ExTRA: Explainable Therapy-Related Annotations

Mat Rawsthorne
University of Nottingham

Tahseen Jilani
HDR-UK
Digital Research Service
University of Nottingham

Jacob Andrews
University of Nottingham

Yunfei Long
University of Essex

Jérémie Clos
University of Nottingham

Sam Malins
Institute of Mental Health
University of Nottingham

Daniel Hunt
University of Nottingham

Abstract

In this paper we report progress on a novel explainable artificial intelligence (XAI) initiative applying Natural Language Processing (NLP) with elements of co-design to develop a text classifier for application in psychotherapy training and practice. The task is to produce a tool that will automatically label psychotherapy transcript text with levels of interaction for patient activation in known psychological processes. The purpose is to enable therapists to review the effectiveness of their therapy session content. We use XAI to increase trust in the model's suggestions and predictions of the client's outcome trajectory. After pre-processing of the language features extracted from professionally annotated therapy session transcripts, we apply a supervised machine learning approach (CHAID) to classify interaction labels (negative, neutral or positive in terms of patient activation). Weighted samples are used to overcome class imbalanced data. The results show this initial model can make useful distinctions among the three labels of patient activation with 74% accuracy and provide insight into its reasoning. This ongoing project will additionally evaluate which XAI approaches are best for increasing the transparency of the tool to end users and explore whether direct involvement of stakeholders improves usability of the XAI interface and therefore trust in the solution.

1 Introduction

It takes a lot of manual effort to quality-assure psychotherapy sessions (Tseng et al., 2017), and therefore assessments of quality are rarely used routinely in psychotherapy practice. This work seeks to produce a tool that can automatically code psychotherapy transcripts, in line with a coding scheme developed by psychotherapists, known to characterise predictors of recovery (Malins et al., 2020a). The tool is also being developed to present explanations of the reasons for the coding decisions it makes. Explaining algorithms to those taking actions based on their outputs is recognised as good practice in data-driven health and care technology (DHSC, 2019). The ExTRA-PPOLATE [1] project is the first step in building tools to optimise scarce resources for provision of mental healthcare (Lorenzo-Luaces et al., 2017) by enabling therapists to adhere to good practice (Waller and Turner, 2016) and deliver care tailored to the patient (Delgadillo et al., 2016).

2 Overall Aims of Programme

The long-term objectives that we aim to achieve throughout our programme are threefold:

Aim 1 To build the foundation for unobtrusive, objective, transdiagnostic measures of patient activation.

Aim 2 To understand the practical trade-offs between classifier accuracy and explainability.

Aim 3 To explore the relationship amongst co-production, transparency and trust in algorithm-informed clinical decision making.

3 Methods

Core project team members were separately surveyed as to their initial hypotheses for key language markers of client-therapist interaction that is deemed helpful, focusing on generating features from different perspectives (see Table 1). These were then reviewed by the whole team and coded into a Python script to extract them from a corpus of transcripts of 120 health anxiety sessions. This created a simple model for identifying key interaction-types of interest (engagement in particular types of conversation) which are predictive of

[1] Explainable Therapy Related Annotations: Patient & Practitioner Oriented Learning Assisting Trust & Engagement

Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2020), pages 11–15,
Dublin, Ireland, 18 December 2020. ©2020 Association for Computational Linguistics

clinical outcomes (Malins et al., 2020a) and could be compared to detailed labels that had been applied to the data by specialist raters in nVivo using the Clinical Interaction Coding Scheme (CICS) (Malins et al., 2020b). Further information on this coding scheme is provided in Appendix A.

4 Data Analysis

Data distribution and model selection The data was skewed, and it was necessary to collapse some similar categories to ensure sufficient representation. We employed Chi-square Automatic Interaction Detection (CHAID), a type of decision tree (DT) classification model that can handle both categorical and numeric data sets. It does not require common statistical assumptions such as normality and non-collinearity (Kass, 1980). For imbalanced data, DT models allow weighting samples according to their importance. A sub-category of the outcome variable having smaller number of samples is assigned higher weight than as compare to other category with larger number of samples. Since positive and neutral category ratings were more common in the dataset than negative ratings, negatively categorised data were weighted for balance.

