Detecting defense mechanisms from Adult Attachment Interview (AAI) transcripts using machine learning

Anthony Tasca, BSc[1], Samantha Carlucci, PhD[2,3], James C. Wiley, BA[1], Matthew Holden, PhD[1], Ahmed El-Roby, PhD[1], & Giorgio A. Tasca, PhD[2,3*]

[1]Carleton University, School of Computer Science & Department of Psychology, Ottawa, Canada

[2]University of Ottawa, School of Psychology, Ottawa, Canada

[3]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Canada

*****Corresponding author: Giorgio A. Tasca, PhD, Email: gtasca@uottawa.ca

Author's Note:

James C. Wiley https://orcid.org/0000-0002-5049-573X

Giorgio A. Tasca: https://orcid.org/0000-0001-7596-0661

Abstract

Objective: Defensive functioning (i.e., unconscious process used to manage real or perceived threats) may play a role in the development of various psychopathologies. It is typically assessed via observer rating measures; however, human coding of defensive functioning is resource-intensive and time-consuming. The purpose of this study was to develop a machine learning approach to automate coding of defense mechanisms from interview transcripts. **Method:** Participants included a clinical sample of women with binge-eating disorder (n = 92) and a community sample without binge-eating disorder (n = 66). We trained and evaluated five RoBERTa-based models to detect the presence of defenses in 16,785 interviewer-participant talk-turn pairs nested within 192 interviews. Separate models were used to detect the presence of any defense and the four most common defenses in this sample (repression, intellectualization, reaction formation, undoing). **Results**: The models were capable of distinguishing defenses (ROC-AUC .82-.90), but were not proficient enough to warrant replacing human coders (PR-AUC .28-.60). Follow-up analysis was performed to assess other practical uses of these models. **Discussion:** Our machine learning models could be used to assist coders. Future research should conduct a deployment study to determine if human coding of defense mechanisms can be expedited using machine learning models.

Keywords: defensive functioning, machine learning, Defense Mechanism Rating Scale, conversation analysis, RoBERTa

Clinical or methodological significance of this article: To our knowledge, this is the first study to use state-of-the-art BERT-based language models to reliably detect some defense mechanisms (repression, intellectualization, reaction formation, undoing) from psychological interview

transcripts. These models may be used to substantially reduce workload and increase accessibility to observer coding for both researchers and clinicians.

Detecting defense mechanisms from Adult Attachment Interview (AAI) transcripts using machine learning

Clinical and research psychologists often rely on observer rating tools to assess unconscious cognitions or behaviors (MacLaren Chorney et al., 2015), as some aspects of these constructs may not be accessible by self-report. However, observer rating is impractical and resource intensive. Furthermore, there are limitations to human or clinical judgement (Yager at al., 2021). Novel approaches are needed to expedite behavioral/observer rating, and advance psychology research, clinical training, and practice.

Recent research has highlighted the value of machine learning in behavioral coding. Machine learning is a type of artificial intelligence whereby a computer algorithm learns to recognize patterns through repeated exposures (Aafjes-van Doorn et al., 2021; Idalski Carcone et al., 2019; Jordan & Mitchell, 2015). Machine learning has widespread utility, including natural language processing (NLP; Bird et al., 2009; Dalal & Zaveri, 2011) which transforms individual words or groupings of words from text into structured data (Pace et al., 2016; Weusthoff et al., 2018). This is often achieved through recurrent neural networks (RNNs), which are dynamic models that accumulate knowledge sequentially and are often applied in handwriting or speech recognition tasks (Kamath et al., 2019). Recent studies have used NLP and RNNs to content code medical and psychotherapy transcripts (Cummins et al., 2019; Ewbank et al., 2020; Ewbank et al., 2021; Gaut et al., 2015; Park et al., 2019). For example, Idalski Carcone and colleagues (2019) developed a machine learning model that accurately and reliably coded specific features of speech (e.g., reflections, affirmations, change talk) from patient-provider interactions in a sample with human immunodeficiency virus (HIV). Similarly, Crangle and colleagues (2019) used machine learning to code four emotions (anger, sadness, joy, and tension) from recordings

of couples in psychotherapy with relative accuracy. These recent advances highlight the potential utility of machine learning for behavioral coding in clinical research and practice (Cummins et al., 2019; Ewbank et al., 2020; Ewbank et al., 2021; Gaut et al., 2015; Park et al., 2019), particularly for complex unconscious processes like defensive functioning.

