An analysis of observation length requirements for machine understanding of human behaviors from spoken language

Sandeep Nallan Chakravarthula$^1$, Brian Baucom$^2$, Shrikanth Narayanan$^1$, and Panayiotis Georgiou$^1$

$^1$Department of Electrical and Computer Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA
$^2$Department of Psychology, College of Social & Behavioral Science, University of Utah, Salt Lake City, UT, USA

Corresponding author:
Sandeep Nallan Chakravarthula
Email address: nallanch@usc.edu

ABSTRACT

Machine learning-based human behavior modeling, often at the level of characterizing an entire clinical encounter such as a therapy session, has been shown to be useful across a range of domains in psychological research and practice from relationship and family studies to cancer care. Existing approaches typically first quantify the target behavior construct based on cues in an observation window, such as a fixed number of words, and then aggregate it over all the windows in that session. During this process, a sufficiently long window is employed so that adequate information is gathered to accurately estimate the construct. The link between behavior modeling and the observation length, however, has not been well studied, especially for spoken language. In this paper, we analyze the effect of observation window length on the quality of behavior quantification and present a framework for determining appropriate windows for a wide range of behaviors. Our analysis method employs two levels of evaluations: (a) extrinsic similarity between machine predictions and human expert annotations, and (b) intrinsic consistency between intra-machine and intra-human behavior relations. We apply our analysis on a dataset of real-life married couple interactions that are annotated for a large and diverse set of behavior codes and test the robustness of our findings to different machine learning models. We find that negative constructs such as blame can be accurately identified from short expressions while those pertaining to positive affect such as satisfaction tend to require slightly longer observation windows. Behaviors that describe more complex personality traits such as negotiation and avoidance are found to require very long observations and are difficult to quantify from language alone. Our findings are generally in agreement with similar work using acoustic vocal cues as well as existing literature in psychology on thin slices and human emotion perception.

1 INTRODUCTION

Human interactions involve social-cognitive abilities of varying levels of complexity such as speech detection, language understanding, emotion recognition and appropriate response generation. Among these, the ability to reliably and accurately assess a person’s behavior by observing their verbal and non-verbal cues is a considerably complex and important one. Such a skill is especially important for both delivery and assessment in psychological research clinical practice as varied as Couples Therapy (Christensen et al., 2004), Addiction Counseling (Baer et al., 2009) and Cancer Care (Reblin et al., 2019). Analysis of the provider and client behaviors in these encounters, often through formal behavioral coding by human experts, is carried out during interactions in order to provide feedback and improve the clinical effectiveness of care. Subsequently, there have been efforts (Narayanan and Georgiou, 2013) to automate...
this behavior annotation (or coding) process using machine learning so that rapid and inexpensive feedback can be provided to the stakeholders. Previous work has shown that automated coding systems are effective at quantifying behaviors from speech and spoken language such as Negativity (Georgiou et al., 2011), Black et al., 2013, Chakravarthula et al., 2015a, Tseng et al., 2017), Depression (Gupta et al., 2014, Morales et al., 2018) and Empathy (Xiao et al., 2012, Gibson et al., 2016, Perez-Rosas et al., 2017). However, there are some critical aspects of this behavior assessment process which humans can handle naturally and easily but machines still cannot, one of which is the notion of how much to observe in order to reliably assess behavior.

When assessing a person’s behavior based on their interaction cues, humans look at factors such as the intensity of expression, context and how frequently the behavior is observed (Baucom et al., 2011). The latter two imply that an appropriately long window is used to observe the cues before making a judgment about the behavior; for lexical cues, we measure the length of this observation window in terms of the number of words spoken. While some behaviors can be assessed based on short-duration cues, others require observations along longer time-scales. For example, one can sense that a person is Angry if they say something as brief as “Shut up!”, but it is difficult to judge whether they are Engaged in a discussion or not unless a longer and more involved conversation is carried out. Based on this, it is intuitive to expect that evaluating different behaviors would require different observation window lengths. Such associations have been exhibited by humans when judging human characteristics such as personality traits (Bluckman and Funder, 1998), non-verbal behaviors (Murphy et al., 2018) and group dynamics (Satterstrom et al., 2019).

However, it is not clear as to how these associations manifest in behaviors or in automated systems that quantify behaviors based on interaction cues. Unlike emotions, which are simple and rapid (Baumeister et al., 2007) and can be reliably observed using short observation windows such as a few seconds (Schuller et al., 2012), a sentence (Zadeh et al., 2018) or a speaker turn (Busso et al., 2008), the rich variety of human behaviors can be much more complicated and long-ranging. Even expert coders in the field of psychological research typically first have to be trained according to domain-specific guidelines or manuals such as CIRS (Heavey et al., 2002) for the study of couple interactions and MITI (Moyers et al., 2003) for motivational interview based counseling before they can start coding patients’ behaviors. This complexity can potentially give rise to uncertainty at the time of assessment, which then necessitates longer observations in order to achieve confident and reliable annotation. Furthermore, the annotation time-frames for coding different behaviors can range from as short as 30 seconds (Heyman, 2004) to as long as 10 minutes (Heavey et al., 2002), demonstrating the potential variability in observation lengths. These facets of behavior coding demonstrate the need for investigating the role of the length of observation for specific behavioral characterization.

![Figure 1. Automated quantification of behavior using a moving-window modeling approach.](image)

This figure shows a process for automatically quantifying behavior from lexical cues. First, during an interaction, interlocutors express multiple behaviors, one of which is \( B_i \), through multiple cues, one of which is language. This language stream is decomposed into its constituent windows of text length \( L \). Then, model \( M(\theta) \) is used to score the text inside each window, resulting in a trajectory of window-level scores. Finally, a functional \( F \) is applied on the trajectory to summarize it as a single interaction-level score.

More importantly, the knowledge of how long an automated system should observe an interaction is vital in applications that rely on moving-window approaches. Such approaches, as shown in Figure 1, typically first compute a “local” behavior score within a fixed-length observation window, such as a few words or speaker turns, and then aggregate local scores from all the windows to obtain a “global” score for behavior. These are especially useful in psychological research settings (Xiao et al., 2012, Gibson et al., 2016, Tseng et al., 2017) where automated systems need to analyze “sessions”, or interactions, at...
fine-grained scales, such as turn-level, but are evaluated at session-level. In such situations, the choice of length of the observation window is important; too short a window can result in noisy or incorrect local scores, since insufficient information is being used, and as a result, the global summary score will be inaccurate, as illustrated in the toy example in Figure 2.

Another scenario where knowledge of the observation length is critical is in multi-speaker interactions where speakers exhibit and influence each others’ behaviors. For example, consider an interlocutor and their partner discussing an important problem, where the interlocutor expresses indifference towards the problem, causing the partner to become angry and yell. While it is possible for a system to detect that the partner is angry just by observing their last speaking turn, in order to understand that the interlocutor was being indifferent, it might need to observe several previous turns from both speakers. Recognizing different behaviors might thus require observations over different time-scales.

![Figure 2. Toy example illustrating the effect of judging how Supportive the statement “honey it’s not your fault please” is at different observation window lengths. When at least 5 words are used, the judgments are accurate. At shorter windows, however, the results are incorrect and noisy across windows; this is especially evident in 1-word windows where each word on its own cannot provide sufficient behavior information.](image)

In this work, we present a systematic analysis of the observation window length for quantifying behavior and how this varies for different behaviors. Specifically, we are interested in empirically identifying the minimum amount of language information, measured in number of words, from which a behavior can be judged. By analyzing how the quality of language-based behavior quantification varies with the length of the observation window, we address the following questions:

1. Do different behaviors need observations of different lengths in order to be quantified from language?
2. How is the nature or type of behavior related to the length of its required observation window?
3. Are the observation length requirements for modeling behaviors from the lexical modality of human speech similar to those in the acoustic modality?

Addressing these requires us to evaluate the system’s behavior modeling scores at multiple window lengths, even when ground-truth annotations might not be available to directly compare against. While evaluating the aggregate score serves as an indirect solution to this problem, it is still limited by the choice and appropriateness of the aggregating mechanism. For this purpose, we propose an additional metric that evaluates with respect to the relations between the ground-truth behavior annotations. For instance, we know that humans perceive Happiness to be similar to Satisfaction but distinct from Sadness; ideally, an automated system’s scores of Happiness, Satisfaction and Sadness would exhibit these relations as well. By investigating how these metrics change with window length, we aim to identify the appropriate amounts of language required to quantify each behavior. We also compare our findings with existing work in psychology and machine learning and identify some general trends based on the observed similarities in results. For our analysis, we choose the Couples Therapy corpus (Christensen et al., 2004) which contains real-life interactions coded for a large and diverse set of behaviors in human interaction.

This paper is structured as follows: In Sec. 2, we describe existing work in psychology and speech processing that has dealt with related problems, following which we formally define our problem of interest in Sec. 3 to distinguish our scope and approach from existing work. Sec. 4.1 and 4.2 describe the process of window-based scoring of behavior from text using two different machine learning methods, an N-gram-based maximum likelihood model and a Deep Neural Network-based estimation model. Following this, Sec. 4.3 introduces the intrinsic and extrinsic metrics for evaluating the behavior...
quantification of the machine learning models and explains the procedure by which these metrics are employed to determine the appropriate observation window lengths for different behavior constructs. Sec. 5 then describes the dataset on which we apply our analysis, the Couples Therapy corpus, which consists of interactions between married couples that are rated by human experts for a large, diverse set of behavior codes. Finally, in Sec. 6 we present and detail the results of our observation window length analysis in both models, followed by a discussion of findings that are consistent across both models and how they address the questions posed earlier. We present our conclusions in Sec. 7 and comment on potential improvements and extensions to this work.

2 RELATED WORK

A body of work in psychology that is related to, but not the same as, our notion of “window length” is the one which studies Thin Slices of observed behavior (Ambady and Rosenthal, 1992). It refers to excerpts or snippets of an interaction that can be used to arrive at a similar judgment of behavior to as if the entire interaction had been used. Essentially, it implies that an entire interaction can be replaced with just a windowed part (which is different from our aim of identifying the best window through which to view the entire interaction). The effect of the location of these slices has been investigated as well; the conventional approach is to situate the slices near the start of the interaction. The effectiveness of thin slices has been observed in many applications, ranging from judging personality traits (Blackman and Funder, 1998) to viewer impressions of TED talks (Cullen and Harte, 2017) such as “funny” and “inspiring”.

Notably, Carney et al. (Carney et al, 2007) studied the accuracy of impressions for Affect, Extraversion and Intelligence at different thin-slice durations, locations, etc. Accuracy was measured as the correlation between the true value of a construct (whether rated or self-reported) and the impression based on the thin slice, and it was observed that, in general, accuracy increased as the slice length increased from 5 seconds to 5 minutes. Furthermore, they found that Negative affect could be assessed with similar accuracies at all slice lengths whereas Positive affect was best assessed only when thicker slices were used. These works provide an encouraging support for analyzing the window length of behaviors along similar lines.

