TIM: A Tool for Gaining Insights into Psychotherapy

Ronan Cummins
Ieso Digital Health
Cambridge, UK
r.cummins@iesohealth.com

Michael P. Ewbank
Ieso Digital Health
Cambridge, UK
m.ewbank@iesohealth.com

Alan Martin
Ieso Digital Health
Cambridge, UK
a.martin@iesohealth.com

Valentin Tablan
Ieso Digital Health
Cambridge, UK
v.tablan@iesohealth.com

Ana Catarino
Ieso Digital Health
Cambridge, UK
a.catarino@iesohealth.com

Andrew D. Blackwell
Ieso Digital Health
Cambridge, UK
a.blackwell@iesohealth.com

ABSTRACT

We introduce and demonstrate the usefulness of a tool that automatically annotates therapist utterances in real-time according to the therapeutic role that they perform in an evidence-based psychological dialogue. This is implemented within the context of an on-line service that supports the delivery of one-to-one therapy. When combined with patient outcome measures, this tool allows us to discover the active ingredients in psychotherapy. In particular, we show that particular measures of therapy content are more strongly correlated with patient improvement than others, suggesting that they are a critical part of psychotherapy. As this tool gives us interpretable measures of therapy content, it can enable services to quality control the therapy delivered. Furthermore, we show how specific insights can be presented to the therapist so they can reflect on and improve their practice.

CCS CONCEPTS

• **Information systems** → **Decision support systems**; • **Computing methodologies** → *Discourse, dialogue and pragmatics*; • **Applied computing** → *Health care information systems.*

KEYWORDS

Psychotherapy; Natural Language Processing; Deep Learning; Telemedicine; eHealth

ACM Reference Format:
Ronan Cummins, Michael P. Ewbank, Alan Martin, Valentin Tablan, Ana Catarino, and Andrew D. Blackwell. 2019. TIM: A Tool for Gaining Insights into Psychotherapy. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308558.3314128

1 INTRODUCTION

Mental health is a substantial healthcare concern for the world's population. It is estimated that one in four people are affected by a mental illness during their lifetime and that mental illness costs the global economy $1 trillion per year in lost productivity alone. Psychotherapy is an effective intervention for a range of mental health problems. In particular, cognitive behavioural therapy (CBT) is a popular and effective evidence-based talking therapy that teaches patients the skills and techniques needed to help overcome and manage their symptoms. While CBT has been shown to be as effective as medication on a range of disorders, only half of those diagnoseable with a specific condition actually recover. In fact, the recovery rates for CBT have largely been stagnant, or even decreasing [8], in the last 30 years. Furthermore, given that the vast majority of therapy happens behind closed doors, it is difficult to objectively measure the quality of therapy being delivered or to identify the aspects of therapy that are most effective.

While there has been a technological revolution in the area of physical health in the last 50 years, where practitioners have been equipped with more accurate instruments, diagnostic tests, and medical devices (e.g. x-rays, MRI, blood tests, DNA sequencing), there has been very little change in how psychotherapy has been delivered. However, it is increasingly being recognised that online applications have a crucial role to play in increasing access to services, alleviating the burden on existing mental health services, and in driving innovation in the mental health domain.

One recent development in the mental healthcare sector in the UK is the use of online systems to support the delivery of psychotherapy. One particular case is internet-enabled CBT (IECBT) which allows a therapist to deliver CBT to a patient over their mobile device via real-time text-based conversation. This is an important development as it means that every interaction between a therapist and patient is recorded. In this work, we present a tool[1] that uses deep learning to automatically categorise therapist utterances according to the role that they play in therapy. In order to train and evaluate our model, human annotations were gathered from 290 hours of therapy. We then applied our tool to a large-scale dataset containing session transcripts from over 14,000 cases of IECBT (approximately 90,000 hours of therapy) to calculate a quantifiable measure of treatment delivered. Finally, using a logistic regression analysis, we investigated the relationship between the quantity (or âĂŸdosageâĂŹ) of each aspect of therapy delivered and clinical outcomes.