Defensive functioning is an unconscious process used to manage real or perceived emotional and cognitive threats (American Psychiatric Association [APA], 1994). Although inherently protective, defense mechanisms vary in terms of their adaptiveness and maturity. For instance, mature defenses allow the individual to acknowledge and effectively confront threats; whereas immature defenses hinder the individual's ability to acknowledge or resolve threats, and may generate additional stress (APA, 2000).

The most widely cited hierarchical model and observer-rating measure of defensive functioning is the Defense Mechanism Rating Scale (DMRS; Perry, 1990). The DMRS contains 30 defense mechanisms categorized into seven levels ranging from least adaptive (i.e., immature) to most adaptive (i.e., mature; Table 1), and is used to code defense mechanisms from recorded interviews and transcripts. The weighted average of all defenses in a single interview/transcript is used to generate an overall defensive functioning (ODF) score (Perry & Henry, 2004), which indicates one's general level or adaptiveness of defensive functioning.

[Table 1 near here]

The DMRS has been used to assess defensive functioning in various clinical (e.g., binge-eating disorder, personality disorders, breast cancer) and non-clinical or community samples (Carlucci et al., 2021; Perry et al., 2015; Perry et al., 2013). For example, Perry and colleagues (2013) found that adults with personality disorders tended to use immature defenses (e.g., devaluation, projection, rationalization, splitting, passive aggression). Furthermore, Carlucci and

colleagues (2021) found that women with binge-eating disorder (BED) had significantly lower ODF scores compared to a control sample of women without BED. Women with BED tended to use mostly mid-range/neurotic defenses (e.g., repression, undoing, intellectualization, reaction formation). Defensive functioning may also play a role in psychological treatment progression and outcomes. For instance, Høglend and Perry (1998) found that compared to patients with less adaptive defensive functioning at baseline, patients with better defensive functioning at baseline had greater improvements in depressive symptoms at six-months post-treatment. As such, it is important to consider the role of defensive functioning when treating patients with mental health problems.

A major benefit of the DMRS is that it facilitates a quantitative, observer-rating assessment of defensive functioning in interviews/transcripts by multiple trained raters (Perry et al., 1993). It can accurately capture aspects of implicit psychological processes that may not be accessible via self-report. Despite these advantages, human coding of defensive functioning is resource-intensive and time-consuming. It can take approximately four hours to code a single one-hour interview/transcript using the DMRS (Carlucci, 2022). These challenges present a barrier to the widespread use of the DMRS and delay the progression of research on defensive functioning and, more broadly, clinical psychology. As such, one goal of this study was to explore an alternative, more time efficient method of coding defenses, while maintaining an adequate level of reliability.

We developed and evaluated a novel machine learning approach to reduce manual workload or automate coding of defense mechanisms from Adult Attachment Interviews (AAI; George et al., 1985). Specifically, we used the Robustly Optimized Bidirectional Encoder Representations from Transformers Pre-Training Approach (RoBERTa; Liu et al., 2019) to build

machine learning models to detect the presence of any defense from AAI transcripts, as well as four separate defense mechanisms (repression, intellectualization, reaction formation, and undoing). Bidirectional Encoder Representations from Transformers-based models are pre-trained on enormous amounts of existing natural language corpora using a combination of modelling tasks (Delvin et al., 2019; Liu et al., 2019). For example, RoBERTa is pre-trained on over 160 GB of uncompressed text data, including all of Wikipedia English (Liu et al., 2019). This corpus can be used as a starting point for training new language-based tasks (Delvin et al., 2019). The effectiveness of our models was evaluated by their ability to accurately classify data when implemented in a new dataset (Aafjes-van Doorn et al., 2021; Bi et al., 2019).