How Decision Trees Work DT models work by recursively partitioning the samples into a number of subsets. The starting node (at the top of the tree) is termed as "root". Any node with outgoing nodes is termed as an internal node, while the nodes without further branches are called "leaves". At each node, the Chi-square test for association is applied and the variable having the strongest association with the outcome variable is selected for further split into leaves. The chosen variable is the one that expresses the strongest discrimination between the different levels of outcome variable. The algorithm keeps dividing the full data set into subsets using the depth-first approach until the stopping criterion is not met (Magidson, 1994).

Validation For internal validation of the model or when no validation data set is available, the model can perform K-fold cross validation. Finally, the results from different K folds were merged to produce a single DT estimation. DT models also offer tree pruning and feature selection based on the Chi-Squared test to prevent overfitting of the model. A "minimum cases" criteria is used for deciding further split of a branch. Discrimination of the original and cross-validated models was evaluated through the generation of Receiver Operating Characteristic (ROC) curves and calculation of C-statistics.

5 Initial Findings

CHAID label classification results are summarised in the classification matrix in Table 2. The overall accuracy of the model was 74% with the highest correct sample classified in the Neutral category. There were a total of 681 negative labels out of a total of 25,823 samples (2.6%). Of these, 60.4% were correctly classified. A larger total of 16,713 samples were recorded for positive labels (64.7% of the total), with a correct classification rate of 69.5%. The performance of the classification could be further enhanced through a more detailed exploration of the language features from the session transcripts, using improved oversampling techniques such as SMOTE and deeper machine learning modelling such as random forest and convolutional neural networks. Furthermore, the interdisciplinary engagement with the data has already helped deepen understanding of both the CICS framework and the classifier model (Páez, 2019) and generated ideas for their refinement.

6 Tool Development

The project uses a fusion of techniques to apply Responsible Research & Innovation (RRI) to the tool's development, specifically:

Incorporating a range of perspectives at multiple levels: The core project team combines the lived experience of a Service User Researcher and Involvement Volunteers (skilled in instrumentation design and plain English summaries) from the Institute of Mental Health, with specialist Clinical Psychology knowledge, Statistical Machine Learning, Psychometrics, Computer Science and Corpus Linguistics expertise. This diversity of experts in the formal and informal language of mental health provide triangulation to ensure the methods and findings make sense (Ernala et al., 2019). Additionally we engaged a Patient & Practitioner Reference Group (PPRG), comprised of 12 people, balanced across key stakeholder groups: patients and carers, clinical psychologists, therapy trainers, and mental health service managers. Dissemination will be via interactive 'roadshow' events with PPRG peer groups to gauge whether they feel the

Perspective	Feature	Impact	Coding
Patient	absolute words, profanity	negative	customised dictionaries
Clinician	positive sentiment	positive	valence and polarity
Linguist	first person pronouns	negative	ratio singular:plural
NLP researcher	utterance length	positive	word, character counts

Table 1: Table Examples of Candidate Language Features
Perspective: professional alignment of the core project team member suggesting the language feature.
Impact: expected relationship between the feature and level of patient engagement in the interaction.
Coding: method used to extract from the text using Python [details available from authors on request].

Observed	Predicted Negative	Predicted Neutral	Predicted Positive	Percent Correct
Negative	411	94	176	60.4%
Neutral	99	7,766	564	92.1%
Positive	1,223	3,871	11,619	69.5%
Overall Percentage	6.7	45.4	47.9	**74%**

Table 2: Initial Results for Classification of Level of Clinical Engagement

co-design process adds to the credibility of the tool.

Agile Science Approach (Hekler et al., 2016) Repeated engagement with end-users is intended to build trust (Carr, 2020) and emulates industry best practice. The project leverages specialist support from a social enterprise[2] on coproduction aspects (Hickey et al., 2018), and a digital health industry partner[3] on user experience (UX) design.

Collaborative (Machine) Learning in the tool and the process: In combination with the agile, participatory approach, the use of Human-in-the-Loop techniques will enable refinement of definitions and expose and explore tacit and latent knowledge in assessment of psychotherapy through direct involvement of domain experts in model development. Through prototyping a person-centred active learning process, we anticipate a two-way exchange of insights which will clarify what helps and what hinder the psychotherapy process.