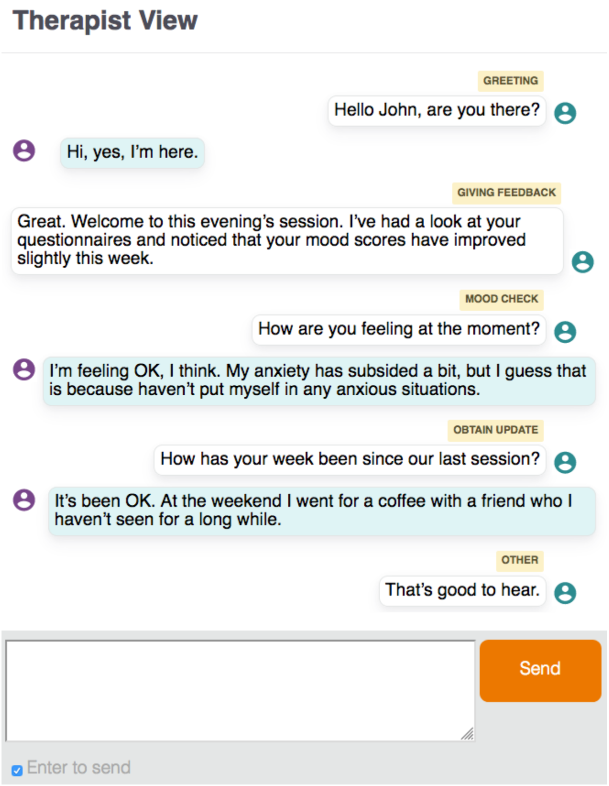
WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3314128

[1] https://youtu.be/TSfuxYctyik

2 BACKGROUND AND ONLINE CBT

While there has been an increase in online forums and web-based applications for mental healthcare, in most cases these deliver self-help materials for patients with mild symptoms and are not a replacement for one-to-one therapy. In IECBT, a patient communicates one-to-one with a qualified CBT therapist using a real-time text-based message system. IECBT has been shown to be clinically effective for the treatment of depression [5] and is currently commissioned in the UK by the National Health Service (NHS) under the Improving Access to Psychological Therapies (IAPT) [2] programme. As CBT is a short-term goal-orientated intervention, the typical number of sessions is between five and 20. As mandated by IAPT, symptoms are measured before each therapy session using self-reported instruments i.e. Patient Health Questionnaire (PHQ-9) and Generalised Anxiety Disorder Assessment (GAD-7). An example segment of therapy delivered using this modality is shown in Figure 1.



Hello John, are you there?

Hi, yes, I'm here.

Great. Welcome to this evening's session. I've had a look at your questionnaires and noticed that your mood scores have improved slightly this week.

How are you feeling at the moment?

I'm feeling OK, I think. My anxiety has subsided a bit, but I guess that is because haven't put myself in any anxious situations.

How has your week been since our last session?

It's been OK. At the weekend I went for a coffee with a friend who I haven't seen for a long while.

That's good to hear.

Figure 1: A screenshot of an example therapy session with labels being applied in real-time to each therapist utterance by our deep learning model (TIM).

Some related work [6, 11] has developed a dialog-act modelling approach for motivational interviewing conducted on spoken transcripts; the aim of which is to measure the fidelity of the session to the therapy model. This differs from our work in terms of the psychotherapy domain, the modality, and the fact that they do not aim to discover which aspects of therapy are most effective.

Table 1: Categories of therapist utterances (* means that the category is stylistic and was determined by a regular expression)

Therapist Utterance Categories	
Mood Check	Planning for the Future
Obtain Update	Elicit Feedback
Bridge	Summarise Session
Risk Check	Greeting
Set Agenda	Arrange next Session
Review Homework	Goodbye
Set Goals	Other
Formulation	*Socratic Questioning
Give Feedback	*Therapeutic Thanks
Change Methods	*Therapeutic Empathy
Perceptions of Change	*Therapeutic Praise
Set Homework	*Collaboration

3 HUMAN ANNOTATION

In conjunction with an experience clinically trained CBT practitioner, a research psychologist developed a set of 19 content-based categories and five stylistic categories that capture the role that therapist utterances play in CBT. The research psychologist annotated 290 hours of therapy (11,221 utterances) applying one or more of the 19 content-based categories to each utterance. The five stylistic categories were identified using regular expressions. A list of all the therapy categories used in the annotation effort are shown in Table 1. Using the annotated data, we trained a deep learning model using 8,859 utterances from 230 therapy sessions with the remaining data being kept for validation (30 transcripts) and testing purposes (30 transcripts). The deep learning model is described in the next section.