Method

Participants

Participants were 158 women recruited from two studies: (1) an uncontrolled treatment study of women with BED (n = 92; Tasca et al., 2013), and (2) a psychological health study on a community sample of overweight/obese and normal weight women (n = 66; Maxwell et al., 2017). The participants' mean age was 43.64 years (SD = 11.63), and most were English-speaking (85.4%), White (90.1%), married (50.0%), employed full-time (67.5%), college or university educated (74.7%), and reported a median family income of $70,000 to $79,000 in Canadian dollars per year.

Some participants completed the AAI at two separate time-points (i.e., baseline/pre-treatment and six months post-treatment), and so there were 192 transcripts in total. Since the purpose of this study was to develop and test the reliability of five machine learning models, and not to assess treatment outcomes related to defensive functioning, we opted to use all available transcripts (i.e., regardless of time-point) to maximize the number of codable talk-turns.

Measures

Demographics

We used a modified version of the Diagnostic Survey for Eating Disorders (DSED; Johnson, 1985) to collect socio-demographic data on age, race/ethnicity, marital status, employment status, education, and median family income.

Defensive functioning

As previously stated, the DMRS (Perry, 1990) is an observer-rating measure that contains 30 defenses categorized into seven levels ranging from least adaptive (i.e., immature) to most adaptive (i.e., mature). As part of this study, we were interested in automating the coding of the four most common defenses (repression, intellectualization, reaction formation, and undoing) within our study sample as identified by human raters (Carlucci et al., 2021), as there were too few data from the other defenses to train a model specifically for them. To handle the detection of the additional 26 defenses from the DMRS, we also trained a binary classification model to detect whether any defense was present in a patient response. Intellectualization and undoing are Obsessional defenses used to separate affect from conscious idea or fact. Specifically, intellectualization allows one to generalize and detach from a painful affect, whereas undoing manifests as ambivalence and is used to minimize guilt. Repression and reaction formation are Other Neurotic defenses used to protect oneself from a threatening experience or idea. Specifically, repression unconsciously inhibits oneself from remembering a traumatic experience, whereas reaction formation substitutes an unwanted thought or feeling for an opposite thought or feeling (Perry, 1990). The DMRS has consistently demonstrated good reliability and validity (Carlucci et al., 2021; Di Giuseppe et al., 2020). The DMRS permits the coding of multiple defenses within the same talk-turn or statement.

Adult Attachment Interviews

The AAI (George et al., 1985) is a semi-structured interview used to assess adult representations of early childhood attachment to primary caregivers. For this study, we only used transcripts from the AAI to elicit and code current defenses with the DMRS.

Procedures

Women with BED were invited to participate in a treatment study via advertisements and through an eating disorders program at a local hospital. A research coordinator screened all interested participants to confirm a diagnosis of BED (Tasca et al., 2013). Only those with BED who met the following inclusion criteria were invited to participate: (a) were proficient in English, (b) did not engage in any purging behaviours, (c) had no history of substance abuse in the past six months, (d) had no comorbid diagnosis of bipolar or psychotic disorder, (e) were not currently pregnant and/or had no plans to become pregnant in the next year, and (f) were not enrolled in psychotherapy or a weight loss program. A community sample of overweight/obese and normal weight women were also recruited via local advertisements for a study on women's health (Maxwell et al., 2017). Again, a research coordinator screened interested participants for presence of BED and other symptoms of psychopathology. Only those who met the inclusion criteria and had no current diagnosis or history of BED were included in the community sample. Participants with BED were administered the AAI at two time-points (i.e., pre-treatment and six months post-treatment), whereas the non-clinical sample only completed the AAI at one time-point. Participants provided written informed consent prior to study enrolment. The XX, XX, and XX Research Ethics Boards approved the original protocols and the current study.

