Quality Assessment of Cognitive Behavioral Therapy Sessions Through Highly Contextualized Language Representations

Nikolaos Flemotomos
Department of Electrical and Computer Engineering
University of Southern California, Los Angeles, CA, USA
flemotom@usc.edu

Abstract

During a psychotherapy session, the counselor typically applies techniques which are codified along specific dimensions (e.g., 'displays warmth and confidence', or 'attempts to set up collaboration'). Those constructs, traditionally scored by trained human raters, reflect the complex nature of psychotherapy and highly depend on the context and the entire history of the discourse. Recent advances in large contextualized language models offer an avenue for accurate in-domain linguistic representations which can lead to robust recognition and scoring of such behavioral constructs, thus leading to better quality of services and supervision. In this work, a BERT-based model is proposed for automatic behavioral scoring of a specific type of psychotherapy, called Cognitive Behavioral Therapy (CBT), where prior work is limited to frequency-based language features and/or short text excerpts which do not capture the unique elements involved in a spontaneous long conversational interaction. In order to provide relevant non-linguistic context, BERT-based representations are further augmented with available therapy metadata, leading to consistent performance improvements.

1 Introduction

Psychotherapy is an intervention based on the verbal communication between the affected individual and a trained professional, aimed at treating mental health disorders. The effectiveness of psychotherapy is widely studied and accepted (Lambert and Bergin, 2002; Perry et al., 1999) and leads millions of people seeking professional help at a yearly basis (Substance Abuse and Mental Health Services Administration, 2019). Cognitive Behavioral Therapy (CBT) (Beck, 2011) is a particular type of psychotherapy that aims at shifting the patient's patterns of thinking by changing maladaptive cognitions and beliefs connected to behavioral problems. It is one of the most popular psychotherapeutic approaches (Gaudiano, 2008) with strong evidence connecting its methods with positive clinical outcomes (Hofmann et al., 2012).

Given its wide popularity and its application to a variety of mental health problems, performance-based measures that ensure high quality of CBT provision are deemed essential (Creed et al., 2016). The gold-standard method for monitoring therapy quality is *behavioral coding* (Bakeman and Quera, 2012), a process during which trained coders listen to audio recordings in order to assess specific therapeutic skills. For CBT, in particular, the most widely used coding scheme is the Cognitive Therapy Rating Scale (CTRS; Vallis et al. (1986)), that defines a set of 11 session-level codes reflecting skills and techniques specific to the intervention. This traditional approach poses strict time- and cost-related limitations to a widespread use into real-world clinical settings, which means that the vast majority of CBT sessions are simply not evaluated.

Recent technological advances have given rise to a digital healthcare era with numerous applications focusing on mental health (Bone et al., 2017). Automatic behavioral coding is a field which has drawn a lot of research interest over the last few years (e.g., Tanana et al. (2016); Gibson et al. (2016); Singla et al. (2018)) and holds promise for more efficient training, more effective supervision, and more positive clinical outcomes. However, despite being one of the most dominant psychotherapy interventions, the literature focusing on computational analysis for CBT sessions is relatively scarce, partly because of limited available data. The existing proposed systems mainly depend on frequency-based and hand-crafted features (Flemotomos et al., 2018; Chen et al., 2020), or study CBT-related constructs appearing in short text excerpts which are not part of an actual therapy session (Barahona et al., 2018). CBT sessions, however, are usually several minutes long (or even longer than an hour), with a typical session consisting of several tens or hundreds of talk turns and utterances. At the same time, the behavioral constructs under examination reflect complex structural, conceptual, and communicative aspects of the therapy that the existing approaches potentially fail to capture.

Inspired by the recent success stories of large pre-trained language models in numerous Natural Language Processing (NLP) tasks (Devlin et al., 2019; Yang et al., 2019; Brown et al., 2020), in this work I am using such models for the downstream task of CBT quality assessment based on the total CTRS score. The total CTRS — equal to the sum of the 11 individual codes — is an aggregate metric used in clinical practice to evaluate a practitioner's degree of competence in delivering CBT. In more detail, a BERT model (Devlin et al., 2019) is adapted to the domain and used to extract semantic representations. Those are passed through a recurrent architecture, trained either to directly classify a session with respect to the total CTRS or to model all the constituent codes in a mult-task approach. Side information from available metadata is also included to the final models, leading to improved predictive power of the system, compared to only using linguistic cues.