Using evidence-based tools to capture key considerations: TrustScapes[4] were used to identify the core factors contributing to trust throughout the model pipeline (data, processing, deployment). Combined with a PROSOCIAL approach[5], this elicited fundamental stakeholder requirements for the qualities of an engaging, interactive feedback interface, and actions needed to mitigate wider concerns about its acceptability. The Software Usability Scale Plus (SUS+ (Bangor et al., 2009)) will be used as a proxy metric to evaluate explainability, supplemented by detailed qualitative feedback through 'think-aloud' exercises (Garcia et al., 2018). Measurement of trust in XAI is a new and developing field (Hoffman et al., 2019; Jacovi et al., 2020; Mohseni et al., 2020) whereas the SUS+ is well established in Human Computer Interaction, seen as a practical compromise to capturing important aspects (Davis et al., 2020) (usability and trust are interdependent (Acemyan and Kortum, 2012)), and has been used as the basis of measures of quality of explanations (Holzinger et al., 2020).

Future directions: The first PPRG workshop also started the process of gathering feedback on what is a good explanation (Danilevsky et al., 2020) and recommendation format, from each perspective (Arya et al., 2019). Over the next 3 months, we will continue to refine the classification tool, and then use it to accelerate annotation of motivation in turns-of-speech in a separate research dataset of anonymised transcripts of mainstream counselling for depression, and update the classification algorithm to increase generalisability of the tool (Topol, 2020). Given that behaviour change involves a degree of persuasion, we will explore whether we can leverage insights from Argumentation Theory to augment the model (Clos et al., 2014) and other, related developments in the field of NLP for mental

[2] Academy for Recovery Coaching CIC

[3] https://virtualhealthlabs.com/

[4] https://UnBIAS.wp.Horizon.ac.uk/fairness-toolkit/

[5] https://prosocial.world

health (e.g. unobtrusive measures of psychological inflexibility (Berkout et al., 2020), empathy (Sharma et al., 2020)). Using Natural Language Generation (NLG) for XAI (Reiter, 2019) we will test whether the model can provide its rationale in plain English matched to terms each perspective understands (Tomsett et al., 2018). We will be exploring the different use cases of justification, improvement, control and discovery (Adadi and Berrada, 2018), and investigating how the predictive ability of engagement language markers relate to those of symptomatology (Losada et al., 2019).

Acknowledgments

This work is funded by a Engineering and Physical Sciences Research Council's (EPSRC) Human Data Interaction (HDI) Network "Future of Mental Health" [6] and support from Health Data Research - UK[7].

This report is supported by the National Institute for Health Research[8] (NIHR Development and Skills Enhancement Award, Dr Sam Malins, NIHR300822). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care.

References

Claudia Ziegler Acemyan and Philip Kortum. 2012. The relationship between trust and usability in systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56(1):1842–1846.

A. Adadi and M. Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques.

Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual sus scores mean: Adding an adjective rating scale. *J. Usability Studies*, 4(3):114–123.

Olga V. Berkout, Angela J. Cathey, and Dmytry V. Berkout. 2020. Inflexitext: A program assessing psychological inflexibility in unstructured verbal data. *Journal of Contextual Behavioral Science*, 18:92–98.

Sarah Carr. 2020. 'ai gone mental': engagement and ethics in data-driven technology for mental health. *Journal of Mental Health*, 29(2):125–130.

Jérémie Clos, Nirmalie Wiratunga, Joemon Jose, Stewart Massie, and Guillaume Cabanac. 2014. Towards argumentative opinion mining in online discussions. In *Proceedings of the SICSA Workshop on Argument Mining (the Scottish Informatics & Computer Science Alliance)*, page 10.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable ai for natural language processing.

Brittany Davis, Maria Glenski, William Sealy, and Dustin Arendt. 2020. Measure utility, gain trust: Practical advice for xai researcher.

Jaime Delgadillo, Omar Moreea, and Wolfgang Lutz. 2016. Different people respond differently to therapy: A demonstration using patient profiling and risk stratification. *Behaviour Research and Therapy*, 79:15–22.

DHSC. 2019. Code of conduct for data-driven health and care technology.

Sindhu Kiranmai Ernala, Michael L. Birnbaum, Kristin A. Candan, Asra F. Rizvi, William A. Sterling, John M. Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 134. ACM.

Francisco Javier Chiyah Garcia, David A Robb, Xingkun Liu, Atanas Laskov, Pedro Patron, and Helen Hastie. 2018. Explainable autonomy: A study of explanation styles for building clear mental models. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 99–108.

Eric B. Hekler, Predrag Klasnja, William T. Riley, Matthew P. Buman, Jennifer Huberty, Daniel E. Rivera, and Cesar A. Martin. 2016. Agile science: creating useful products for behavior change in the real world. *Translational behavioral medicine*, 6(2):317–328.