DMRS coding procedure

Eleven trained raters used the DMRS to code defensive functioning from AAI transcripts. The raters received formal training over two to four months, which included familiarizing oneself with the DMRS and its associated defenses, engaging in practice coding with audio recorded interviews and transcripts, and meeting adequate reliability with other raters and a defense mechanism expert. Perry and Henry (2004) recommended that each rater code a minimum of 10 training recordings/transcripts not associated with the current study. Once the raters achieved adequate reliability, they went on to code AAI audio recordings and transcripts for current use of defense mechanisms independently and also in rotating groups of two to three raters to maintain reliability. It can take four hours to code a single one-hour transcript (Carlucci, 2022). The transcripts were de-identified to mask interview time-point (baseline versus six months post-treatment) and weight/diagnosis (BED, overweight/obese without BED, normal weight without BED). All raters participated in supervisory calibration sessions throughout coding to discuss any concerns and prevent rater drift. (For more information on training guidelines, please refer to Perry and Henry [2004]).

Dataset generation procedure

The human coded transcripts were processed by a computer program to generate a dataset of interviewer-participant talk-turns to train and evaluate the machine learning models. The coded transcripts were stored as Microsoft Word document files using a pre-defined method for separating talk-turns and coding the content. The dataset generation program assumed that each talk-turn was separated into sequential paragraphs in the Word document. It also assumed that bolded paragraphs represented interviewer talk-turns and non-bolded paragraphs represented participant talk-turns. These parsing strategies aligned with the formatting of the transcripts. If a participant talk-turn paragraph had a comment associated with it in the Word document, then the

program searched that comment for the name of a defense mechanism and associated that defense with the given talk-turn. Again, this aligned with how raters recorded defense mechanisms. The resulting dataset contained an example for every interviewer question/participant answer pair present in the original transcripts, and a label for each of these examples. These question/answer pairs were then used in the machine learning procedure in which each interviewer talk-turn was assumed to be a question or follow-up question, and each participant talk-turn assumed to be an answer to that question.

Machine learning procedure

We used RoB-RT (Barbieri et al., 2020), a version of RoBERTa that is pre-trained on text from Twitter posts, to model the relationship between a question/answer pair and a coded defense mechanism. We used a model pre-trained on Twitter because it contains more conversational data than Wikipedia or other sources used to train RoBERTa. Thus, we expected it to produce a better statistical model of language for our specific task.

Figure 1 shows a high-level process flow of the machine learning procedure. The inputs were an interviewer question and participant response talk-turn pair. The tokenized pair or sequence was then fed into the RoB-RT model and the participants' response was classified as exhibiting or not exhibiting a given defense mechanism.

[Figure 1 near here]

A binary classification model was trained to detect the presence of any defense. This model outputted the probability that a given answer contained any of the defenses that could be coded by the DMRS. Separate binary classification models were also trained and evaluated to dectect the four defense mechanisms (repression, intellectualization, reaction formation, and undoing) of interest. Each model outputted the probability that a given answer contained one of

these defenses. We then tested different probability thresholds for determining whether a defense was present. If a given answer had a probability above said threshold, then it was classified as containing the specific defense (Kutner & Neter, 2005; Saito & Rehmsmeier, 2015). The models were trained for 5 epochs using a weighted binary cross entropy loss function, with a batch size of 24, and an AdamW (Loshchilov & Hutter, 2017) optimizer with a learning rate of $2e{-}5$ and an epsilon of $1e{-}8$. These hyperparameters were selected in accordance with recommendations from Delvin and colleagues (2019), who noted that a learning rate of $2e$-5 across 4 epochs worked well across multiple fine-tuning tasks. In this study, we used 5 epochs instead of 4 to allow for a buffer of additional learning at the cost of some computational resources. The weight applied to the positive class in the loss function is equal to the number of negative samples divided by the number of positive samples in the training data for each model. A batch size of 24 was used as this was the maximum batch size that could be achieved with the available hardware. A 10-fold stratified cross-validation was used to measure the area under the receiver operating characteristic curve (ROC-AUC), area under the precision-recall curve (PR-AUC), accuracy, true positive rate, false positive rate, true negative rate, false negative rate, and F_1 score of each model. Additionally, we enforced patient-out cross-validation (i.e. different talk-turns from the same patient were restricted to the same validation fold, so data from the same patient could never appear in both training and testing sets).