There has also been a great deal of work in psychology on studying how humans perceive and process events characterized by concepts such as good and bad. Specifically, there exists a notion that “bad” is “stronger” than “good” (Baumeister et al., 2001), meaning that undesirable or unpleasant events have a greater impact on an individual than desirable, pleasant ones. The behavior constructs that we are interested in analyzing are similar to concepts that have been shown to exhibit this phenomenon in previous works. For instance, Oehman et al. (Oehman et al., 2001) found that people detected threatening faces more quickly and accurately than those depicting friendly expressions. In a similar experiment, Krull et al. (Krull and Dil, 1998) showed videos of either happy or sad individuals to participants and reported that happiness evoked more spontaneous inferences while sadness drew slower ones. This shows that different concepts are perceived differently, depending on their valence; hence, in this work, we investigate how the nature of different behavior constructs is tied to aspects of their expression and perception in language such as window length, aggregation mechanism, etc.

Some works that are related to ours are those that have investigated the accuracy of behavior prediction using acoustic vocal cues. Xia et al. (Xia et al., 2015) found that as the observation window used to compute acoustic features was increased from 2 seconds to 50 seconds, the classification accuracy generally improved, with Negative and Positive behaviors gaining the most. Similar results were reported by Li et al. (Li et al., 2019) who classified behaviors such as Acceptance, Negativity and Sadness by employing emotion-based behavior models on acoustic features. In their models, the receptive field (measured in seconds) of a 1-D Convolutional Neural Network-based system served as the observation window for vocal cues. In general, they found that behaviors relating to negative affect, such as Negativity, were classified more accurately than behaviors such as Acceptance and Sadness. They also observed that increasing the receptive field from 4 seconds to 64 seconds generally resulted in better classification, with Sadness performing best at 16 seconds while Negativity performed best at 64 seconds. Other efforts have addressed related aspects; for instance, Lee et al. (Lee et al., 2012) examined whether the behavior annotation process is driven more by a gradual, causal mechanism or by isolated salient events, which imply the use of long and short observation windows respectively.

While these works have contributed to a better understanding of the effect of observation windows, they are limited, including in the variety of constructs that are analyzed. Notably, they mostly focus
on acoustic and vocal cues and not enough on the linguistic characteristics. Hence, the novelty of our work lies in (1) analyzing the effect of observation lengths in the lexical modality and (2) performing this analysis using a large and diverse set of real-life human behavior constructs. Through our analysis, we aim to understand the relation between the nature of the behavior and how long an automated system needs to observe the language through which it is being expressed.

3 PROBLEM STATEMENT

Let’s suppose that we want to model behavior \( B_i \) using a set of observed data samples \( D \), where each sample is a sequence of words. Let \( A_i \in \mathbb{R}^{|D|} \) be the ground-truth annotations of \( B_i \) in \( D \) and let \( C \) be a metric that is used to evaluate the quantification results against \( A_i \); the higher the value of \( C \), the better the model fit.

Let \( E_{\text{lang}}(B_i) \) represent the degree to which \( B_i \) is actually expressed in language. Let \( M \) denote a machine learning model (e.g., Deep Neural Network) that estimates a scalar value from a sequence of words and \( \theta \) represent its learnable parameters (e.g., weights). Finally, let \( L \) represent the window length at which \( B_i \) is observed in language and \( F \) denote a statistical functional that maps a sequence of scalar values to a single scalar value. Then, the quality of behavior quantification can be expressed as:

\[
Q_i = C(F(M(E_{\text{lang}}(B_i), L, D; \theta)), A_i)
\] (1)

Our goal is to identify the best window length \( L_i^* \), for each \( B_i \), at which \( Q_i \) is maximized:

\[
\Rightarrow L_i^* = \arg \max_L Q_i
\]

\[
= \arg \max_L C(F(M(E_{\text{lang}}(B_i), L, D; \theta)), A_i)
\] (2)

We argue that \( Q_i \) can be high only when \( E_{\text{lang}}(B_i), L, M, \theta \) and \( F \) are all appropriate together. If even one of them is flawed or incompatible with the rest, then it would adversely affect \( Q_i \), as explained below:

- \( E_{\text{lang}}(B_i) \): The behavior \( B_i \) must be sufficiently expressed in the lexical channel to begin with; otherwise, it might not be possible to observe it using lexical cues alone (for example, if it is instead primarily expressed through nonverbal vocal cues such as laughter).

- \( L \): The observation window must be long enough to observe \( B_i \); otherwise, the incomplete information from partial observations can lead to noisy or incorrect estimates.

- \( M \): The model must be well-suited for capturing \( B_i \). For example, quantifying a behavior that is based on the actions of both speakers requires a model that looks at both speakers; using a single-speaker model instead might provide incomplete information and, hence, inaccurate estimates.

- \( \theta \): The model must be well-trained; otherwise, its estimates might be inaccurate. This is dependent on the training process, the amount and quality of data used to estimate parameters, etc.

- \( F \): The aggregating functional must be well-suited for summarizing \( B_i \); otherwise, the resulting overall scores might not match the ground-truth annotations \( A_i \). For example, the functional mode, which identifies the most frequently occurring value, might not be appropriate for summarizing a behavior that is expressed very infrequently.

We reason that a high value of \( Q \) is indicative of all the aforementioned factors being appropriate and a low value is indicative of a limitation in at least one of these factors. Based on this, we now proceed to analyze the variation in \( Q \) for different behaviors as window length \( L \) is varied.

4 METHODOLOGY

This section provides the details of our methodology, starting with the process for creating windowed samples of language from the interaction transcripts. Then, we describe the machine learning models used in this work to score the window-level text samples for degrees of different behavior constructs. Following this, we describe two metrics for evaluating the quality of the models’ window-level scores and, finally, we detail the procedure for employing these metrics to identify the appropriate window lengths for different behaviors.
4.1 Windowed Scoring of Text
To score the text $T_i$ of an interaction containing $O_i$ words using an observation window of length $L_j$, we first decompose it into its constituent windows. If $O_i > L_j$, then we get $O_i - L_j + 1$ windows, each containing $L_j$ words; else, we get just one window. Then, each window is scored independently using model $M$. Assuming that $T_i = <s> w_1, w_2, \ldots, w_{O_i} <s>$, where $<s>$ and $<s>$ represent start and end respectively, there are $O_i$ distinct observation window lengths at which it can be scored, as shown in the first column of Table. 1.

| Observation Window | Window Decomposition | No. of Scores | Resolution |
|--------------------|----------------------|--------------|------------|
| interaction-length | $<s> w_1, w_2, \ldots, w_{O_i} <s>$ | 1 | Very Coarse |
| $O_i$-1 word       | $<s> w_1, w_2, \ldots, w_{O_i-1}$, $w_3, \ldots, w_{O_i} <s>$ | 2 | Coarse      |
| 2-word             | $<s> w_1, w_2$, $w_2, w_3$, $w_3, \ldots, w_{O_i-1}, w_{O_i} <s>$ | $O_i$-1 | Fine        |
| 1-word             | $<s> w_1$, $w_2$, $w_2, w_3$, $w_3, \ldots, w_{O_i-1}, w_{O_i} <s>$ | $O_i$ | Very Fine   |

Table 1. Scoring a sentence with $O_i$ words using observation windows of different lengths

The window with the coarsest resolution is the “interaction-length” window since it scores the entire interaction as a whole, resulting in a single score for $T_i$. On the other hand, the “1-word” window provides the finest resolution possible since a score is generated for each word of the sentence, resulting in a trajectory of scores for $T_i$. In this work, we test the following observation window lengths: $\{3, 10, 30, 50, 100\}$ where 3 represents a 3-word window and interaction represents a window spanning the entire length of the interaction. We qualitatively refer to 3 and 10 as “short” windows, 30 as “medium”-length and 50 and 100 as “long” windows.

4.2 Behavior Quantification Models
From Eqn. 2, we see that the analysis of window lengths depends on the model $M$; therefore, it is possible that the results of our analysis might be different for different choices of $M$. In such a scenario, running an analysis using just one modeling framework might be limiting, since the results might be reflective of the model and not necessarily the behaviors. This can be addressed by testing different models and observing traits that are specific to each model as well as traits that are consistent across different models, thereby providing a more comprehensive understanding about how behaviors are expressed through language and how they are affected by the observation window length. Hence, in this work, we conduct our analysis using two different modeling frameworks: (1) Maximum Likelihood N-gram models and (2) Deep Neural Networks, the details of which are provided in this section.

4.2.1 Maximum-Likelihood N-gram
Model
We use a Maximum Likelihood method closely following [Frank and Hall 2001; Rozgić et al. 2011], where N-gram Language Models (LMs) are employed to map text onto a scalar value. N-gram LMs compute the joint probability of a sequence of words or, equivalently, the probability of a word given the context of the preceding $n-1$ words. Given a sequence of $M$ words $W = \{w_1, w_2, \ldots, w_M\}$, its N-gram probability is given by:

$$P(W) = P(w_1, w_2, \ldots, w_{M}) \approx \prod_{j=1}^{M} P(w_j | w_{j-1}, w_{j-2}, \ldots, w_{j-n+1})$$

(3)

We assume that the text corpus is annotated using ratings in the interval $[I, K]$ where $I$ indicates the lowest degree (“absence of behavior”) and $K$ indicates the highest degree (“strong presence of behavior”). First, we partition the corpus into $A_i$ and $\overline{A}_i$ such that $A_i$ contains text whose annotations lie in the interval $[I, i+I]$ while $\overline{A}_i$ contains those in the interval $(i+I, K]$. Then, we train N-gram LMs on $A_i$ and $\overline{A}_i$ to obtain a binary classifier, denoted as the $i^{th}$ pair. We repeat this for all integer-thresholded partitions $\{A_i, \overline{A}_i\} \forall i \in \{1, K-2\}$, shown in Table 2, to obtain $K$-2 binary classifier pairs.
| Pair \ Rating | 1 - 2 | 2 - 3 | \ldots | (K-1) - K |
|--------------|------|------|---------|----------|
| 1st          | $A_1$ | $A_1$ |         |          |
| 2nd          | $A_2$ | $A_2$ |         |          |
| \ldots       |       |       |         |          |
| (K-2)th      | $A_{K-2}$ | $A_{K-2}$ |       |          |

Table 2. Integer-thresholded partitions for behavior ratings ranging from 1 to K

For a text sequence $W$, whose behavior score, $x$, we want to find, the $i$th pair provides likelihoods $P_{A_i}(W)$ and $P_{\bar{A}_i}(W)$:

$$P_{A_i}(W) = P(W \mid 1 \leq x \leq i+1)$$  \hspace{1cm} (4)

$$P_{\bar{A}_i}(W) = P(W \mid i+1 < x \leq K)$$  \hspace{1cm} (5)

