziping Zhao, Lukas Stappen, Nicholas Cummins, and Björn W. Schuller. 2019. A Hierarchical Attention Network-Based Approach for Depression Detection from Transcribed Clinical Interviews. In *Interspeech 2019*, pages 221–225. ISCA.

Margaret Mcallister, Beth Matarasso, Barbara Dixon, and C Shepperd. 2004. Conversation starters: re-examining and reconstructing first encounters within the therapeutic relationship. *Journal of Psychiatric and Mental Health Nursing*, 11.

Michelle Morales, Stefan Scherer, and Rivka Levitan. 2017. A Cross-modal Review of Indicators for Depression Detection Systems. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 1–12, Vancouver, BC. Association for Computational Linguistics.

Anastasia Pampouchidou, Kostas Marias, Fan Yang, Manolis Tsiknakis, Olympia Simantiraki, Amir Fazlollahi, Matthew Pediaditis, Dimitris Manousos, Alexandros Roniotis, Georgios Giannakakis, Fabrice Meriaudeau, and Panagiotis Simos. 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 - AVEC '16*, pages 27–34, Amsterdam, The Netherlands. ACM Press.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015. Beyond LDA: Exploring Supervised Topic Modeling for Depression-Related Language in Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 99–107, Denver, Colorado. Association for Computational Linguistics.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using Topic Modeling to Improve Prediction of Neuroticism and Depression in College Students. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1348–1353, Seattle, Washington, USA. Association for Computational Linguistics.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards Assessing Changes in Degree of Depression through Facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125, Baltimore, Maryland, USA. Association for Computational Linguistics.

Chris Segrin. 1990. A meta-analytic review of social skill deficits in depression. *Communication Monographs*, 57(4):292–308.

Ivan Sekulic and Michael Strube. 2019. Adapting Deep Learning Methods for Mental Health Prediction on Social Media. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 322–327, Hong Kong, China. Association for Computational Linguistics.

Laura Smith, Salvatore Giorgi, Rishi Solanki, Johannes Eichstaedt, H. Andrew Schwartz, Muhammad Abdul-Mageed, Anneke Buffone, and Lyle Ungar. 2016. Does 'well-being' translate on Twitter? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2042–2047, Austin, Texas. Association for Computational Linguistics.

Bo Sun, Yinghui Zhang, Jun He, Lejun Yu, Qihua Xu, Dongliang Li, and Zhaoying Wang. 2017. A Random Forest Regression Method With Selected-Text Feature For Depression Assessment. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge - AVEC '17*, pages 61–68, Mountain View, California, USA. ACM Press.

Jackson Tolins and Jean E. Fox Tree. 2014. Addressee backchannels steer narrative development. *Journal of Pragmatics*, 70:152–164.

Chrystal Vergara and John E. Roberts. 2011. Motivation and goal orientation in vulnerability to depression. *Cognition and Emotion*, 25(7):1281–1290.

A. H. Weinberger, M. Gbedemah, A. M. Martinez, D. Nash, S. Galea, and R. D. Goodwin. 2018. Trends in depression prevalence in the USA from 2005 to 2015: widening disparities in vulnerable groups. *Psychological Medicine*, 48(8):1308–1315.

Le Yang, Dongmei Jiang, Lang He, Ercheng Pei, Meshia Cédric Oveneke, 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 - AVEC '16*, pages 89–96, Amsterdam, The Netherlands. ACM Press.

Le Yang, Dongmei Jiang, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. 2017. Multimodal Measurement of Depression Using Deep Learning Models. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge - AVEC '17*, pages 53–59, Mountain View, California, USA. ACM Press.

V. H. Yngve. 1970. On getting a word in edgewise. *In Chicago Linguistics Society, 6th Meeting*, pages 567–578.

Justine Zhang, Robert Filbin, Christine Morrison, Jaclyn Weiser, and Cristian Danescu-Niculescu-Mizil. 2019. Finding Your Voice: The Linguistic Development of Mental Health Counselors. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 946–947.

A Appendices

A.1 Continuous representation of utterances

For continuous representation using the GloVe model, we use the pretrained 100-dimensional embeddings (Pennington et al., 2014). The representation of an utterance is computed as the average of embeddings for words in the utterance, with $\mathbf{0}^{100}$ used to represent words not in the pretrained vocabulary. Based on the pretrained vocabulary, contractions (e.g. "can't") are decomposed. For continuous representation with the

BERT model, utterances are split into sequences of sub-word tokens following the authors' specifications (Devlin et al., 2019), and the pretrained BERT (Base, Uncased) model computes a 768-dimensional position-dependent representation.

A.2 Training the BERT Model

For the *BERT* model, all interviews were truncated to fit the maximum sequence length of the pretrained BERT model (Base, Uncased): 512 subword tokens. Truncation occurs by alternating between removing prompt and response tokens until the interview length in tokens is adequate.

Devlin et al. (2019) suggest trying a limited number of combinations of learning rate and training epochs to optimize the BERT classification model. Specifically, the paper recommends combinations of 2, 3, or 4 epochs and learning rates of 2E-5, 3E-5, and 5E-5. We noted that validation and test scores were surprisingly low (significantly below random) using these combinations, and posited that the small number of suggested epochs could have resulted from the authors only evaluating BERT on certain types of datasets. Accordingly, we evaluated up to 50 epochs with the suggested learning rates and selected a learning rate of 2E-5 with 15 epochs based on validation results.

A.3 Exclusion of prompts

The goal of removing prompts is to prevent a classifier from identifying participants as depressed based on certain prompts simply being present in the interview, such as "How long ago were you diagnosed [with depression]?" While some prompts are clear indicators, early tests showed that even with these prompts removed, other prompts were predictors for the participant being depressed for no obvious reason, indicating a bias in the design in the interview. Rather than using arbitrary means to determine whether prompts could be predictive, we used a machine-learning based algorithm to identify and remove predictive prompts from interviews.

After the division of interviews into turns as described in Section 3.1, we extracted the set of distinct prompts $P_{distinct}$ from all interviews (with no additional preprocessing). We then iteratively performed 10 logistic regression experiments using the same set of splits described in Section 4.2.

In a given experiment, each interview was represented as an indicator vector with $|P_{distinct}|$ dimensions, such that position p is set to 1 if prompt $p \in \{1, \cdots, |P_{distinct}|\}$ is present in the interview, and 0 otherwise. Logistic Regression was optimized on the vector representations for the training interviews. The predicted F1 score for the depressed class on the validation set was recorded for each experiment.

The average weight vector for the 10 Logistic regression models was computed. The prompt corresponding to the highest weight was removed from $P_{distinct}$ and added to a separate set D of predictive prompts. The process was repeated until the mean validation F1 score was less than the random baseline for the dataset (see Section 4.3).

The final set of 31 prompts D had to be removed from the dataset before the baselines and proposed approaches could be evaluated. The design of the DAIC interview posed a challenge, however: the same prompt can appear in many interviews, but preceded by unique interjections by the interviewer, such as "mhm," "nice," and "I see". We refer to this interjections as "prefixes." We manually compiled a list of 37 prefixes that commonly reoccur in interviews. For all interviews, if a prompt from $P_{distinct}$ occurred in the interview after prefixes were ignored, then both the prompt and its corresponding response were removed from the interview before training. This resulted in an removing an average of 13.64 turns from each interview in the dataset.