Typically, ROC-AUCs over .80 are considered excellent in diagnostic test assessment (i.e., the model correctly assigned higher probabilities to true positives; Mandrekar, 2010). However, this does not necessarily mean the model is able to perform accurate classification (Kutner & Neter, 2005; Saito & Rehmsmeier, 2015). The advantage of a PR-AUC is its interpretability; when used on an imbalanced dataset, the PR-AUC score of a random estimator

is equal to the positive class ratio (Saito & Rehmsmeier, 2015), providing a concrete baseline for comparison. Like ROC-AUC, higher PR-AUC scores indicate better model performance, though cut-offs for labelling quality are somewhat ambiguous. Accordingly, we first looked at the ROC-AUC to judge the theoretical performance of our models. Afterwards, we judged the practical capabilities of our models using the PR-AUC. The models with the highest PR-AUC score on a 15% validation set across the five epochs were selected for testing within each cross-validation fold. We also reported accuracy, which represents the sum of true positives (i.e., talk-turns that are correctly classified as having a defense) and true negatives (i.e., talk-turns that are correctly classified as not having a defense) divided by the total number of talk-turns. Finally, the F_1 score represents the harmonic mean of the precision (i.e., the number of correct positive predictions divided by the number of total positive predictions) and recall (i.e., the number of correct positive predictions divided by all positive samples) of a model. Unlike accuracy, the F_1 score does not consider true negatives but is more sensitive to false negative and false positive predictions. Classification thresholds were selected by maximizing the F_1 score on the validation data for each fold and then applying a macro average. For more information on each evaluation metric, please refer to Raschka and Mirjalili (2017) and Saito and Rehmsmeier (2015).

Results

Preliminary results

[Table 2 near here]

We used the (1, *k*) model (Koo & Li, 2016) to assess inter-rater reliability between pairs of human raters. Table 2 shows the corresponding intra-class correlations (ICCs) for each defense mechanism (repression, intellectualization, reaction formation, and undoing) across 65 reliability-coded interview transcripts. All ICC scores were above the threshold of 0.60,

indicating good inter-rater reliability (Landis & Gary, 1977). Carlucci and colleagues (2021) reported similar results on the same sample, with ICCs of 0.74 and 0.89 for the Obsessional (e.g., intellectualization, undoing) and Other Neurotic (e.g., repression, reaction formation) defense levels, respectively. These preliminary results provide evidence of reliability for the data used to train and evaluate the machine learning models.

Main results

[Table 3 near here]

A total of 16,875 talk-turns were identified across 192 AAI transcripts recorded at baseline/pre-treatment and six months post-treatment in a sample of women with and without BED. There was a highly skewed class imbalance in the data (as shown in Table 3).

[Table 4 near here]

Table 4 displays the macro-averaged cross-validation performance metrics for each model, along with the accuracy scores of a pessimistic model that always predicts the negative class. All our classifiers achieved a ROC-AUC score above 0.82, which is considered in the excellent range for diagnostic test assessments (Mandrekar, 2010). High ROC-AUC scores indicate that each model was able to distinguish answers that have a higher probability of containing a specific defense relative to all other defenses. The low false positive rates (i.e., the number of incorrect positive predictions divided by the total number of negative samples) indicated that the models will falsely identify a talk-turn that does not have a defense mechanism only 4% to 6% of the time.

The PR-AUC scores also showed a significant improvement of our models over random binary estimators, which would have scores ranging from 0.04 to 0.26. The accuracy of each model was high; though this metric is misleading in the presence of unbalanced data. For

example, the accuracy score of a pessimistic model for a given defense is simply 1 minus the Positive Ratio. This ends up producing approximately equivocal accuracy scores relative to our trained models (see Table 4). This indicates that our models should not be favourably evaluated by their high accuracy scores, given class imbalance inflates this metric. We note that low accuracy would be reason to suspect poor performance, but high accuracy is not necessarily suggestive of high performance in the face of unbalanced data. The F_1 scores of our models were more in line with the PR-AUC scores, since both metrics are more critical of false negative and false positive predictions.