From these, we calculate the posterior probability of each cumulative interval using Bayes Rule, as shown in Eq. [6]. For the prior probabilities, we do not use any domain-specific knowledge about the behaviors of interest, instead assuming them to be uniformly distributed i.e. $X \sim \text{Uniform}(1, K)$. We then compute posteriors for each adjacent interval using Eq. [7] resulting in a probability mass function of the behavior score:

$$P(x \leq i + 1 \mid W) = P(1 \leq x \leq i+1 \mid W)$$

$$= \frac{P_{A_i}(W)P(1 \leq x \leq i+1)}{P_{A_i}(W)P(1 \leq x \leq i+1) + P_{\bar{A}_i}(W)P(i+1 < x \leq 9)}$$  \hspace{1cm} (6)

$$P(i < x \leq i+1 \mid W) = P(x \leq i+1 \mid W) - P(x \leq i \mid W)$$  \hspace{1cm} (7)

Since the $i$th point in Eq. [7] denotes the interval $[i,i+1]$, we represent it using its midpoint $i + \frac{1}{2}$. Finally, the behavior score $x$ of text sequence $W$ is computed as simply the expected value:

$$x = \sum_{i=1}^{K-1} \left( i + \frac{1}{2} \right) P(i < x \leq i+1 \mid W)$$  \hspace{1cm} (8)

**Training**

Similar to previous work (Georgiou et al., 2011; Chakravarthula et al., 2015b), we implement Maximum Likelihood models through 3-gram LMs trained with Good-Turing discounting using the SRILM toolkit (Stolcke, 2002). A leave-one-couple-out scoring scheme is used where models are trained on data from $N-1$ couples and subsequently used to score data from the $N$th couple. Ideally, we would train and test a separate model at each window length; for example, the N-gram model at window length 100 would be trained on sequences of 100 consecutive words. However, this is infeasible due to the curse of dimensionality in N-gram models, where the amount of data required for training increases exponentially with the order N (Bengio et al., 2003). Therefore, we train a single 3-gram model, i.e. at window length 3 words, and use the same model for testing at all five window lengths. As we show in Sec. 6.1.3, this difference between training and testing window lengths does not result in automatic degradations at longer window lengths.

**4.2.2 Neural Model Estimation**

**Model**

We use a recurrent modeling framework similar to the one used by Tseng et al. (Tseng et al., 2016) for classifying Negative behavior from language, but with a few changes: the Long Short-Term Memory (Hochreiter and Schmidhuber, 1997) unit is replaced with a Gated Recurrent Unit (GRU) (Cho et al., 2014) since it is similar but with fewer trainable parameters. In addition, we replace the word2vec (Mikolov et al., 2013) embeddings with ELMo (Peters et al., 2018) since it offers the advantages of deeper, contextual and character-level representations. Finally, while (Tseng et al., 2016) post-processed the system outputs using Support Vector Regression, we do not use
such transformations since we are interested in analyzing the properties of the system outputs themselves.

At runtime, given a sequence of $K$ words $W = \{w_j, w_{j+1}, \ldots, w_{j+K-1}\}$ words from the $j^{th}$ observation window, we pass their embeddings $M \in \mathbb{R}^{K \times U}$ through a GRU to obtain a fixed-length hidden representation $h_K \in \mathbb{R}^V$ for that window, where $U$ and $V$ are the dimensions of ELMo and the GRU hidden representation respectively. The representation $h_K$ is then passed through a fully connected layer, followed by a ReLUn (Krizhevsky and Hinton, 2010) layer, resulting in a scalar value that represents the behavior score of the $j^{th}$ window. We use the ReLUn layer in order to ensure that our predictions are bounded, similar to the ground truth annotations $A$. Our neural estimation model is shown in Figure 3.

![Figure 3. Neural model that estimates the behavior score of a sequence of words in a $K$-length observation window](image)

- **Training**

  We trained and tested our Neural model at two window lengths: 3-word (i.e. 3 unrolled time steps), representing a short window, and 30-word (30 unrolled time steps), representing a medium-length window, with zero padding performed at the end as required. We also attempted to train at long windows (50 and 100 words) but were unable to obtain well-trained, converged models for all the behaviors; hence, we only show results for 3-word and 30-word in this paper. A separate model was trained for each behavior construct, without any shared parameters or layers, so as to ensure that the results for each construct would be indicative of that construct only.

  For every word, ELMo provides embeddings from 3 layers, each of size $U = 1024$, and their corresponding mixing weights; we obtain a single 1024-dimensional embedding through a scaled, weighted sum of all 3 embeddings as recommended by Peters et al. (Peters et al., 2018). These mixing weights are softmax-normalized, similar to attention (Graves, 2013) weights, and, along with a scaling factor, were trained using backpropagation, similar to all other parameters. The size of the GRU hidden representation $V$ was tuned to be either 10 or 100 and the sample minibatch size was set to 64. Dropout (Srivastava et al., 2014) of 0.2 was applied before the fully connected layer and we used $L1$ loss in conjunction with Adam (Kingma and Ba, 2015) optimizer to train our models.

  We initially used grid search for tuning the learning rate but it was observed that the rates at which training remained stable varied greatly across different behaviors. Addressing this in a task-agnostic manner would necessitate a large search space and exacerbate the time and computational cost of training. Hence, we replaced the grid search with a task-specific tuning scheme; specifically, we implemented the Cyclic Learning Rate schedule as proposed by Smith (Smith, 2017). For each behavior and model configuration, we first performed a “range test” to determine the minimum and maximum learning rates at which training remained stable. We then cyclically varied the learning rate between its minimum and maximum value during training, saving a model checkpoint at the
end of every epoch when the learning rate would be at its lowest. Finally, at testing time, inspired by Huang et al. (Huang et al., 2017), instead of using just the last or the best checkpoint, we used an ensemble average of all of them.

To further reduce the time and computational costs of training, we employed a 6-fold nested cross-validation setup where in every test fold, four folds were used to train the model while the fifth fold was used to optimize the model hyper-parameters, learning rate range, etc. Similar to the setup with N-gram models, we ensured that no dyad appeared in more than one fold.

4.3 Analysis

4.3.1 Metrics

1. **Behavior Construct Similarity**

   Our first proposed metric, the Behavior Construct Similarity (BCS), is a direct representation of Eqn. [1] with Spearman’s Correlation \( R \) as the choice of the evaluation metric \( C \). For behavior \( B_i \) and window length \( L_j \), it is computed as:

   \[
   BCS_i(j) = R(F(M(E_{lang}(B_i), L_j, D; \theta)), A_i) \tag{9}
   \]

   BCS measures how similar the system’s interaction-level scores are to the ground truth annotations, over all the data samples; the higher the value, the more similar the two are. Hence, by observing how it varies with window length, we can obtain an idea of which ones are appropriate for each behavior. The procedure for computing BCS is detailed in Algorithm 1.

   **Algorithm 1 Behavior Construct Similarity**

   ```
   1: for each behavior \( B_i \) in \( B \) do
   2:     for each window length \( L_j \) in \( L \) do
   3:         for each functional \( F_k \) in \( F \) do
   4:             Initialize empty lists \( G, H \)
   5:             for each interaction \( T_j \) in \( T \) containing \( O_l \) words do
   6:                 Score \( T_j \) at window length \( L_j \) to get trajectory of scores \( S = \{S_1, S_2, \ldots, S_{max(1, O_l-L_j+1)}\} \)
   7:                 Compute aggregate score \( F_k(S) \)
   8:                 Append \( F_k(S) \) to \( G \)
   9:                 Append ground-truth annotation \( A_i \) to \( H \)
   10:         end for
   11:     end for
   12:     Compute Spearman Correlation \( R_k(j) \) between \( G \) and \( H \)
   13: end for
   14: \( BCS_i(j) = \{R_k(j) \forall k\} \)
   15: end for
   ```

   We use Spearman’s Correlation since it captures monotonic similarity and is, thus, apt for relating together ordinal variables such as the ground truth annotations \( A_i \). For the functional \( F \), we test 3 statistics: *median*, *minimum* and *maximum*; the former because it has been shown to be useful for summarizing behaviors such as *Negative* in previous works (Tseng et al., 2016, 2018) and the latter two since they can capture characteristics such as frequent/rare expression, large/small fluctuations, etc. We did not use the *mean* because in some instances, we observed that the system’s window-level scores were \( \alpha \)-stable distributed with \( \alpha \leq 1 \), for which the *mean* is undefined (Nolan, 2003); the other three statistics, however, are still defined.

2. **Behavior Relationship Consistency**

   The BCS metric performs a direct validation of the window-level scores by evaluating their aggregate against ground-truth annotations. However, this means that it is limited by the choice of the functional \( F \) used for aggregation, since a poor choice of \( F \) can result in inaccurate results and, hence, low BCS, even if a sufficiently long window is used. In such a scenario, relying on BCS alone would lead to an incorrect conclusion that the window length is still not sufficient. One way to resolve this would be to perform a similar evaluation at the window-level, without any aggregation,
but it is not possible since we do not have ground truth annotations at the window-level to compare against.

Instead, we propose to evaluate predictions of constructs based on whether they are related to each other the same way as the constructs themselves. This is motivated by previous work by Thornton et al. (Thornton and Tamir 2017) who, while studying emotional transitions, found that similar emotions (e.g. Anger and Disgust) frequently tend to co-occur whereas dissimilar ones (e.g. Anger and Joy) do not. Extending this principle to our scenario, we argue that if our system is indeed able to accurately quantify related behaviors, then we should be able to observe those same relations in the predictions as well. The more consistent the relations between the predictions are with the relations between the actual behaviors, the more accurate we can consider the window-level predictions, and hence the more appropriate the window length, to be.