4 DEEP LEARNING ARCHITECTURE

We developed a deep learning model (see Figure 2) to automatically classify each utterance into one or more of the 24 categories. Firstly, we used word2vec [10] on a preprocessed version of the entire data set of over 90,000 transcripts (approx. 200M words) to learn word embeddings that are suited to the domain of psychotherapy. We preprocessed our data by tokenizing according to whitespace and punctuation and then by lower-casing all tokens (we kept punctuation symbols as separate tokens). This resulted in a vocabulary of 89,260 words, each represented as a continuous dense vector of length 200.

We modelled each utterance in a transcript as a sequence of word embeddings and fed these into a bidirectional long short-term memory [7] (BiLSTM). We used max-pooling over the hidden states of the BiLSTM to encode each utterance as a fixed-length vector. In order to model each utterance in the context of the entire transcript, we fed each of the fixed-length utterance representations in a transcript into another BiLSTM and used the hidden state at each time-step to feed into our final output layer. For both our BiLSTM stages, we used a hidden dimension of 400 and used dropout of 0.5 [4]. A fully connected layer maps each fixed length utterance-in-context representation into a vector of length 23 (23 classes with

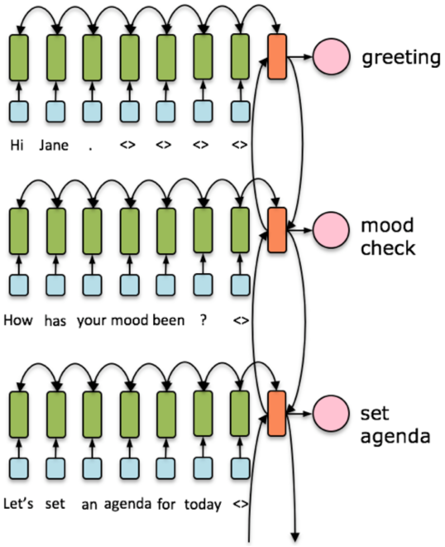


Figure 2: A diagram of the deep learning architecture where word representations (blue) are transformed via an LSTM into word-in-context representations (green) which are max-pooled into utterance representations (orange) and are fed into another LSTM to create utterance-in-context representations (pink).

the 'Other' class modelled as all zeros) with a sigmoid activation function used on this output. Conceptually, each category is modelled as a binary classifier. The training labels were obtained from manual annotation for 19 of the categories, and from high-precision regular expression for the remaining five categories. We call our model TIM (Therapy Insights Model).

5 EXPERIMENTS

We compared our model to a number of other baseline systems for sentence classification. As a simple baseline, we used a multilayer perceptron (MLP) that uses simple unigram lexical features as input. This MLP had one hidden layer the size of which was tuned on the validation set. As a more advanced baseline we used StarSpace [12] which is reported to outperform FastText [9] on a number of classification tasks. On the validation set, we found that a bigram model worked best. We also used a BiLSTM with max-pooling as another baseline [3]. For this baseline, we tuned the hidden dimension and dropout on the validation set. These three baselines treat each utterance in isolation and have no wider contextual information available during classification.

5.1 Classification Results

Table 2 shows the results (F1-score) of the various automated approaches to the utterance annotation task for the 19 manually annotated categories. The MLP that uses the simple lexical features is the worst performing model. The best non-contextual model is the BiLSTM. Overall, the best performing model is TIM which uses contextual information from the wider transcript when annotating an utterance.