[Table 5 near here]

The ROC-AUC scores are much higher than our F_1 and PR-AUC scores because F_1 and PR-AUC scores do not incorporate true negatives into their calculation, which are abundant in our highly unbalanced dataset. High ROC-AUC scores indicate that these models can correctly assign higher probabilities to talk-turns wherein there are defenses, while lower F1 and PR-AUC scores suggest fully automating the coding process using these models is not feasible.

To explore this concretely, we examined how each model performed using differing probability thresholds for classifying defenses. For example, if we use the probability threshold of 5%, any talk-turn with a predicted probability of 5% or greater is classified as having the defense. The resulting data are displayed in Table 5.

We note that—at a threshold of 50%—the true positive rates (recall) range from 41.8% to 72.8%. This means that, if we were to fully automate coding using these models, we would miss approximately three-fifths to one-fourth of defenses, depending on which model was used. This degree of error is too high for automated coding. We can adjust our true positive rate by sliding our probability thresholds down, but this incurs a trade-off in the false positive rate. For example,

at a threshold of 5%, our true positive rate was 91.0%, but our false positive rate was 60.0%, for repression. While correctly classifying 91.0% of defenses is desirable, we would also be misclassifying more than half of non-defenses as defenses. Again, this degree of error is undesirable for full automation.

Accordingly, we considered an alternative application of these models, wherein sections of a transcript that might have defenses are flagged, and coders only review those flagged sections. This way of using the models would assist human coders as opposed to fully replace them. This approach is typical for prediction problems where the goal is to reduce the number of cases that require more detailed investigation (e.g., Alvin et al., 1998; Baghdasaryan et al, 2022). For example, consider that breast cancer detection is a 2-step process wherein (a) patients are screened via mammography (a non-invasive procedure) and (b) if flagged as positive then patients undergo surgical biopsy (a more accurate but highly invasive test; Alvin et al., 1998). The mammography is not perfectly accurate, but requires less work and is better for the patient's quality of life. It is therefore easier to start with a mammography, and to perform a more thorough follow-up check through biopsy if the mammography comes back positive. Analogously, if talk-turns that might contain defenses are highlighted by an algorithm for investigation (meaning all others can be ignored), this reduces the amount of work that human coders must perform. Here, we are shifting our focus from machine automated coding to machine assisted coding.

To this end, we examined the reduction in unnecessary workload (true negative rate) and defenses missed (false negative rate) at various thresholds (also displayed in Table 5). Using a classification threshold of 5%, we can reduce the number of question/answer pairs that have no defenses from the rater's workload by 35.8% to 92.3%, while only missing 3.6% to 34.0% of

defenses from participant talk-turns, depending on which model is used. Looking at the model predicting the presence of any defense as a specific example, at a threshold of 5% we can cut 35.8% of answers that have no defenses from the coder's workload, while only missing 3.6% of answers that actually have defenses.

Discussion

Individuals with psychopathology tend to use lower level/less adaptive defenses compared to non-clinical samples (Carlucci et al., 2021; Perry et al., 2013), which can negatively impact symptomology and treatment progression or outcomes (Høglend & Perry, 1998). As such, it is important to consider defensive functioning when working with clinical samples. However, using an objective, observer rated measure to assess and monitor defensive functioning can be time consuming and labour intensive, which represents a significant barrier to research in this domain. To address this shortcoming, we developed five machine learning models to code defense mechanisms from AAI transcripts in a sample of women with and without BED.

Each model was able to distinguish adequately its respective defense mechanism from the other defenses among participant responses to interviewer questions. This is consistent with other recent studies on the use of machine learning for content coding of psychotherapy sessions (Ewbank et al., 2020; Gaut et al., 2015; Hawa et al., 2020). For example, Ewbank and colleagues (2020) used a deep learning model to accurately classify therapist utterances into 24 categories. These data were then used to study the association between therapist utterances and clinical outcomes in a sample of adults receiving internet-based psychotherapy. Relatedly, Hawa and colleagues (2020) used machine learning to identify signs of depression from therapy transcripts. These studies illustrate the potential utility of machine learning to inform diagnosis and treatment (Hawa et al., 2020), and reduce bias in clinical settings. For example, clinicians could

use machine learning to supplement their own judgements with supporting indicators of a patient's defensive functioning.