We formalize the above intuition in the form of our second proposed evaluation metric, the Behavior Relationship Consistency (BRC). The BRC metric is defined for a pair of constructs and measures how close the Spearman Correlation between their ground-truth annotations is, over all the data samples, to the Spearman Correlation between their window-level scores; the full procedure is detailed in Algorithm 2. For a pair of behaviors $B_i$ and $B_j$ and window length $L_k$, it is calculated as:

$$BRC_{i,j}(k) = 1 - \frac{|Q^*_i - Q^*_j|(k)|}{2}$$  \hspace{1cm} \text{(10)}

where $Q^*_i = R(A_i, A_j)$

and $Q^*_j = R(M(E_{lang}(B_i), L_k, D; \theta), M(E_{lang}(B_j), L_k, D; \theta))$

\begin{algorithm}
\caption{Behavior Relationship Consistency}
\begin{algorithmic}[1]
\State \textbf{for} each behavior $B_i$ in B \textbf{do}
  \State \hspace{1cm} \textbf{for} each behavior $B_j$ in B \textbf{do}
  \State \hspace{2cm} \textbf{for} each window length $L_k$ in L \textbf{do}
  \State \hspace{3cm} Initialize empty lists $C_i$, $C_j$, $G_i$, $G_j$
  \State \hspace{3cm} \textbf{for} each interaction $T_i$ in T containing $O_i$ words \textbf{do}
  \State \hspace{4cm} Score $T_i$ at window length $L_k$ to get scores
  \State \hspace{4cm} $S^i_{T_i} = \{S^i_{1}, S^i_{2}, \ldots, S^i_{\max(1,O_i-L_k+1)}\}$ for $B_i$
  \State \hspace{4cm} $S^j_{T_j} = \{S^j_{1}, S^j_{2}, \ldots, S^j_{\max(1,O_j-L_k+1)}\}$ for $B_j$
  \State \hspace{4cm} Append $S^i_{T_i}$ to $C_i$, $S^j_{T_j}$ to $C_j$
  \State \hspace{3cm} Append ground-truth annotation $A^i_j$ to $G_i$, $A^j_i$ to $G_j$
  \State \hspace{2cm} \textbf{end for}
  \State \hspace{1cm} Compute Spearman Correlations $R_C(i, j)$ between $C_i$ and $C_j$, $R_G(i, j)$ between $G_i$ and $G_j$
  \State \hspace{1cm} $BRC_{i,j}(k) = 1 - \frac{|R_C(i, j) - R_G(i, j)|}{2}$
  \State \hspace{1cm} \textbf{end for}
\State \textbf{end for}
\State \textbf{end for}
\end{algorithmic}
\end{algorithm}

It can be seen that $Q^*_j$ in Eqn. 10 is similar to $Q_i$ from Eqn. 1; both evaluate the window-level scores of behavior $B_i$, but the former evaluates against predictions of behavior $B_j$ whereas the latter evaluates against the ground truth annotations of behavior $B_i$. Naturally, this means that the effectiveness of BRC for $B_i$ is directly dependent on the accuracy of predictions of $B_j$; the more accurate they are, the more effective it will be. For this reason, BRC is an indirect validation of the window-level scores, unlike BCS. However, BRC can be computed directly on the window-level scores without involving any kind of aggregation; hence, it does not suffer from the same limitation as BCS. In the following section, we describe how we perform our analysis using both BCS and BRC in a manner that combines their advantages while mitigating their disadvantages.

\subsection*{4.3.2 Procedure}

Now that we have defined the metrics with which to analyze behaviors, we proceed to employ them in the following multi-stage fashion, depicted in the flowchart in Figure 4.

\begin{itemize}
\item \textbf{Step 1:}
\item \textbf{Step 2:}
\item \textbf{Step 3:}
\item \textbf{Step 4:}
\item \textbf{Step 5:}
\item \textbf{Step 6:}
\item \textbf{Step 7:}
\item \textbf{Step 8:}
\item \textbf{Step 9:}
\item \textbf{Step 10:}
\item \textbf{Step 11:}
\item \textbf{Step 12:}
\item \textbf{Step 13:}
\item \textbf{Step 14:}
\item \textbf{Step 15:}
\item \textbf{Step 16:}
\end{itemize}
Figure 4. Flowchart depicting the step-by-step procedure through which we analyze the appropriate window lengths for a target behavior \( B_i \). First, in stages 1 and 2, we examine whether the BCS metric, which describes how well \( B_i \)’s predictions by the system match its annotations by human experts, contains clues about window length. If not, then in stages 3 and 4, we check whether the BRC metric, which describes how much the system’s predictions resemble human expert annotations in terms of \( B_i \)’s relations with other behaviors, can provide the required information. If not, we make no determinations about \( B_i \)’s window length. Blue boxes represent standard operations while green boxes represent final conclusions from our analysis.

- **Stage 1:** We start by identifying behaviors for which all the factors listed in Sec. 3, such as window length, functional, etc. are appropriate. For such behaviors, their estimates by the system are highly correlated with human ratings at all window lengths. Therefore, given behavior \( B_i \) and a set of window lengths \( L \) being tested, we check whether its BCS, which was defined in Eqn. 9, is greater than a threshold \( Y_1 \) at all lengths. The higher this threshold \( Y_1 \), the better estimated \( B_i \) is considered to be; in our work, we used \( Y_1 = 0.59 \) since the highest BCS that both of our models achieved was around 0.6.

\[
\text{Identify } B_i: BCS_i(j) > Y_1 \forall j \in L
\]

We refer to these behaviors as “reference” behaviors, denoted using \( B_{ref} \). Since they are considered to be estimated well at all window lengths, we set the appropriate window length to be the shortest one in \( L \).

\[
\text{Window length } L_i = \min \{ L \}
\]
• **Stage 2:** In Stage 2, we focus on behaviors whose BCS did not exceed $Y_1$, i.e. they were either low or medium. This could be attributed to multiple factors - inappropriate window length, ill-suited model, insufficient expression of behavior in language, etc. However, with all other factors kept constant, any change in BCS must be solely due to the effect of the window length. Therefore, in this stage, we observe how the BCS of a behavior changes over the window lengths in $L$ and pick the one at which it is maximum. Specifically, for each behavior in Stage 2, we check if its maximum BCS is significantly larger than its minimum BCS.

$$\text{Identify } B_i : \max \{ BCS_i(j) \} > \min \{ BCS_i(j) \} \quad \text{(significant, } p < 0.05)$$

Window length $L_i = \arg \max_j BCS_i(j)$

We check statistical significance by following the recommendations outlined by Diedenhofen et al. [Diedenhofen and Musch, 2015]. Since we are comparing correlations between ground truth annotations and system estimates from the same data at different window lengths, they are considered to be dependent overlapping correlations. Accordingly, we calculate the 95% confidence interval for differences in dependent overlapping correlations using Zou’s method [Zou, 2007]: the change in BCS from minimum to maximum is significant if the interval does not contain 0; else, it is not significant.

• **Stage 3:** For those behaviors $B_i$ whose BCS neither exceeded $Y_1$ nor significantly changed with window length, we instead examine whether their BRC, which was defined in Eqn. 10, is high at any window length. Specifically, we check if it is higher than a threshold $Y_2$ and pick the shortest window length at which this is true.

$$\text{Identify } B_i : \exists k \in L : BRC_i(k) > Y_2$$

Window length $L_i = \min \{ k \in L : BRC_i(k) > Y_2 \}$

A value of BRC greater than $Y_2$ indicates that $B_i$ exhibits similar relations in the system’s scores as in the ground truth annotations. $Y_2$ should be as close to 1 as possible, since we would like our system’s scores to exhibit exactly the same relations as human annotations; in our work, we use $Y_2 = 0.95$.

$BRC_i(k)$ is computed with respect to the $B_{ref}$ behaviors that were identified in Stage 1 since they are well-estimated and are, thus, more reliable than other behaviors. In cases where $B_{ref}$ consists of multiple behaviors, we compute $BRC_i(k)$ as a weighted sum of individual BRCs, as shown below:

$$BRC_i(k) = \sum_{B_j \in B_{ref}} \alpha_i(k) BRC_j(k)$$

where

$$\alpha_i(k) = \frac{BCS_i(k)}{\sum_{B_m \in B_{ref}} BCS_m(k)}$$

We see that for every behavior in $B_{ref}$, its BRC with $B_i$ is scaled by a normalized weight that is proportional to its BCS. This is done in order to assign greater importance to better estimated behaviors than worse ones, thereby increasing the reliability of the resulting BRC metric.

• **Stage 4:** This is the last stage of our analysis and the behaviors we are left with are those with low and unchanging BCS as well as low BRC. While we could not identify appropriate window lengths (at which BRC is very high), it is nevertheless useful to understand which lengths are better or worse than others. Therefore, similar to Stage 2, we examine whether the BRC of $B_i$ shows a tendency to increase or decrease as window length changes and identify where it is maximum.

$$\text{Identify } B_i : \max_k \{ BRC_i(k) \} > \min_k \{ BRC_i(k) \} \quad \text{(significant, } p < 0.05)$$

Window length $L_i = \arg \max_k BRC_i(k)$

Since we are comparing correlations between system estimates of different behaviors at different window lengths, these fall under the category of independent groups [Diedenhofen and Musch, 2015]. Similar to Stage 2, we check statistical significance using 95% confidence intervals as described in [Zou, 2007], but this time, for differences between independent correlations.
• **End:** For behaviors that do not show a significant change in BRC in of Stage 4, we do not analyze them any further and simply conclude that we are unable to make any determinations at this point about their window length properties. In the next section, we describe the data corpus on which we apply our analysis.

### 5 DATASET

The Couples Therapy project (Christensen et al., 2004) involved 134 real-life chronically distressed couples attending marital therapy over a period of up to 1 year. Its dataset consists of hundreds of realistic interactions as well as a rich and diverse set of annotations characterizing the behavior of the participants in these interactions. This offers an attractive and challenging domain for behavior analysis, which is why we select it as the area of focus for this paper. We note, though, that the analysis and techniques presented in this work can be also applied to other domains involving machine learning for human-derived and human-centered attributes, such as emotion and sentiment.

| **CIRS 2 Code** | **Description** |
|-----------------|-----------------|
| Acceptance      | Indicates understanding, acceptance, respect for partner’s views, feelings and behaviors |
| Perspective     | Tries to understand partner’s views, feelings by clarifying and asking to hear them out |
| Responsibility  | Implies self-power over feelings, thoughts, behaviors on issue being discussed |
| External        | Softens criticism of partner by attributing their undesirable behaviors to external origins |
| Define          | Articulates problems clearly, facilitates everyone’s participation in problem solving process |
| Solution        | Suggests specific solutions that could solve the problem |
| Negotiates      | Offers compromises or bargains |
| Agreement       | States terms of agreement, willingness to follow them with partner |
| Blame           | Blames, accuses, criticizes partner and uses critical sarcasm and character assassinations |
| Change          | Requests, demands, nags, pressures for change in partner |
| Withdrawal      | Generally non-verbal, becomes silent, refuses to respond, discuss, argue, defend |
| Avoidance       | Minimizes importance and denies existence of problem, diverts attention, delays discussion |
| Discussion      | Discusses problem, shows engagement, interest and willingness in discussing issue |

| **SSIRS Code** | **Description** |
|----------------|-----------------|
| Positive       | Overtly expresses warmth, support, acceptance, affection, positive negotiation |
| Negative       | Overtly expresses rejection, defensiveness, blaming, and anger |
| Anger          | Expresses anger, frustration, hostility, or resentment during the interaction |
| Belligerence   | Quarrels, argues, verbalizes nasty comments and mean rhetorical questions |
| Disgust        | Shows disregard, scorn, lack of respect and makes patronizing and insulting comments |
| Sadness        | Cries, sighs, speaks in a soft or low tone, expresses unhappiness and disappointment |
| Anxiety        | Expresses discomfort and stress, answers with short yes/no responses without elaboration |
| Defensiveness  | Deflects criticism by defending self, accusing partner of similar behavior |
| Affection       | Expresses warmth and caring for partner, speaks warmly, uses endearments |
| Satisfaction   | Feels satisfaction about how topic of discussion is defined, discussed, and resolved |
| Dominance      | Commands course of interaction, dominates conversation, changes subject frequently |
| Solicits Suggestions | Shows interest in and seeks partner’s suggestions, help in handling issue |
| Instrumental Support | Offers positive advice for clear, concrete actions to support partner |
| Emotional Support | Emphasizes feelings, builds confidence, and raises self-esteem in partner |