G Hickey, S Brearley, T Coldham, S Denegri, G Green, S Staniszewska, D Tembo, K Torok, and K Turner. 2018. Guidance on co-producing a research project. *Southampton: INVOLVE*.

Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2019. Metrics for explainable ai: Challenges and prospects.

[6] https://hdi-network.org/fmh/

[7] https://www.hdruk.ac.uk/

[8] https://www.nihr.ac.uk/

Andreas Holzinger, André Carrington, and Heimo Müller. 2020. Measuring the quality of explanations: The system causability scale (scs). *KI - Künstliche Intelligenz*, 34(2):193–198.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2020. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai.

Gordon V Kass. 1980. An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(2):119–127.

Lorenzo Lorenzo-Luaces, Robert J. DeRubeis, Annemieke van Straten, and Bea Tiemens. 2017. A prognostic index (pi) as a moderator of outcomes in the treatment of depression: A proof of concept combining multiple variables to inform risk-stratified stepped care models. *Journal of Affective Disorders*, 213:78–85.

David E Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of erisk at clef 2019: Early risk prediction on the internet (extended overview). In *CLEF (Working Notes)*.

Jay Magidson. 1994. The chaid approach to segmentation modeling: Chi-squared automatic interaction detection. In Richard P. Bagozzi, editor, *Advanced methods of marketing research*, pages 118–159. Blackwell, Cambridge, UK.

Sam Malins, Nima Moghaddam, Richard Morriss, and Thomas Schröder. 2020a. Extending the use of routine outcome monitoring: Predicting long-term outcomes in cognitive behavioral therapy for severe health anxiety. 30(5):662–674. PMID: 31438807.

Sam Malins, Nima Moghaddam, Richard Morriss, Thomas Schröder, Paula Brown, Naomi Boycott, and Chris Atha. 2020b. Patient activation in psychotherapy interactions: Developing and validating the consultation interactions coding scheme (cics). *Journal of Clinical Psychology*, 76(4):646–658.

Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2020. A multidisciplinary survey and framework for design and evaluation of explainable ai systems.

Andrés Páez. 2019. The pragmatic turn in explainable artificial intelligence (xai). *Minds and Machines*, 29(3):441–459.

Ehud Reiter. 2019. Natural language generation challenges for explainable ai.

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support.

Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. Interpretable to whom? a role-based model for analyzing interpretable machine learning systems.

Eric J. Topol. 2020. Welcoming new guidelines for ai clinical research. *Nature Medicine*, 26(9):1318–1320.

Shao-Yen Tseng, Brian R. Baucom, and Panayiotis G. Georgiou. 2017. Approaching human performance in behavior estimation in couples therapy using deep sentence embeddings. In *INTERSPEECH*, pages 3291–3295.

Glenn Waller and Hannah Turner. 2016. Therapist drift redux: Why well-meaning clinicians fail to deliver evidence-based therapy, and how to get back on track. *Behaviour Research and Therapy*, 77:129–137.

Appendix A. Further information on CICS

A recently developed tool was deemed suitable for automation using NLP because of its focus on turn-by-turn language use in psychological therapy. The Consultation Interaction Coding Scheme (Malins et al., 2020b) offers reliable turn-by-turn assessment of interaction-types, incorporating both client and therapist responses. Using the CICS, in-session therapist-client turns-of-speech are first categorized into one of seven interaction types (action planning; evaluations of self or therapy; information discussion; noticing change or otherwise; problem description; problem analysis; structuring) and then rated -2 to +2 based on the degree of patient activation observable in the interaction. Positive ratings indicate high patient activation and engagement; negative ratings indicate low patient activation and disengagement. A series of studies have now indicated that CICS-rated psychological therapy interactions at initial sessions predict wellbeing across the course of therapy and a range of health outcomes across 12-month follow-up. This means that language features in the interactions at the first sessions of psychological therapy predicted health anxiety, generalised anxiety, depression, quality of life, general health, functioning, and somatic symptoms up to 12 months later. Specifically, if clients gave more positive evaluations of themselves or the therapy at initial sessions then better outcomes followed. Similarly, where clients were more actively engaged in structuring initial sessions and choosing session tasks, health improvements were greater. Conversely, larger proportions of initial sessions spent describing problems (as opposed to more active discussion of what might be done with problems) predicted poorer outcomes.