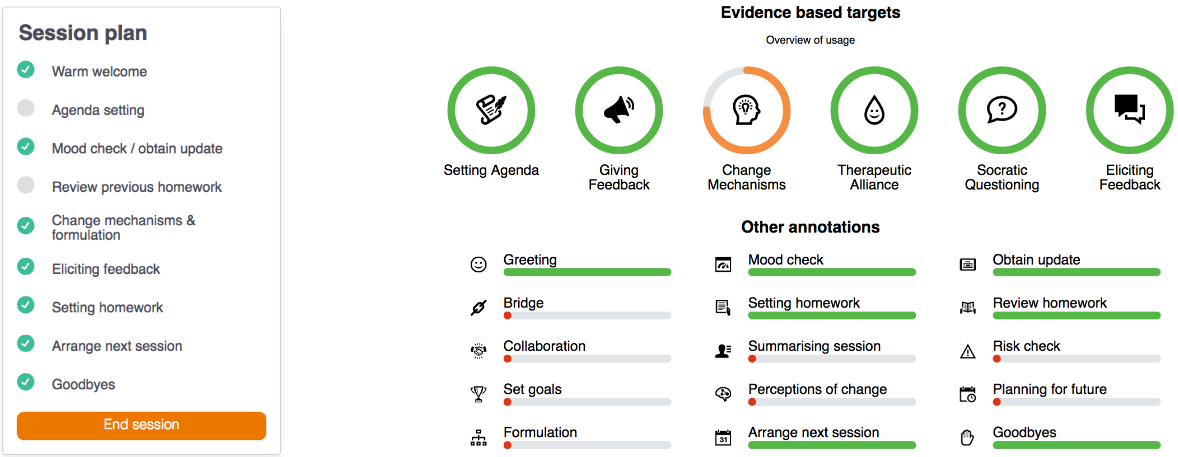
Table 2: F1 scores for the manually annotated categories on test data for the three baselines and TIM.

Category	#	MLP	SS	BiLSTM	TIM
Arrange next Session	45	0.68	0.84	0.86	0.87
Bridge	12	0.50	0.52	0.60	0.56
Change Methods	428	0.62	0.69	0.65	0.68
Elicit Feedback	42	0.30	0.34	0.55	0.61
Formulation	44	0.35	0.39	0.54	0.65
Give Feedback	74	0.26	0.39	0.37	0.37
Goodbye	41	0.84	0.91	0.87	0.92
Greeting	28	0.81	0.89	0.95	1.00
Mood Check	24	0.55	0.56	0.66	0.68
Obtain Update	35	0.37	0.43	0.52	0.61
Other	293	0.52	0.60	0.58	0.63
Perceptions of Change	20	0.12	0.37	0.25	0.25
Planning for the Future	24	0.20	0.21	0.35	0.49
Review Homework	37	0.21	0.43	0.39	0.40
Risk Check	15	0.52	0.81	0.70	0.72
Set Agenda	53	0.54	0.63	0.64	0.77
Set Goals	12	0.25	0.29	0.59	0.64
Set Homework	52	0.29	0.44	0.47	0.60
Summarise Session	0	n/a	n/a	n/a	n/a
macro-avg-F1		**0.44**	**0.54**	**0.58**	**0.64**
micro-avg-F1		**0.39**	**0.53**	**0.56**	**0.66**

6 OUTCOME RELATED INSIGHTS

In order to understand the therapy features that are associated with reduced symptoms, we performed a logistic regression on statistically reliable improvement. Patients are defined as having reliably improved if they have dropped by 6 points on the PHQ-9 scale or 4 points on the GAD-7 scale (with no reliable increase in the other) [2]. Patients scoring above 9 on the PHQ-9 or above 7 on the GAD-7 were classed as meeting the clinical threshold at assessment and were included in this regression analysis (approx. 14,000 cases). The TIM tool was used to automatically annotate all sessions (excluding the final session) for those patients (approx. 90,000 hours of therapy). Then the number of words of each therapy feature in Table 1, averaged over all sessions attended by a patient (excluding the final session), was used as a measure of the 'dosage' of that feature administered to that patient (case), while controlling for the number of sessions and other patient variables [1].