Table 3. Description of behavior codes in Couples Therapy corpus

#### 5.1 Description of Corpus

The corpus consists of audio and video recordings with manual transcriptions of each couple discussing topics of marital distress in 10-minute interactions. Each couple had at least 2 interactions or “sessions”, once with each participant leading the discussion on a topic of their choice, and the total number of sessions per couple ranged from 2 to 6. In each session, both the husband and the wife were rated for a total of 31 CIRS (Heavey et al. 2002) and SSIRS (Jones and Christensen, 1998) behaviors by trained human annotators with a sense of what “typical” behavior is like during these interactions. The annotators were asked to observe both verbal and nonverbal expressions when rating each behavior independently and in many cases, different annotators rated different behaviors. Each behavior in each session was rated by 3 to 9 annotators, with most of them being rated by 3 to 4. The rating was done on a Likert scale from 1 to 9, where 1 represents “absence of behavior”, 5 represents “some presence of behavior” and
9 represents “strong presence of behavior”. More details about the recruitment, data collection and the annotations can be found in (Christensen et al., 2004; Baucom et al., 2011).

Consistent with previous work (Lee et al., 2010; Georgiou et al., 2011; Black et al., 2013; Lee et al., 2014; Tseng et al., 2017), for each participant and behavior, we take the average of the annotators' ratings as the true rating in that session. Therefore, each speaker’s data sample consists of the manual transcription of their utterances and their behavior ratings in that session. Of the 31 behaviors that were rated, we discard 4 of them, such as “Is the topic of discussion a personal issue?” and “Is the discussion about husband’s behavior?”, since they are tied more to the topic of interaction and less to the speaker’s behavior. The resulting set of 27 behaviors that will be analyzed in this work are listed in Table 3 and categorized as follows:

- **Couples Interaction Rating System 2 (CIRS2):** This set contains 13 codes that describe a speaker when interacting with their partner about a problem.

- **Social Support Interaction Rating System (SSIRS):** This set contains 14 codes that measure emotional features and ratings of the interaction.

We only consider those sessions where both the speaker and the partner are rated for all 27 behaviors, resulting in 1325 data samples. Since the content and nature of interaction vary from one couple to another, the number of words spoken by a person during the whole session ranges from around 50 to 2500, as shown in Figure 5.

![Figure 5. Histogram of number of words per speaker per session](image)

**5.2 Behavioral Grouping**

Based on the descriptions of these behaviors, we can see clear relations between some of them; for instance, Negative is similar to Blame and Anger but opposite to Positive and Affection. Withdrawal is similar to Avoidance but opposite to Discussion, etc. These notions of similarity and polarity arise as a result of how humans perceive them. Using them, we can identify different groups of behaviors, with each group representing a concept that is common to the behaviors in it. Such a grouping would be beneficial since it can lend more interpretability to our analysis; for example, we can investigate a link between the common trait of a group and the number of words required to express the behaviors in it. Hence, we identify groups of related behaviors, shown below, using k-means clustering as described in Algorithm 3.

Figure 6 shows the clustering of the 27 Couples behaviors based on how similarly they were rated by human annotators, as described in Sec. 5.2. Each cell shows the Spearman Correlation for a pair of behaviors. As expected, we observe that Negative is positively correlated with Blame and Anger and negatively correlated with Positive and Affection. Similarly, Withdrawal is positively correlated with Avoidance and negatively correlated with Discussion. On the other hand, some behaviors do not appear to be strongly related to each other, such as the pairs (Sadness, Positive) and (Disgust, Negotiates).

We tested different numbers of clusters and found that 4 clusters provided the most coherent grouping. This also matches a different study by Sevier et al. (Sevier et al., 2008) on the same dataset, which derived 4 scales of behavior using a Principal Component Analysis-based approach: Negativity, Withdrawal, Positivity and Problem-Solving. Hence, we name our 4 groups of behaviors in similar fashion: The
Algorithm 3 Similarity-based Grouping of Behaviors

1: Calculate $R_{global} = R_G$ as in Algorithm 2
2: for number of clusters $N$ in \{2, 3, \ldots, K\} do
3:     for $D$ random initializations do
4:         Run K-Means clustering on $R_{global}$ with $N$ clusters, store clustering $C^D_N$
5:     end for
6:     Pick most frequent clustering $C_N = \arg \max_D \text{Count}(C^D_N)$
7: end for
8: Pick Behavior Grouping $C = \arg \min_N S_N$

Figure 6. Behavior Grouping: Clustering of Couples Therapy behaviors into 4 groups based on their similarity to each other. The cell in the $(i, j)$ position shows the Spearman Correlation between human ratings of the corresponding behaviors $i$ and $j$. Yellow (Blue) indicates highly positive (negative) correlation. Non-diagonal gray cells indicate that correlation is not statistically significant ($p < 0.05$)

first group contains behaviors such as Discussion, Negotiates, Responsibility and Solutions. Since these are strongly tied to a back-and-forth problem-specific style of interaction, we refer to them as Problem-Solving behaviors. The second group consists of behaviors such as Anger, Blame, Disgust and Negative; hence, we refer to it as Negative. Similarly, we name the third group Positive since it contains Affection, Positive, Satisfaction, etc. The last group contains behaviors such as Anxiety, Change, Sadness and Withdrawal. Since a majority of these are related to “dysphoria”, which is a state of unease or unhappiness, we call this last group Dysphoric.

6 RESULTS & DISCUSSION

In this section, we present the results of our analysis on the Couples Therapy dataset. We first discuss the findings from each model separately following which we comment on behavior characteristics that are shared across both models. Finally we review our findings with respect to how they address the questions posed in Sec. I and how they compare with findings from other related work.
6.1 N-gram Model

6.1.1 Analysis of BCS

Figure 7 shows the BCS of the N-gram model scores at the 5 observation window length values that were tested - 3, 10, 30, 50 and 100 words. Every behavior is represented by a trajectory and each point on the trajectory represents the Spearman Correlation between ground truth annotations and the aggregated model scores at that window length. While we tested three functionals for aggregation - minimum, median and maximum - we only use the one that performed best, on average, across all window lengths for our analysis. Hence, we only show the best performing functional for each behavior in the BCS plot; all correlations were found to be statistically significant ($p < 0.05$).

![Behavior Construct Similarity (BCS)](image)

**Figure 7.** Behavior Construct Similarity (BCS): Spearman Correlation between human annotations and functional-aggregated scores of N-gram model at different observation window lengths. All correlations were found to be statistically significant ($p < 0.05$)

The best performing behaviors with the N-gram model are Acceptance and Blame, with BCS values greater than 0.6 at nearly all window lengths; hence, we use these as reference behaviors during our analysis. With respect to behavior groups, we see that the Negative behaviors are, on average, the best estimated ones, followed by Positive behaviors. The BCS for Problem-Solving behaviors varies from moderate (0.45 for Solutions) to extremely low (0.06 for External). Finally, with the exception of Change, the BCS for Dysphoric behaviors is, in general, extremely low. This matches previous studies which have found that behavioral constructs related to negative and positive affect tend to be estimated well from low-level lexical features. From these results, we can now also see that they are, in fact, much better estimable than higher-level and more complex behaviors related to dysphoria and problem-solving. This could be due to factors such as these behaviors not being expressed sufficiently in language or their expression in language, even if sufficient, being too complex to be modeled using N-gram phrases and simple statistics.

In evaluating the choice of functionals, median appears to be the best aggregation method for nearly every behavior except for a few such as dominance and withdrawal, for whom maximum and minimum respectively work better. This agrees with previous findings (Tseng et al., 2016, 2018) which also found the median to be a useful measure of aggregate behavior. In cases where median is the best functional, we see that the BCS does not change much even from varying from the shortest window length to the longest one possible. This, however, does not imply that all window lengths are equally appropriate for such behaviors. As shown in Figure 7, the scores from the N-gram model tend to be symmetrically distributed, a pattern which was also reported in (Tseng et al., 2016). As a result, any change in scores resulting from changes in the window length would not be reflected by the median and, hence, the BCS would not change, giving the false impression that all windows are equally appropriate. Therefore, in order to gain more insights into the effect of window length on behavior quantification, we proceed to examine the
6.1.2 Analysis of BRC
Table 4 displays the Spearman Correlation between behaviors in the ground-truth annotations and in the N-gram scores at different window lengths. From these correlations, we calculate the BRC for each behavior using Eqns. 10 and 11. The Ratings and Scores from Table 4 serve as $Q^*$ and $Q'$ in Eqn. 10 and since we have 2 reference behaviors, Acceptance and Blame, we use Eqn. 11 to calculate the weighted BRC where the Reference Weights serve as the normalized weights $\alpha$. This now enables us to observe changes in the quality of the N-gram model scores which were otherwise not reflected in the BCS. For instance, for the behavior Solutions, the BCS is nearly the same, around 0.45 at all the window lengths. However, when compared to its ground truth correlations with Acceptance and Blame (0.282 and -0.184 respectively), we see that the N-gram score correlations at 50 words (0.37 and -0.153 respectively) are more similar to them than those at 3 words (0.323 and 0.006 respectively). Therefore, based on this, we can conclude that a window length of 50 words is more appropriate for scoring the behavior Solutions than 3 words. In this manner, we perform our analysis using both BCS and BRC and obtain the final set of results which are displayed in Figure 9. Each behavior is shown against its appropriate window length and arranged by behavior groups so as to also demonstrate group-level patterns.

6.1.3 Analysis of window lengths
We see that a majority of behaviors tend to perform best at short window lengths, typically fewer than 10 words, with nearly half of them at just 3 words. Nearly a third of the behaviors perform best when scored using long observation windows, such as Solutions at 50 words and Avoidance at 100 words. Among the behavior groups, we see that Negative behaviors, on average, appear to require much shorter observation windows than the rest. This matches our intuition about the emotional, short-term nature of these behaviors that lends itself to brief expressions.