The proposed system is evaluated on a set of more than 1,000 real-world CBT sessions, recorded and automatically transcribed, which are accompanied by human annotations. To the best of my knowledge, this is the first attempt to use linguistic information extracted from BERT-like architectures, not only in CBT, but for behavioral code prediction in general. Experimental results show consistent improvements over baseline approaches.

2 Datasets

The Beck Community Initiative (BCI) partners provide high-quality psychotherapy training to community clinics and, through this work, have generated a large archive of recorded CBT sessions (Creed et al., 2016), many of which are accompanied by CTRS scores. Out of those, 292 sessions have been sent for professional transcription. The selection of the particular sessions was done so that the audio quality is above a certain threshold and there is a fair representation of sessions across the entire range of the total CTRS scale. I used those human-transcribed sessions to adapt and evaluate an automatic speech transcription pipeline, which was later used to transcribe a total of 1,118 CBT sessions (including the 292 already mentioned). This number comes from the initial pool of available sessions after excluding those marked as non-English and the ones for which not all the 11 CTRS codes were available.

The transcription pipeline is based on (Flemotomos et al., 2020), after getting adapted to the CBT data. Adaptation was based on 100 transcribed sessions which were not further used (leaving 1,018 CBT sessions for further experimentation). The final speech recognition error, when the pipeline is evaluated on the remaining 192 human-transcribed sessions, is 45.81%, with the therapist-attributed error being 41.19%. Even though the error is high, those numbers are inflated since they are highly affected by fillers and other idiosyncrasies of conversational speech.

Each one of the 11 CTRS codes listed in Table 1 is scored by a trained human coder on a 7-point Likert scale (0-6). In clinical practice, any CTRS score above or equal to 4 indicates competency on that behavioral construct and any score lower than 4 indicates room for improvement. Additionally, giving equal importance to all the 11 CTRS dimensions, CBT researchers take into consideration the total CTRS which is the sum of the 11 components. A total CTRS above or equal to 40 indicates competent delivery of CBT, whereas a score less than 40 could suggest, for example, that additional training is required for the particular practitioner. The focus of this work is on the binary classification problem of the total CTRS (below/above 40). The distribution of all the CTRS codes is given in Figure 1. Even though the total CTRS follows an approximately normal distribution, after binarization the problem is unbalanced with the dataset becoming skewed towards the class with non-competent CBT delivery (total CTRS below 40), which contains 76.23% of the sessions.

For all the sessions, a limited amount of metadata is also available. In particular, the variables taken into consideration in this work are:

1. the clinic where the session took place: The dataset totally consists of sessions delivered by 383 therapists across 25 different clinics.

2. the specific clinical program type: The sessions are clustered into 16 distinct programs, including school-based, family, geriatric, etc.

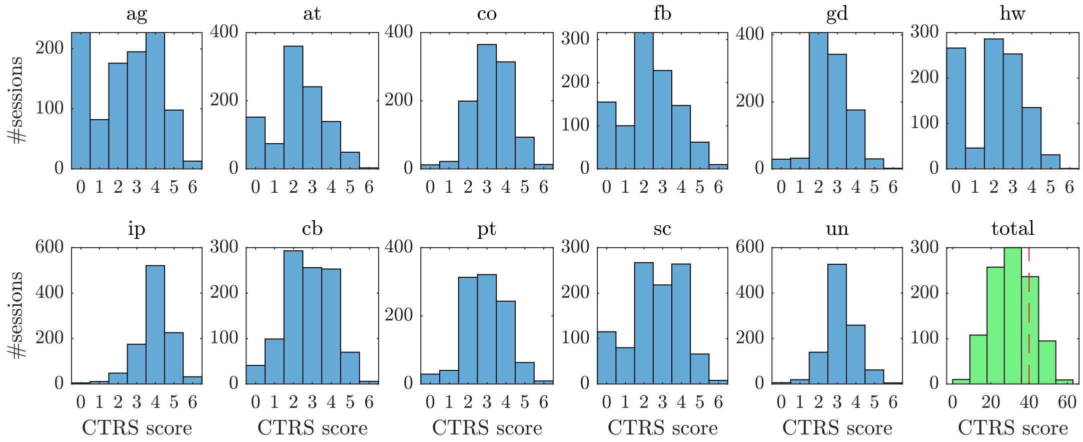

Figure 1: Distribution of the 11 CTRS codes (and the total CTRS) across the Likert scale.

abbr.	meaning
ag	agenda
fb	feedback
un	understanding
ip	interpersonal effectiveness
co	collaboration
pt	pacing and efficient use of time
gd	guided discovery
cb	focusing on key cognitions & behaviors
sc	strategy for change
at	application of techniques
hw	homework

Table 1: The 11 CBT quality codes defined by CTRS.