The results revealed a significant relationship between a number of features and statistically reliable improvement. 'Therapeutic Praise', 'Planning for the Future', 'Perceptions of Change', 'Set Agenda', 'Elicit Feedback', 'Review Homework' and 'Give Feedback' were all positively associated with reliable improvement. 'Change Methods' - a CBT-specific feature of therapy was also positively related with improvement. By contrast, increases in'non-therapy' related content ('Other', 'Greeting' and 'Goodbye') were negatively related with improvement, as were 'Therapeutic Empathy', 'Risk Check' and 'Bridge'. The output from this analysis can be used in a number of important clinical ways, including quality control, clinical decision support, and clinical supervision.



(a) clinical decision support

(b) therapist feedback

Figure 3: A screen-shot of in-session clinical decision support (a) and of therapist feedback (b) after a particular session (green indicates a suitable amount of that particular content was delivered).

6.1 Automated Quality Control

The output of the logistic regression model enables automated quality assurance. Using the logistic regression model as a predictive model, we can estimate the probability of recovery for any therapy session based on the quantity of the specific categories. Space limits the inclusion of a screen-shot of the output of a daily ranking of sessions.

6.2 Clinical Decision Support

The real-time nature of the automatic annotation tool allows us to monitor the delivery of crucial elements in live therapy. For example, therapists can be reminded to set an agenda or to set homework when they have not done so. Figure 3a shows an example of this in terms of a session plan which is updated as items are addressed by the therapist.

6.3 Automated Clinical Supervision

Clinical supervision allows therapists to discuss their practice with an experienced supervisor as part of their professional development. Given the data-driven nature of our insights, we can supply our therapists with summaries and insights for every session that they deliver. Figure 3b is an example of a session summary presented after a session which indicates which aspects of therapy are sufficient and which may be lacking. This can help therapists reflect on the specific session or on the therapy they deliver more generally.

7 CONCLUSION

This paper demonstrates a tool (TIM) that automatically annotates CBT transcripts. Along with paired outcome data, it engenders a sea-change in how psychotherapy can evolve in the digital age. We have demonstrated how it can be used to inform clinical best practice and furthermore, we outline three use cases for how this technology can be used in practice.

REFERENCES

[1] Ana Catarino, Sarah Bateup, Valentin Tablan, Katherine Innes, Stephen Freer, Andy Richards, Richard Stott, Steven D Hollon, Samuel Robin Chamberlain, Ann Hayes, et al. 2018. Demographic and clinical predictors of response to internet-enabled cognitive–behavioural therapy for depression and anxiety. *BJPsych open* 4, 5 (2018), 411–418.

[2] David M Clark. 2011. Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: the IAPT experience. *International review of psychiatry* 23, 4 (2011), 318–327.

[3] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364* (2017).

[4] Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*. 1019–1027.

[5] Tad Hirsch, Christina Soma, Kritzia Merced, Patty Kuo, Aaron Dembe, Derek D. Caperton, David C. Atkins, and Zac E. Imel. 2018. "It's Hard to Argue with a Computer": Investigating Psychotherapists' Attitudes Towards Automated Evaluation. (2018), 559–571.

[6] Tad Hirsch, Christina Soma, Kritzia Merced, Patty Kuo, Aaron Dembe, Derek D Caperton, David C Atkins, and Zac E Imel. 2018. It's hard to argue with a computer: Investigating Psychotherapists' Attitudes towards Automated Evaluation. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, 559–571.

[7] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[8] Tom J Johnsen and Oddgeir Friborg. 2015. The effects of cognitive behavioral therapy as an anti-depressive treatment is falling: A meta-analysis. *Psychological Bulletin* 141, 4 (2015), 747.

[9] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759* (2016).

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[11] 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 (2016), 43–50.

[12] Ledell Yu Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. 2018. Starspace: Embed all the things!. In *Thirty-Second AAAI Conference on Artificial Intelligence*.