The remaining three groups consist of behaviors that are expressed over a range of lengths, from short (Acceptance) to long ones (Avoidance). Among these, Positive and Dysphoric behaviors mostly work best at short observations (10 words or fewer). For Positive behaviors, this is likely due to their affective content which, while not as brief as negative ones, is nevertheless short-term. Dysphoric behaviors, on the other hand, are characterized by a lack of participation and expression and are thus, likely to be marked by brief expressions, which could be why they tend to do best at short window lengths.

Finally, Problem-Solving behaviors are evenly split between either very short (3 words) or very long windows (50 - 100 words). This is a little surprising since we would normally expect them to be

![Figure 8. Sample distribution of Dominance and External scores at window lengths 3, 100 and session. In both behaviors, the median 3-gram as well as 100-gram scores are very similar to the session-level scores, possibly as a result of the nearly-symmetrical distribution of scores.](image-url)
| Reference Behavior | Acceptance | Blame |
|--------------------|------------|-------|
|                    | Reference Weight |        |        |
|                    | 0.501 | 0.507 | 0.516 | 0.514 | 0.516 | 0.499 | 0.493 | 0.484 | 0.486 | 0.484 |
| Target Behavior    |        |        |        |        |        |        |
| Discussion         | 0.178 | -0.007 | 0.015 | 0.018 | 0.012 | 0.005 | 0.085 | -0.21 | -0.191 | -0.174 | -0.161 | -0.146 |
| External           | 0.256 | 0.181 | 0.166 | 0.192 | 0.216 | 0.256 | -0.073 | 0.212 | 0.176 | 0.12 | 0.08 | 0.015 |
| Negotiates         | 0.237 | 0.199 | 0.198 | 0.227 | 0.252 | 0.292 | -0.083 | 0.181 | 0.138 | 0.076 | 0.033 | -0.036 |
| Perspective        | 0.099 | 0.068 | 0.041 | 0.021 | 0.006 | -0.022 | - | 0.141 | 0.124 | 0.123 | 0.127 | 0.139 |
| Responsibility     | 0.328 | 0.198 | 0.221 | 0.257 | 0.284 | 0.329 | -0.239 | 0.176 | 0.102 | 0.029 | -0.016 | -0.092 |
| Solicit-suggestions| 0.379 | 0.297 | 0.305 | 0.349 | 0.381 | 0.434 | -0.23 | 0.083 | 0.013 | -0.067 | -0.116 | -0.195 |
| Solutions          | 0.282 | 0.323 | 0.326 | 0.351 | 0.37 | 0.397 | -0.184 | 0.096 | -0.057 | -0.117 | -0.153 | -0.207 |
| Anger              | -0.653 | -0.499 | -0.564 | -0.627 | -0.66 | -0.71 | 0.673 | 0.736 | 0.739 | 0.762 | 0.778 | 0.806 |
| Belligerence       | -0.646 | -0.483 | -0.538 | -0.598 | -0.632 | -0.684 | 0.677 | 0.738 | 0.735 | 0.756 | 0.773 | 0.8 |
| Defensiveness      | -0.564 | -0.446 | -0.487 | -0.548 | -0.585 | -0.644 | 0.522 | 0.639 | 0.635 | 0.666 | 0.689 | 0.729 |
| Disgust            | -0.66 | -0.424 | -0.485 | -0.545 | -0.579 | -0.632 | 0.69 | 0.707 | 0.708 | 0.729 | 0.745 | 0.773 |
| Negative           | -0.729 | -0.656 | -0.708 | -0.757 | -0.782 | -0.819 | 0.693 | 0.774 | 0.784 | 0.809 | 0.825 | 0.852 |
| Affection          | 0.582 | 0.283 | 0.299 | 0.348 | 0.384 | 0.44 | -0.352 | 0.152 | 0.081 | -0.004 | -0.058 | -0.143 |
| Agreement          | 0.347 | 0.269 | 0.259 | 0.284 | 0.307 | 0.34 | -0.295 | 0.118 | 0.072 | 0.008 | -0.032 | -0.097 |
| Define             | 0.517 | 0.439 | 0.497 | 0.556 | 0.59 | 0.646 | -0.353 | -0.536 | -0.552 | -0.594 | -0.622 | -0.674 |
| Positive           | 0.67 | 0.575 | 0.619 | 0.674 | 0.705 | 0.753 | -0.547 | -0.201 | -0.297 | -0.39 | -0.44 | -0.522 |
| Satisfaction       | 0.563 | 0.456 | 0.486 | 0.538 | 0.572 | 0.624 | -0.537 | -0.095 | -0.182 | -0.27 | -0.32 | -0.401 |
| Support-emotional  | 0.492 | 0.239 | 0.257 | 0.301 | 0.333 | 0.386 | -0.326 | 0.181 | 0.119 | 0.042 | -0.008 | -0.086 |
| Support-instrumental | 0.46 | 0.32 | 0.342 | 0.397 | 0.432 | 0.487 | -0.343 | 0.089 | 0.004 | -0.089 | -0.143 | -0.23 |
| Anxiety            | -0.299 | -0.107 | -0.146 | -0.182 | -0.201 | -0.233 | 0.17 | 0.425 | 0.408 | 0.413 | 0.417 | 0.426 |
| Avoidance          | -0.17 | 0.004 | -0.015 | -0.022 | -0.022 | -0.022 | 0.085 | 0.333 | 0.307 | 0.287 | 0.273 | 0.253 |
| Change             | -0.474 | -0.47 | -0.528 | -0.579 | -0.606 | -0.653 | 0.7 | 0.703 | 0.704 | 0.726 | 0.743 | 0.77 |
| Dominance          | -0.144 | -0.201 | -0.202 | -0.209 | -0.216 | -0.237 | 0.293 | 0.174 | 0.224 | 0.234 | 0.24 | 0.254 |
| Sadness            | -0.131 | -0.059 | -0.069 | -0.075 | -0.074 | -0.07 | 0.198 | 0.394 | 0.354 | 0.328 | 0.312 | 0.288 |
| Withdrawal         | -0.164 | 0.019 | - | 0.003 | 0.008 | 0.014 | - | 0.293 | 0.268 | 0.241 | 0.224 | 0.2 |

Table 4: Spearman Correlation between reference and target behaviors used to calculate the BRC metric (Eqn. 10) for the N-gram model. Ratings and Scores refer to ground-truth correlations \( Q \) and N-gram model score correlations \( Q' \) respectively in Eqn. 10. All correlations statistically significant \( (p < 0.05) \) unless marked as -.
Figure 9. Appropriate observation window lengths of behaviors as determined by our analysis when quantified using an N-gram model. Differences between model performance at window lengths were found to be statistically significant \((p < 0.05)\)

mostly, if not all, long-range due to their extended, back-and-forth nature. Upon inspection, we found that the 3-word-window behaviors in this group exhibited their highest BCS at that window length but their highest BRC at longer window lengths. This indicates that while the model could be capturing them better at longer windows, the aggregating functional performs better at shorter window lengths. Hence, for such behaviors, perhaps more sophisticated and better-suited functionals could exploit the full potential of using long window lengths.

Lastly, we note here that a-priori, it could have been conjectured that the N-gram model would always prefer 3-word or short windows over longer ones, simply by virtue of its 3-gram training. However, we see that this is not the case and that, in fact, more than half of the behaviors perform better at windows longer than 3 words when using the N-gram model. Therefore, this demonstrates that our findings are not merely an artifact of the nature of the model or training being used.

6.2 Neural Model

6.2.1 Analysis of BCS

Figure 10 shows the BCS of the Neural model scores at the two observation window lengths that were tested, 3 and 30 words. For each behavior, a bar represents the Spearman Correlation between ground truth annotations and the aggregated Neural model score at that window length. Similar to the N-gram model, we used the best performing, on average, functional for our analysis and display it for each behavior in the BCS plot; all correlations were found to be statistically significant \((p < 0.05)\).

The best performing behavior with the Neural model is *Blame*, with BCS values close to 0.6 at both window lengths; hence, it is used as the *reference behavior* in our analysis. Once again, we see that the best estimated behaviors belong to the *Negative* group, followed by *Positive*. Interestingly, however, we observe low-to-moderate BCS for both *Dysphoric* behaviors, which are generally subtle and non-verbal, as well as *Problem-Solving* behaviors, which are generally verbose. This shows that the contextual embeddings of ELMo in conjunction with the long-term, non-linear processing of the GRU are able to handle both scenarios’ diverse linguistic requirements equally well.

With respect to functionals, we see that the *median* is once again the best performing one, observed in almost all behaviors except for a few such as *Solutions* and *Withdrawal*, for which *maximum* and *minimum* respectively work better. We also see that the BCS for some behaviors changes noticeably as the window length increases from 3 to 30 words. Technically, this variation can be attributed to not just the change in window length but also the quality of the model trained at that window length. However, since we tuned for the best Neural model at each window length, we assume that the quality of training is similar across window lengths and that the variation in BCS is mostly due to the change in window length.
Figure 10. Behavior Construct Similarity (BCS): Spearman Correlation between human annotations and functional-aggregated scores of Neural model at different observation window lengths. All correlations were found to be statistically significant ($p < 0.05$)

6.2.2 Analysis of BRC

Table 5 shows the Spearman Correlations between behaviors in the ground truth annotations and Neural model scores. From these, we calculate the BRC using Eqn. 10, since we have just one reference behavior, (Blame). Then, using both BCS and BRC, we analyze the Neural model scores of all the behaviors, the results of which are shown in Figure 11. As before, each behavior is shown against its appropriate window length and arranged by behavior groups. Since we are comparing just two window lengths, 3 and 30 words, the resolution of our analysis here is slightly coarse and doesn’t necessarily reflect exhaustive trends. For instance, a behavior that actually requires 10-word windows might perform better at 30 words than at 3 words simply because of the increased context and not because it is best observed at 30 words. Hence, the window length results in the Neural model should be interpreted in a relative light, i.e. short window vs longer window.

6.2.3 Analysis of window lengths

From Figure 11, we can see that most of the behaviors perform best at the 30-word window length, with less than a quarter performing well with the 3-word window. For such behaviors, this indicates that the longer context is able to provide more useful information to the Neural model than what a short observation window could have. Among the behavior groups, we see, once again, that Negative and Positive behaviors require, on average, shorter observation windows than the other two groups. Interestingly, however, we see that the trend has slightly been altered; a smaller percentage of Negative behaviors perform better at short window lengths with the Neural model than the Positive behaviors. We inspected the analysis metrics of Negative behaviors to understand why this was happening, and found that while the BRC was higher at 3 words for most of them, their BCS was higher at 30 words. This suggests that the functional median is probably better suited for aggregating the Neural model scores at 30 words than at 3 words, and that, therefore, further investigation may be required to find a suitable functional at short window lengths.