As shown from those examples, the program type often carries information about characteristics of the population from where the patients are sampled.

3. the assessment time with respect to when the CBT-focused training of the corresponding therapist took place: Each therapist participating in the program attends a workshop organized by BCI to receive CBT training. It is expected that counselors adhere more to CBT-related skills after their training and the degree of their competency gets higher as they get more experience. Thus, the availability of such information can be useful for the task of CTRS prediction. There are totally 7 timestamps characterizing a session along this dimension (e.g., pre-workshop, post-workshop, one/three/six months after workshop).

Even though this is the largest corpus of CBT sessions ever used for computational analysis, this is probably still not enough for a sufficient adaptation and training of the models. To that end, I additionally employ a set of 4,269 recorded sessions automatically transcribed from a university counseling center (Flemotomos et al., 2020). We will denote this as the UCC set. Those sessions span a wide range of psychotherapy approaches (including, but not limited to, CBT) and have not been coded following the CTRS. Despite the expected differences between the two domains (e.g., the UCC sessions are focused on concerns common among college students, such as anxiety, exams, etc.), several common linguistic patterns in psychotherapy are expected to be shared, so this set is deemed suitable to adapt the BERT model which will be used to extract utterance-level linguistic representations of the CBT sessions. An additional advantage of using the UCC sessions for adaptation is that they have been transcribed using the same transcription pipeline as the CBT sessions, so BERT is expected to be fine-tuned not only on the psychotherapy domain, but also on transcription-specific errors. The size of the two datasets in terms of duration and number of utterances/words is provided in Table 2.

3 Method

3.1 Single-task approach

As mentioned, the current work is focused on the binary classification problem of low vs. high total CTRS score. Thus, it is natural, as a first step, to build a model viewing the problem as a single task where the output is exactly a binary variable denoting whether the therapist is considered to have successfully adhered to CBT-related skills or not.

dataset	number of sessions	session duration in min (mean ± std)	talk turns per session (mean ± std)	words per talk turn (mean ± std)
CBT	1,018	41.5 ± 14.2	431.1 ± 229.3	12.7 ± 24.3
UCC	4,269	49.8 ± 11.5	438.3 ± 200.7	15.4 ± 29.6

Table 2: Size of the datasets to be used to train and evaluate the proposed models.

Since the majority of the codes defined by CTRS only depend on therapist behavior and are not directly related to the patient (e.g. 'did the therapist set a clear agenda for the session?'), we can focus on the utterances (talk turns) assigned to the former. Such an architecture is illustrated in Figure 2.

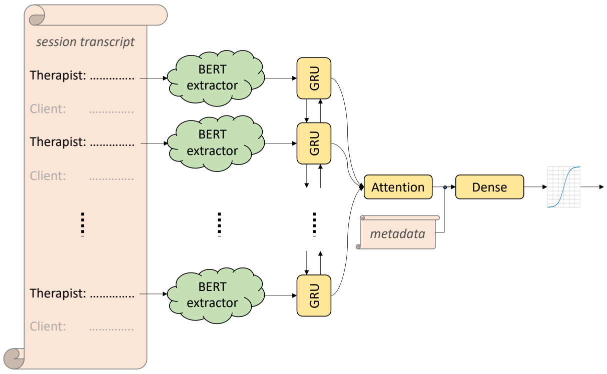

Figure 2: Proposed architecture for total CTRS score classification following a single-task approach when only the utterances attributed to the therapist are used.

First, a pre-trained BERT model is adapted to the psychotherapy domain by continuing training on an external, in-domain dataset (in our case, the UCC data). This can then be used to extract fixed-dimensional linguistic representations for each available utterance by average-pooling the last layer.

The sequence of utterance representations forms the input to a bidirectional recurrent layer that accordingly outputs a sequence of hidden vectors, each one of which takes into consideration not only the corresponding utterance but the entire context of the session (i.e., what is said before and after the utterance). In particular, the hidden vector corresponding to each recurrent cell is considered to be the concatenation of the forward and backward outputs of the specific cell. This sequence is fed to an additive self-attention layer (Bahdanau et al., 2015) which models the entire session as a weighted average of the information encoded in the hidden vectors, thus learning which parts of the session are useful in order to construct a meaningful (with respect to the final task of overall CTRS prediction) session representation. This representation can now be concatenated with the available session-level metadata information, represented by one-hot variables. Finally, a sigmoid non-linearity is applied after a dense layer, which gives the desired output. The network is optimized based on the binary cross-entropy loss function.

3.2 Multi-task approach

The architecture described in Section 3.1 does not take into account what exactly the total CTRS represents, which is estimated as the unweighted sum of the 11 individual CTRS codes. However, different codes typically represent completely different CBT skills which are related to specific linguistic patterns and are often applied by the therapist during different parts of the session. For example, the therapist is expected to set an appropriate *agenda* towards the beginning of the session that includes specific target problems the patient is concerned about. Similarly, an important aspect of a successful CBT session is incorporating *homework* relative to the therapy. That includes reviewing previous homework (typically done towards the beginning of the session) and assigning new homework for the coming week (typically done towards the end of the session). Finally, there are codes which reflect communicative skills expected to be displayed throughout the entire session. For instance, the therapist is expected to thoroughly understand the patient and properly communicate this *understanding* through appropriate verbal responses.

In order to implicitly incorporate such knowledge in the network, I propose following instead a multi-task approach, as depicted in Figure 3. The first steps (BERT-based feature extraction and bidirectional recurrency) are the same as in the single-task approach described in Section 3.1. However, instead of directly modeling the total CTRS score, the network of Figure 3 tries to separately model each one of the 11 codes, with each code defining a "task" for the network.

In more detail, the sequence of the hidden vec-

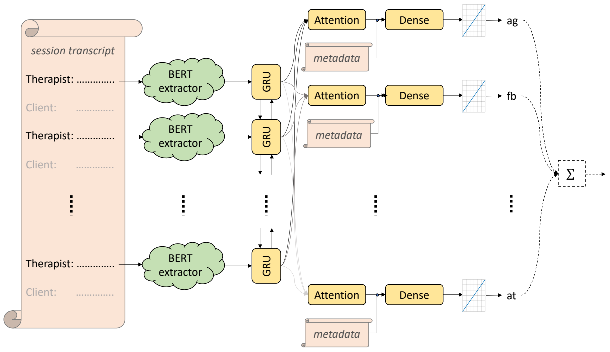

Figure 3: Proposed architecture for total CTRS score classification following a multi-task approach modeling each CTRS code when only the utterances attributed to the therapist are used.

tors from the recurrent layer is shared across all the tasks and is the input to 11 different attention layers, each one associated with a specific CTRS code. That way, the network can attend to what is important for the prediction of a particular code. As previously, metadata information in the form of one-hot encoded variables is concatenated to the context vector learned by the attention layers and is passed through a final dense layer with linear (instead of sigmoid) activation function. So, a continuous output — restricted in the range $[0, 6]$ — represents each CTRS code. Those codes can later be added and the binarized sum corresponds to the desired classification outcome (low vs. hight total CTRS). The loss function to be optimized during training is

$$L = \sum_{i=1}^{11} L_i$$

where L_i is the mean squared error associated with the i-th code.

An additional advantage we get following this approach is enhanced interpretability, an aspect of high importance, especially in systems to be used in real-world clinical settings. Instead of viewing the overall system as a black box giving information about the total CTRS score, we can now track specific attributes which led to the classification of a session as competent or not. For example, if such a system is used to provide feedback to counselors, this can be targeted to specific areas in need for improvement.

4 Experiments and Results

4.1 BERT fine-tuning

For this work I employ the pre-trained baseBERT model[1] as a feature extractor. This is fine-tuned by continuing training on the set of 4,269 UCC sessions (Table 2) after a random $90\% - 10\%$ train-eval split at the session level. Two adapted BERT models are built and evaluated: i) a model fine-tuned on all the available utterances, called *psychBERT*, and ii) a model fine-tuned only on the therapist-attributed utterances, called *therapistBERT*. In both cases, a maximum utterance length of 64 tokens was assumed and fine-tuning took place for 10,000 steps with a learning rate equal to $2 \cdot 10^{-5}$ and minibatch size equal to 64.

When evaluated on the CBT sessions, the accuracy on the next sentence prediction task is given in Table 3. As shown, adaptation leads to substantial improvements, both in the cases of psychBERT and therapistBERT. The large performance gap when baseBERT is used comes at no surprise: The higher accuracy for the task when all the utterances are taken into consideration (compared to the case when only the therapist utterances are evaluated) is due to the fact that the base model can more accurately represent naturalistic conversations (e.g., questions-answers), compared to predicting the next utterance of a specific person, skipping one of the interlocutors. However, after fine-tuning, the system does an almost equally good job for the two cases (and even slightly better when only applied to the therapist utterances).

model	CBT set all utterances	CBT set therapist utterances
baseBERT	60.03	40.00
psychBERT	69.53	–
therapistBERT	–	71.66

Table 3: Next sentence prediction accuracy (%) before and after BERT adaptation when evaluated on the CBT dataset.