In the Problem-Solving and Dysphoric groups, we see that their behaviors tend to perform, on average, better at longer window lengths than short ones. This is in line with our intuition and expectations about the nature of these behaviors being either subtle or marked by complex construction, both of which necessitate the use of medium-to-long window lengths in order to resolve the ambiguities involved. Furthermore, since our Neural model uses a GRU, which was originally designed to handle long-context dependencies such as the ones seen in these behaviors, it makes sense that it would work better when fed information from a longer observation window.
Reference Behavior | Blame
---|---
| Ratings | Scores (3) | Scores (30)
External | -0.073 | -0.18 | -0.134
Negotiates | -0.083 | -0.381 | -0.263
Perspective | - | 0.063 | 0.0902
Responsibility | -0.239 | -0.311 | -0.346
Solicit-suggestions | -0.23 | -0.388 | -0.42
Solutions | -0.184 | -0.306 | -0.356
| Anger | 0.673 | 0.678 | 0.659
| Belligerence | 0.677 | 0.605 | 0.55
| Defensiveness | 0.522 | 0.408 | 0.454
| Disgust | 0.69 | 0.634 | 0.421
| Negative | 0.693 | 0.732 | 0.457
| Acceptance | -0.75 | -0.656 | -0.618
| Affection | -0.352 | -0.481 | -0.528
| Agreement | -0.295 | -0.464 | -0.347
| Positive | -0.547 | -0.61 | -0.59
| Satisfaction | -0.537 | -0.465 | -0.464
| Support-emotional | -0.326 | -0.541 | -0.358
| Support-instrumental | -0.343 | -0.571 | -0.39
| Anxiety | 0.171 | 0.369 | 0.349
| Avoidance | 0.085 | - | -0.005
| Change | 0.7 | 0.76 | 0.725
| Dominance | 0.293 | 0.394 | 0.425
| Sadness | 0.198 | 0.204 | 0.177
| Withdrawal | - | -0.118 | -0.035

Table 5. Spearman Correlation between reference and target behaviors used to calculate the BRC metric (Eqn. 10) for the Neural model. **Ratings** and **Scores** refer to ground-truth correlations $Q^*$ and Neural model score correlations $Q'$ respectively in Eqn. 10. All correlations statistically significant ($p < 0.05$) unless marked as -

6.2.4 Analysis of ELMo layer utilization

Lastly, we perform a qualitative analysis of linguistic characteristics for each behavior group, by inspecting the trained parameters of our Neural models. Specifically, we examine the ELMo layer mixing weights that, as described in Sec. 4.2.2, were used to fuse the embeddings from 3 layers, for each word, into a single input embedding. Since these are softmax-normalized weights, they indicate relatively how much each layer was utilized in order to estimate that behavior. Generally, it has been observed in deep neural networks that lower layers tend to learn simpler representations such as edges in images and phrase-level information in text whereas higher layers tend to learn more complicated representations such as objects in images and semantic features in text (Olah et al., 2017; Jawahar et al., 2019). In particular, the higher layers in ELMo have been found to model complex characteristics of language such as polysemy and semantics better than the lower layers (Peters et al., 2018). Therefore, observing how these weights differ for each behavior group can illuminate how their linguistic characteristics are different.

The 3 layers in ELMo, from lower to higher, are the input token layer and 2 bidirectional language model (biLM) layers, which we denote as 0, 1 and 2 respectively. Their mixing weights can be denoted using the 3-tuple \( (x_1, x_2, x_3) \) where \( x_1 \) and \( x_2 \) are the weights of layers 1 and 2 respectively. Figure 12 displays the scatter plot of \( x_1 \) and \( x_2 \), averaged over all model checkpoints, from each test fold and for each behavior in a group. Each point \((x_1, x_2)\) below the diagonal corresponds to a unique 3-tuple; for example, \((0.1, 0.8)\) corresponds to \( \{0.1, 0.8, 0\} \), which can be interpreted as the model using layer 2, mostly, to estimate that behavior; similarly, \((0.33, 0.33)\) corresponds to \( \{0.34, 0.33, 0.33\} \) which means that the model used all 3 layers to the same extent. We use different colors to highlight regions where a particular layer has the largest weight among all 3 of them; for example, the red region consists of all the points corresponding to 3-tuples where layer 0 has the largest weight, while the blue region represents
Figure 11. Appropriate observation window lengths of behaviors as determined by our analysis when quantified using an Neural model. Differences between model performance at window lengths were found to be statistically significant ($p < 0.05$).

Figure 12. Scatter plot of ELMo layer mixing weights from trained Neural models of behaviors of every group. Colored regions denote a particular layer having the largest weight among all 3 layers. For example, the red region represents the space of all weights where layer 0’s is the largest, whereas blue represents the space where layer 2’s weight is the largest.

Those 3-tuples where layer 2 has the largest weight.

Firstly, we see that Problem-Solving behaviors overwhelmingly tend to use information mostly from either just the highest layer or the top 2 layers. This means that the models used to estimate these behaviors rely mostly on the biLM representations which, as we noted earlier, pertain to complex and high-level characteristics of language. This agrees with our findings from the window length analysis where Problem-Solving behaviors performed best when observed using longer windows, since complex expression patterns are difficult to capture using short observations. Next, we see that Negative behaviors tend to either use all 3 layers in near-equal proportions or lean slightly towards layer 2. This indicates that even though they may favor short windows over longer ones, they are nevertheless characterized...
by both low-level and high-level linguistic features. Similarly, we see that Dysphoric behaviors are mostly clustered around the point (0.33,0.33), implying that their expressions involve a balanced mix of all aspects of language. Finally, we see that Positive behaviors occupy a broad spectrum of expression characteristics, from heavily relying on Layer 2, as evinced by the cluster around the point (0.2,0.6), to not utilizing it much, as seen by the cluster near the point (0.4,0.2). This suggests that the expression of positive affect in language is a multifaceted and diverse phenomenon, which probably explains why we found that Positive behaviors favored both short as well as long observation windows in both the N-gram and Neural models.

This concludes our discussion of the behavior window length results and linguistic characteristics with respect to specific models. Next, we identify results that are consistent across both modeling frameworks and comment on how they inform general patterns about the expression of behaviors in language.

6.3 Behavior Group Characteristics

![Figure 13](image)

Figure 13. Comparison of highest BCS from both models at any window length for different behaviors

...
In the case of **Problem-Solving** and **Dysphoric** behaviors, we see that the Neural model performs as well as, if not better than, the N-gram model. This is intuitive since these behaviors are more complex and ambiguous than affect-based ones and are, thus, not as amenable to being understood solely from short phrases. In addition, we saw in our analysis that both behavior groups performed better mostly at longer window lengths than at short ones. This suggests that understanding these behaviors requires the ability to handle long context dependencies in a sophisticated, possibly non-linear manner, which is precisely the advantage that the Neural model offers over the N-gram one. Moreover, the observation that **Sadness** requires longer observation windows than **Positive** is similar to the finding by Krull et al. (Krull and Dil [1998]) that sad faces evoked less spontaneous reactions than happy faces.

Finally, we observe that the best functional for aggregating some behaviors in both of these groups was found to be either the **maximum** or the **minimum**, which represent extreme deviations from the “typical” value. In particular, **Dominance** was best aggregated as the **maximum**, consistently across both models, while **Withdrawal** was best aggregated as the **minimum**. As seen from their descriptions in Sec. ??, this finding seems to agree with the polar nature of these two behaviors: **Dominance** is overt and involves strong participation in the interaction whereas **Withdrawal** is subtle and primarily characterized by a lack of participation. Our results are thematically aligned with previous works such as (Lee et al. [2012]) which studied how humans use different processes for different behaviors when forming an overall impression over the course of an interaction.

We compare these findings to the results on acoustic vocal cues obtained by Li et al. (Li et al. [2019]) and observe that they agree on the modeling accuracy trends but diverge with respect to the observation window lengths. Specifically, in both spoken language and vocal cues, negative behaviors were generally associated with higher modeling accuracies than the other behavior groups. However, in contrast to our findings, Li et al. reported that behaviors such as **Blame** and **Negativity** performed better with longer observation windows for vocal cues while **Sadness** performed best at shorter windows. This suggests that negative behaviors are sufficiently expressed in both lexical and acoustic modalities but over different time-scales, in which case multi-scale approaches might be beneficial when fusing information across modalities. On the other hand, dysphoric behaviors such as **Sadness** do not appear to be strongly detected in either modality regardless of how long they are observed, hence motivating the need for additional information, such as from the video modality.

7 CONCLUSIONS & FUTURE WORK

In this paper, we analyzed how long a system needs to observe conversational language cues, measured in number of words, in order to quantify different behaviors. We proposed an analysis framework and associated evaluation metrics that can be used to analyze behaviors in scenarios where ground-truth annotations are not available to compare against at every possible window length. We applied our analysis to the Couples Therapy dataset which contains a rich and diverse set of behaviors observed in real-life interactions. We also examined the robustness of our analysis to two different behavior scoring methods, a Maximum Likelihood N-gram model and a Deep Learning model. Finally, we compared our findings with those from similar work in speech processing and psychology and addressed pertinent issues related to the nature of human behavior expression in spoken language.

Based on our analysis, we found that behaviors related to negative affect can be well captured even with short lexical observations while those related to positivity require more information. On the other hand, behaviors involving complex deliberations and back-and-forth dialogue tend to require much longer observation windows in order to be accurately understood while those related to dysphoric affect are, in general, difficult to quantify from language alone even when observed using long windows. Finally, when trying to obtain an overall, aggregate quantification of behavior based on information over a period of time, we find that different behaviors seem to be best captured by different functionals that might be indicative of their nature.

The findings from this work are of relevance not only to machine learning applications that attempt to quantify the behavior content of conversational language, but also to research studies in psychological research and clinical practice domains that deal with manual annotations of behavior. For instance, our work has shown that negative and positive behaviors constructs appear to require short to medium-length amounts of information in order to be accurately quantified. Hence, future studies might find it beneficial, both in terms of cost as well as time, to consider which types of behaviors are of primary importance when deciding on the length of interactions to be collected. Studies focused on such behaviors may be
able to elicit and measure such behaviors over relatively brief periods of time whereas studies focused on dysphoria and problem-solving behaviors likely will require considerably longer intervals that generate larger amounts of text required for accurate text analysis.

The next step in this work would be to extrinsically evaluate our findings in different behavior estimation and classification tasks and checking if they translate into gains in performance, compared to just using the same window length for all behaviors. It is also worth investigating how the window length requirements change when employing a multimodal analysis system that used both lexical as well as acoustic cues. As a supplement to this work, we would like to crowdsource human annotations of how accurately humans can assess different behaviors using different amounts of text information, thus conducting a study similar to those involving thin slices. A related effort in that direction would also be to test on datasets with dialog acts and utterance-level annotations of behavior. Finally, we plan on investigating if functionals which mimic the expected manner in which humans perceive some behaviors, such as those involving primacy and recency [Steiner and Rain [1989]], might be a better fit for different behavior constructs and if so, identifying which functionals work best when quantifying different constructs.