4.2 Experimental Setup

The models are built and trained using Tensorflow (Abadi et al., 2016). In any case, an Adam optimizer with initial learning equal to 0.001 is employed. The models are trained for a maximum of

[1] https://github.com/google-research/bert

200 epochs with early stopping based on validation loss (and with patience set equal to 10 epochs). When focusing only on therapist-attributed utterances, the maximum sequence length (session length) is set to 256 utterances and a minibatch size equal to 128 is used. When all the utterances are taken into consideration, the minibatch size is 64 and the maximum sequence length is set to 512 utterances.

All the results reported are based on a 10-fold cross validation scheme so that there is no therapist overlap between the folds (the patient IDs are not known). Since there is a considerable class imbalance, I chose as evaluation metric the macro-averaged F_1 score.

4.3 Results and Discussion

The experimental results, following various combinations of the proposed models and techniques, are given in Tables 4 and 5. Comparing the two Tables, it is apparent that the inclusion of the patient utterances does not usually provide additional useful information for the task of the total CTRS prediction. Even though psychotherapy is a dyadic interaction and one would assume that the entire history of the dialog could improve the predictive power of the system, the results support the initial hypothesis that focusing on therapist-only language is sufficient to assess therapist-related behaviors within the proposed framework.

utterance representation	metadata info	single-task	multi-task
baseBERT	✗	60.48	59.68
	✓	65.96	70.83
psychBERT	✗	65.85	63.03
	✓	69.56	**71.99**

Table 4: F_1 score (%) based on 10-fold CV when all the utterances are used.

Additionally, it is shown that, as expected, in most cases the fine-tuned BERT extractor yields better linguistic representations than the base model, at least with respect to our final goal. Overall, the best results are achieved when we use the multi-task approach with the fine-tuned BERT model and after providing the available metadata information. However, it is interesting to note that, while metadata information is consistently beneficial to the system, it is not always clear whether the multi-task approach leads to improved results, specifically when metadata information is not provided. Given the availability of such side information, though, the multi-task architecture does indeed boost the overall performance. This is likely due to the fact that, in this case, metadata improves the robustness while estimating each one of the codes, thus improving the overall robustness.

utterance representation	metadata info	single-task	multi-task
baseBERT	✗	61.71	64.09
	✓	68.46	71.64
therapistBERT	✗	65.68	61.70
	✓	69.55	**72.40**

Table 5: F_1 score (%) based on 10-fold CV when only the utterances assigned to the therapist are used.

It should be highlighted, here, that, while non-linguistic side information proves to be highly beneficial for the particular dataset, further investigation is required to study how each specific metadata variable affects the overall result and, importantly, which variables are expected to be readily available in the general case of CBT quality assessment “in the wild”. For instance, even though the assessment time with respect to CBT training appears to be a reasonable proxy of CBT quality, should we expect that such information be always provided to the system? In a real-world scenario, such decisions could actually be informative of how an interface used in clinical settings should be built, i.e., what therapist-related information should be asked for during a new user registration.

Finally, it is important to note that, most of the time, the value of the used metadata variables was known to the human annotators. So, it is not clear at this point whether the performance boost is due to actual useful complementary information that such non-linguistic variables carry or due to modeling annotator bias. For example, annotators may be biased towards specific clinics because of exceptional reputation, or they may be stricter when evaluating therapists who are not adequately experienced in CBT techniques.

5 Ethical and Practical Implications

When dealing with such sensitive topics, like psychotherapy and automatic evaluation of one’s performance, it is important to step back and reflect on the implications of our work. Speech and language processing models keep getting better with an unprecedented pace, and the same is expected for the downstream tasks that those models are used for. But which are the key areas we should focus on when using those models and when developing techniques that exploit people's data and affect user's lives? At least three questions need to be answered with respect to the specific application we are dealing with in this work:

1. *Is it acceptable to collect and use patients' sensitive data for training?* I believe that, since being able to offer better quality of psychotherapy services would have a tremendous positive impact to the society and to the patients themselves, the answer here would be yes, but only if necessary conditions are met. There is no doubt that psychotherapy sessions contain extremely sensitive information, since patients often build a trust bond with their therapist, unbosom themselves, and disclose thoughts and secrets they are afraid or ashamed to share even with friends and family. Thus, it is of utmost importance that such data be treated with extreme caution. Of course, the current study is governed by restrictions imposed by the relevant Institutional Review Board, while all participating therapists and patients are asked to sign a detailed consent form. However, this only provides a formal framework and constitutes just a first step. It should be a moral obligation of each individual researcher working with such data to treat them cautiously, and not only because of external restrictions imposed to them. Additionally, both therapists and patients should have the right to withdraw during or after the study. Finally, all data should be de-identified, as much as possible, with respect to patients.

2. *What if such a system is used to blindly evaluate a therapist?* Since CTRS provides quantifiable metrics for quality assessment, the proposed system can be used to evaluate a therapist's performance and their adherence to specific psychotherapeutic skills. In a dystopian scenario, that could mean therapists loosing their jobs because of not meeting minimum standards and students being disappointed because of getting "low scores" from some automated system. It should be made clear that our goal is not to replace human supervision, but rather augment the supervisor's efficiency and additionally offer a tool for self-assessment. Moreover, it is important that the users be adequately trained to understand the meaning of automatically generated evaluation scores. This is why the focus should be on highly interpretable models.

3. *What if the system is wrong? Are there any explicit additional requirements before using such a system in clinical settings?* I believe it is important for a practical realistic system to incorporate confidence metrics and quality safeguards. The described system depends on a series of machine learning models where things can simply go wrong. Establishing confidence metrics for the quality of the automatic transcription (e.g., speech recognition - induced errors) and the final CTRS prediction (e.g., applying thresholds on the final sigmoid non-linearity) would enhance the transparency of the models and would help practitioners trust them and introduce them into the clinical world.

6 Prior Work

6.1 Automatic behavioral coding

There has been an increasing interest in developing systems for automatic psychotherapy evaluation over the last few years, focusing on both acoustic (e.g., Black et al. (2013); Nasir et al. (2018)) and textual information. Depending on the domain, coding procedures may be applied at different resolutions, i.e., at the utterance (e.g., Atkins et al. (2014); Pérez-Rosas et al. (2017)) or at the session level. I am limiting this short overview in the proposed language-based approaches for session-level codes, since this is the focus of the current project.

Early works in the field employed n-gram models (Georgiou et al., 2011; Xiao et al., 2014) and domain-specific semantic features (Gibson et al., 2015) coulped with maximum likelihood (Xiao et al., 2014), maximum entropy (Xiao et al., 2015), and SVM (Gibson et al., 2015) classifiers. Linguistic similarities between the therapist and the patient have been also studied as a useful predictor for therapist behaviors (Lord et al., 2015; Nasir et al., 2019). Deep learning techniques opened up the way for more accurate language modeling and better performance for the behavioral coding task (Gibson et al., 2016; Tseng et al., 2016).

In the CBT domain, Flemotomos et al. (2018) compared various linguistic features on a limited dataset of therapy transcriptions, both manually and automatically derived, and demonstrated that simple language representations, like tf-idf features, can achieve competitive results. However, the dataset was selected to showcase a use case focusing on the two extremes of the rating scale with mostly very low and very high CTRS values. The same dataset was utilized by Gibson et al. (2019), who used additional sessions from a different psychotherapy domain (with a different coding scheme) in a multi-task setting. Chen et al. (2020) improved the tf-idf based approach by enhancing the features with information again distilled from a different domain. A large corpus of written posts from an online platform was used by Barahona et al. (2018), who examined several deep learning approaches for CBT-related mental health concept understanding. However, online posts are typically much shorter than an actual CBT session and exhibit a more well-defined structure than a spontaneous conversational interaction.

6.2 Highly contextualized language models

Large pre-trained language models have lately led to several developments and breakthroughs in numerous NLP tasks, including text classification, text generation, question-answering, and natural language inference. Those language models are usually built based on the concept of Transformer (Vaswani et al., 2017). Using several stacked Transformer blocks, systems like GPT (Radford et al., 2018) and BERT (Devlin et al., 2019) were able to push the limits of NLP.

BERT opened up a new era in NLP with several variants having been proposed, which are usually targeted at specific tasks and applications, or address certain BERT limitations. In its original form, for instance, BERT is only able to handle relatively short pieces of text. DocBERT (Adhikari et al., 2019) was proposed to address this limitation by focusing on the task of document classification. Psychotherapy code prediction is actually a variant of document classification, with the "document", however, being a dialogue. ToD-BERT (Wu et al., 2020) has been specifically proposed to incorporate the power of BERT in modeling task-oriented dialogues. Similarly, DialoGPT (Zhang et al., 2019) builds upon GPT-2 (Radford et al., 2019) focusing on dialogues, but for the task of response generation.