REFERENCES

Ambady, N. and Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. Psychological bulletin, 111(2):256.
Baer, J. S., Wells, E. A., Rosengren, D. B., Hartzler, B., Beadnell, B., and Dunn, C. (2009). Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors. Journal of substance abuse treatment, 37(2):191–202.
Baucom, K. J., Sevier, M., Eldridge, K. A., Doss, B. D., and Christensen, A. (2011). Observed communication in couples two years after integrative and traditional behavioral couple therapy: Outcome and link with five-year follow-up. Journal of consulting and clinical psychology, 79(5):565.
Baumeister, R. F., Bratslavsky, E., Finkenauer, C., and Vohs, K. D. (2001). Bad is stronger than good. Review of general psychology, 5(4):323–370.
Baumeister, R. F., Vohs, K. D., Nathan DeWall, C., and Zhang, L. (2007). How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. Personality and social psychology review, 11(2):167–203.
Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. Journal of machine learning research, 3(Feb):1137–1155.
Black, M. P., Katsamanis, A., Baucom, B. R., Lee, C.-C., Lammert, A. C., Christensen, A., Georgiou, P. G., and Narayanan, S. S. (2013). Toward automating a human behavioral coding system for married couples’ interactions using speech acoustic features. Speech communication, 55(1):1–21.
Blackman, M. C. and Funder, D. C. (1998). The effect of information on consensus and accuracy in personality judgment. Journal of Experimental Social Psychology, 34(2):164–181.
Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4):335.
Carney, D. R., Colvin, C. R., and Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. Journal of Research in Personality, 41(5):1054–1072.
Chakravarthula, S.-N., Baucom, B., and Georgiou, P. (2018). Modeling interpersonal influence of verbal behavior in couples therapy dyadic interactions. Proc. Interspeech 2018, pages 2339–2343.
Chakravarthula, S. N., Gupta, R., Baucom, B., and Georgiou, P. (2015a). A language-based generative model framework for behavioral analysis of couples’ therapy. In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, pages 2090–2094. IEEE.
Chakravarthula, S. N., Xiao, B., Imel, Z. E., Atkins, D. C., and Georgiou, P. G. (2015b). Assessing empathy using static and dynamic behavior models based on therapist’s language in addiction counseling. In Sixteenth Annual Conference of the International Speech Communication Association.
Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 103–111.
Christensen, A., Atkins, D., Berns, S., Wheeler, J., Baucom, D., and Simpson, L. (2004). Traditional
versus integrative behavioral couple therapy for significantly and chronically distressed married couples. *Journal of Consulting and Clinical Psychology*, 72(2):176–191.

Cullen, A. and Harte, N. (2017). Thin slicing to predict viewer impressions of ted talks. In *Proceedings of the 14th International Conference on Auditory-Visual Speech Processing*.

Diedenhofen, B. and Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PloS one*, 10(4):e0121945.

Frank, E. and Hall, M. (2001). A simple approach to ordinal classification. In *European Conference on Machine Learning*, pages 145–156. Springer.

Georgiou, P. G., Black, M. P., Lammert, A. C., Baucom, B. R., and Narayanan, S. S. (2011). “that’s aggravating, very aggravating”: Is it possible to classify behaviors in couple interactions using automatically derived lexical features? In *International Conference on Affective Computing and Intelligent Interaction*, pages 87–96. Springer.

Gibson, J., Can, D., Xiao, B., Imel, Z. E., Atkins, D. C., Georgiou, P., and Narayanan, S. (2016). A deep learning approach to modeling empathy in addiction counseling. *Commitment*, 111:21.

Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Gupta, R., Malandrakis, N., Xiao, B., Guha, T., Van Segbroeck, M., Black, M., Potamianos, A., and Narayanan, S. (2014). Multimodal prediction of affective dimensions and depression in human-computer interactions. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 33–40. ACM.

Heavey, C., Gill, D., and Christensen, A. (2002). Couples interaction rating system 2 (cirs2). *University of California, Los Angeles*, 7.

Heyman, R. E. (2004). Rapid marital interaction coding system (rmics). In *Couple observational coding systems*, pages 81–108. Routledge.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., and Weinberger, K. Q. (2017). Snapshot ensembles: Train 1, get m for free. In *International Conference on Learning Representations*.

Jawahar, G., Sagot, B., Seddah, D., Unicomb, S., Iniguez, G., Karsai, M., Léo, Y., Karsai, M., Sarraute, C., Fleury, É., et al. (2019). What does bert learn about the structure of language? In *57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy*.

Jones, J. and Christensen, A. (1998). Couples interaction study: Social support interaction rating system. *University of California, Los Angeles*, 7.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Krizhevsky, A. and Hinton, G. (2010). Convolutional deep belief networks on cifar-10. *Unpublished manuscript*, 40(7):1–9.

Krull, D. S. and Dil, J. C. (1998). Do smiles elicit more inferences than do frowns? the effect of emotional valence on the production of spontaneous inferences. *Personality and Social Psychology Bulletin*, 24(3):289–300.

Krzyzaniak, S. L., Colman, D. E., Letzring, T. D., McDonald, J. S., and Biesanz, J. C. (2019). The effect of information quantity on distinctive accuracy and normativity of personality trait judgments. *European Journal of Personality*.

Lee, C.-C., Black, M., Katsamanis, A., Lammert, A. C., Baucom, B. R., Christensen, A., Georgiou, P. G., and Narayanan, S. S. (2010). Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In *Eleventh Annual Conference of the International Speech Communication Association*.

Lee, C.-C., Katsamanis, A., Black, M. P., Baucom, B. R., Christensen, A., Georgiou, P. G., and Narayanan, S. S. (2014). Computing vocal entrainment: A signal-derived pca-based quantification scheme with application to affect analysis in married couple interactions. *Computer Speech & Language*, 28(2):518–539.

Lee, C.-C., Katsamanis, A., Georgiou, P. G., and Narayanan, S. S. (2012). Based on isolated saliency or causal integration? toward a better understanding of human annotation process using multiple instance learning and sequential probability ratio test. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Li, H., Baucom, B., and Georgiou, P. (2019). Linking emotions to behaviors through deep transfer
learning. submitted to PeerJ Computer Science.
McCrae, R. R., Costa Jr, P. T., and Busch, C. M. (1986). Evaluating comprehensiveness in personality systems: The california q-set and the five-factor model. Journal of Personality, 54(2):430–446.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119.
Morales, M. R., Scherer, S., and Levitan, R. (2018). A linguistically-informed fusion approach for multimodal depression detection. In NAACL HLT, page 13.
Moyers, T. B., Martin, T., Manuel, J. K., Miller, W. R., and Ernst, D. (2003). The motivational interviewing treatment integrity (miti) code: Version 2.0. Retrieved from Verf¨ubar unter: www. casaa. unm. edu [01.03. 2005].
Murphy, N. A., Hall, J. A., Ruben, M. A., Frauendorfer, D., Schmid Mast, M., Johnson, K. E., and Nguyen, L. (2018). Predictive validity of thin-slice nonverbal behavior from social interactions. Personality and Social Psychology Bulletin, page 0146167218802834.
Narayanan, S. and Georgiou, P. G. (2013). Behavioral signal processing: Deriving human behavioral informatics from speech and language. Proceedings of the IEEE, 101(5):1203–1233.
Nolan, J. (2003). Stable distributions: models for heavy-tailed data. Öhman, A., Lundqvist, D., and Esteves, F. (2001). The face in the crowd revisited: a threat advantage with schematic stimuli. Journal of personality and social psychology, 80(3):381.
Olah, C., Mordvintsev, A., and Schubert, L. (2017). Feature visualization. Distill, 2(11):e7.
Pérez-Rosas, V., Mihalcea, R., Resnicow, K., Singh, S., and An, L. (2017). Understanding and predicting empathic behavior in counseling therapy. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 1426–1435.
Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237.
Reblin, M., Baucom, B. R., Clayton, M. F., Utz, R., Caserta, M., Lund, D., Mooney, K., and Ellington, L. (2019). Communication of emotion in home hospice cancer care: Implications for spouse caregiver depression into bereavement. Psycho-Oncology.
Rozgić, V., Xiao, B., Katsamanis, A., Baucom, B., Georgiou, P. G., and Narayanan, S. (2011). Estimation of ordinal approach-avoidance labels in dyadic interactions: Ordinal logistic regression approach. In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2368–2371. IEEE.
Satterstrom, P, Polzer, J. T., Kwan, L. B., Hauser, O. P., Wiruchnipawan, W., and Burke, M. (2019). Thin slices of workgroups. Organizational Behavior and Human Decision Processes, 151:104–117.
Schuller, B., Valster, M., Eyben, F., Cowie, R., and Pantic, M. (2012). Aavec 2012: the continuous audio/visual emotion challenge. In Proceedings of the 14th ACM international conference on Multimodal interaction, pages 449–456. ACM.
Sevier, M., Eldridge, K., Jones, J., Doss, B. D., and Christensen, A. (2008). Observed communication and associations with satisfaction during traditional and integrative behavioral couple therapy. Behavior therapy, 39(2):137–150.
Smith, L. N. (2017). Cyclical learning rates for training neural networks. In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 464–472. IEEE.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1):1929–1958.
Steiner, D. D. and Rain, J. S. (1989). Immediate and delayed primacy and recency effects in performance evaluation. Journal of Applied Psychology, 74(1):136.
Stolcke, A. (2002). Srilm-an extensible language modeling toolkit. In Seventh international conference on spoken language processing.
Thornton, M. A. and Tamir, D. I. (2017). Mental models accurately predict emotion transitions. Proceedings of the National Academy of Sciences, 114(23):5982–5987.
Tseng, S.-Y ., Baucom, B., and Georgiou, P. (2017). Approaching human performance in behavior estimation in couples therapy using deep sentence embeddings. In Proceedings of Interspeech. August
Tseng, S.-Y., Chakravarthula, S. N., Baucom, B. R., and Georgiou, P. G. (2016). Couples behavior modeling and annotation using low-resource lstm language models. In INTERSPEECH, pages 898–902.

Tseng, S.-Y., Li, H., Baucom, B., and Georgiou, P. (2018). Honey, i learned to talk: Multimodal fusion for behavior analysis. In Proceedings of the 2018 on International Conference on Multimodal Interaction, pages 239–243. ACM.

Xia, W., Gibson, J., Xiao, B., Baucom, B., and Georgiou, P. G. (2015). A dynamic model for behavioral analysis of couple interactions using acoustic features. In Sixteenth Annual Conference of the International Speech Communication Association.

Xiao, B., Can, D., Georgiou, P. G., Atkins, D., and Narayanan, S. S. (2012). Analyzing the language of therapist empathy in motivational interview based psychotherapy. In Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific, pages 1–4. IEEE.

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. (2018). Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2236–2246.

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. Psychological methods, 12(4):399.