Domain-specific BERT variants have been also developed for particular fields which use, for example, specialized vocabulary (e.g., Lee et al. (2020); Lee and Hsiang (2020)). In the clinical domain, Alsentzer et al. (2019) adapted the BERT embeddings both on general clinical corpora and on discharge summaries in particular. Similarly, Huang et al. (2019) adapted the BERT model on clinical notes for the task of hospital readmission prediction. However, those adaptation processes were based on written text and do not focus on medical conversations, such as psychotherapy.

7 Conclusion

In this work I introduced a model for quality assessment of psychotherapy sessions based on fine-tuned BERT representations. The focus throughout the analysis was on the binary classification of CBT sessions with respect to the overall CTRS score. Two main architectures were proposed and compared. One was based on a single-task approach directly modeling the total CTRS as the binary output. The other exploited the definition of the total CTRS as the sum of 11 constituent scores, and was instead based on a multi-task approach where each score defined a task. Additionally, non-linguistic information was given to the models in the form of metadata variables modeling critical session and therapist characteristics. Experimental results showed that the best performance is achieved employing the multi-task network with metadata information. Finally, a set of ethical considerations and practical recommendations for the research community was proposed. Moving forward, I am confident that similar systems will be proved invaluable in clinical practice, leading to more efficient training and supervision, improved quality of services, and, eventually, more positive clinical outcomes.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283.

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398*.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

David C Atkins, Mark Steyvers, Zac E Imel, and Padhraic Smyth. 2014. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, 9(1):49.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Roger Bakeman and Vicenç Quera. 2012. Behavioral observation. In H Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, and K. J. Sher, editors, *APA handbook of research methods in psychology, Vol. 1. Foundations, planning, measures, and psychometrics*, pages 207–225. American Psychological Association, Washington, DC.

Lina M Rojas Barahona, Bo-Hsiang Tseng, Yinpei Dai, Clare Mansfield, Osman Ramadan, Stefan Ultes, Michael Crawford, and Milica Gasic. 2018. Deep learning for language understanding of mental health concepts derived from cognitive behavioural therapy. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 44–54.

J.S. Beck. 2011. *Cognitive behavior therapy: Basics and beyond*. Guilford Press, New York, NY, USA.

Matthew P Black, Athanasios Katsamanis, Brian R Baucom, Chi-Chun Lee, Adam C Lammert, Andrew Christensen, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2013. Toward automating a human behavioral coding system for married couples interactions using speech acoustic features. *Speech communication*, 55(1):1–21.

Daniel Bone, Chi-Chun Lee, Theodora Chaspari, James Gibson, and Shrikanth Narayanan. 2017. Signal processing and machine learning for mental health research and clinical applications [perspectives]. *IEEE Signal Processing Magazine*, 34(5):189–196.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Zhuohao Chen, Nikolaos Flemotomos, Victor Ardulov, Torrey A Creed, Zac E Imel, David C Atkins, and Shrikanth Narayanan. 2020. Feature fusion strategies for end-to-end evaluation of cognitive behavior therapy sessions. *arXiv preprint arXiv:2005.07809*.

T.A. Creed, S.A. Frankel, R.E. German, K.L. Green, S. Jager-Hyman, K.P. Taylor, A.D. Adler, C.B. Wolk, S.W. Stirman, S.H. Waltman, et al. 2016. Implementation of transdiagnostic cognitive therapy in community behavioral health: The beck community initiative. *Journal of consulting and clinical psychology*, 84(12):1116–1126.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Nikolaos Flemotomos, Victor R Martinez, Zhuohao Chen, Karan Singla, Victor Ardulov, Raghuveer Peri, James Gibson, Michael P Tanana, Panayiotis Georgiou, Jake Van Epps, Tad Lord, Sarah P Hirsch, Zac E Imel, David C Atkins, and Shrikanth Narayanan. 2020. “Am i a good therapist?” automated evaluation of psychotherapy skills using speech and language technologies. Manuscript under review.

Nikolaos Flemotomos, Victor R Martinez, James Gibson, David C Atkins, Torrey Creed, and Shrikanth S Narayanan. 2018. Language features for automated evaluation of cognitive behavior psychotherapy sessions. In *INTERSPEECH*, pages 1908–1912.

Brandon A Gaudiano. 2008. Cognitive-behavioural therapies: achievements and challenges. *Evidence-based mental health*, 11(1):5–7.

Panayiotis G Georgiou, Matthew P Black, Adam C Lammert, Brian R Baucom, and Shrikanth S Narayanan. 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.

James Gibson, David Atkins, Torrey Creed, Zac Imel, Panayiotis Georgiou, and Shrikanth Narayanan. 2019. Multi-label multi-task deep learning for behavioral coding. *IEEE Transactions on Affective Computing*.

James Gibson, Doğan Can, Bo Xiao, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth S Narayanan. 2016. A deep learning approach to modeling empathy in addiction counseling. *Proc. Annual Conference of the International Speech Communication Association*, pages 1447–1451.

James Gibson, Nikolaos Malandrakis, Francisco Romero, David C Atkins, and Shrikanth S Narayanan. 2015. Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. In *Proc. Annual Conference of the International Speech Communication Association*.

Stefan G Hofmann, Anu Asnaani, Imke JJ Vonk, Alice T Sawyer, and Angela Fang. 2012. The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive therapy and research*, 36(5):427–440.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Michael J Lambert and Allen E Bergin. 2002. The effectiveness of psychotherapy. In Michel Hersen and William Sledge, editors, *Encyclopedia of Psychotherapy*, volume 1, pages 709–714. Elsevier Science, USA.

Jieh-Sheng Lee and Jieh Hsiang. 2020. Patent classification by fine-tuning bert language model. *World Patent Information*, 61:101965.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Sarah Peregrine Lord, Elisa Sheng, Zac E Imel, John Baer, and David C Atkins. 2015. More than reflections: empathy in motivational interviewing includes language style synchrony between therapist and client. *Behavior therapy*, 46(3):296–303.

Md Nasir, Brian Baucom, Shrikanth Narayanan, and Panayiotis Georgiou. 2018. Towards an unsupervised entrainment distance in conversational speech using deep neural networks. *Proc. Annual Conference of the International Speech Communication Association*, pages 3423–3427.

Md Nasir, Sandeep Nallan Chakravarthula, Brian RW Baucom, David C Atkins, Panayiotis Georgiou, and Shrikanth Narayanan. 2019. Modeling interpersonal linguistic coordination in conversations using word movers distance. *Proc. Annual Conference of the International Speech Communication Association*, pages 1423–1427.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence An, Kathy J Goggin, and Delwyn Catley. 2017. Predicting counselor behaviors in motivational interviewing encounters. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1128–1137.

J Christopher Perry, Elisabeth Banon, and Floriana Ianni. 1999. Effectiveness of psychotherapy for personality disorders. *American Journal of Psychiatry*, 156(9):1312–1321.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Karan Singla, Zhuohao Chen, Nikolaos Flemotomos, James Gibson, Dogan Can, David C Atkins, and Shrikanth Narayanan. 2018. Using prosodic and lexical information for learning utterance-level behaviors in psychotherapy. In *Proc. Annual Conference of the International Speech Communication Association*, pages 3413–3417.

Substance Abuse and Mental Health Services Administration. 2019. *Key substance use and mental health indicators in the United States: Results from the 2018 National Survey on Drug Use and Health*. Center for Behavioral Health Statistics and Quality, Rockville, MD.

Michael Tanana, Kevin A Hallgren, Zac E Imel, David C Atkins, and Vivek Srikumar. 2016. A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of substance abuse treatment*, 65:43–50.

Shao-Yen Tseng, Sandeep Nallan Chakravarthula, Brian R Baucom, and Panayiotis G Georgiou. 2016. Couples behavior modeling and annotation using low-resource lstm language models. In *INTERSPEECH*, pages 898–902.

T.M. Vallis, B.F. Shaw, and K.S. Dobson. 1986. The cognitive therapy scale: psychometric properties. *Journal of consulting and clinical psychology*, 54(3):381–385.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Chien-Sheng Wu, Steven Hoi, Richard Socher, and Caiming Xiong. 2020. Tod-bert: Pre-trained natural language understanding for task-oriented dialogues. *arXiv preprint arXiv:2004.06871*.

Bo Xiao, Daniel Bone, Maarten Van Segbroeck, Zac E Imel, David C Atkins, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2014. Modeling therapist empathy through prosody in drug addiction counseling. In *Proc. Annual Conference of the International Speech Communication Association*.

Bo Xiao, Zac E Imel, Panayiotis G Georgiou, David C Atkins, and Shrikanth S Narayanan. 2015. "Rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing. *PloS one*, 10(12).

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.