Despite these promising results and implications, it may not be feasible to fully automate the DMRS coding process with our current machine learning models. That is, if we were to fully automate coding using these models, we would miss three-fifths to one-fourth of defenses. This degree of error is too high for automated coding. We can adjust our true positive rate by sliding our probability thresholds down, but this incurs a trade-off in the false positive rate. For example, at a threshold of 5%, our true positive rate for repression was 91.0%, but our false positive rate was 60.0%. While correctly classifying 91.0% of defenses is desirable, we would also be misclassifying more than half of non-defenses as defenses. Again, this degree of error is undesirable for full automation.

Instead, our machine learning models may be used to make observer coding more accessible to researchers by substantially reducing the workload required to manually code defense mechanisms. These models could highlight talk-turns that contain a possible defense. The rater would then only review these highlighted sections, rather than the entire transcript. This would reduce the number of question/answer pairs that a human rater would need to read to code defense mechanisms. Looking at repressions for example, we can reduce the rater's workload by 40.0%, while only missing 9.0% of participant talk-turns that include instances of repression. For the other defenses (intellectualization, undoing, and reaction formation), unnecessary workload could be reduced by 75.2% to 92.3%, while only missing approximately one-fifth to one-third of defenses. Under this framework, these models could be used to assist human raters rather than fully replace them.

Machine learning may also be used to expedite both the DMRS training and coding processes. Typically, it takes several months to complete training and over four hours to code a single AAI transcript (Carlucci, 2022). These challenges pose a barrier to the assessment of defensive functioning in both research and practice. Although self-report measures of defensive functioning (e.g., the DMRS-SR-30; Di Giuseppe et al., 2020) may be more expeditious than coding transcripts, there are potential issues of validity in using self-report to assess unconscious processes (Davidson & MacGregor, 1998). Observer rating measures allows psychologists to capture these implicit behaviors or cognitions more accurately, which is important given that defensive functioning may play a role in treatment progression and outcomes for those with mental health problems (Høglend & Perry, 1998).

In essence, machine learning may be a useful tool to supplement human coding of defensive functioning from the DMRS. Furthermore, it may increase accessibility to objective measures of defensive functioning for researchers and clinicians, and may clear time to collect data from larger, more representative samples with a wider range of defense mechanisms. Although these applications are conjectural, the current study underscores the potential utility of machine learning in the field of defensive functioning.

Limitations & future directions

This study was not without its limitations. First, although we used internal cross-validation (i.e., evaluated our machine learning models via a training dataset and a testing dataset—both of which came from the same dataset), there was not enough available data to test out-of-sample external validation; therefore, we cannot generalize our findings to other datasets or samples (Sammut & Webb, 2017).

Second, we only trained and assessed our models on AAI transcripts and not audio recordings. Future studies should incorporate audio to extract more semantic information (e.g., tone, emotion) or meaning from each talk-turn or possible instance of defensive functioning.

Third, we did not fully assess the practical application of each model. In future, researchers should conduct a deployment study to determine if DMRS training or human coding of defense mechanisms can be expedited by applying the pertinent model.

Fourth, our findings should be interpreted with caution, as any biases within the human raters, participants, and/or the training dataset propagate to machine learning models. For example, Hutchinson and colleagues (2020) found that their NLP models harboured harmful biases toward persons with disabilities. As such, any biases inherent to our human raters or study samples may be perpetuated in our machine learning models. Future researchers should include a more diverse set of raters and participants to safeguard against potential biases.

Fifth, we note that we included interviewer questions as part of talk-turns, which was found to boost performance slightly in initial tests. This suggests that a given model's ability to detect a defense has a degree of dependency on the question being asked. Therefore our findings may not generalize well to free-flow conversation or different kinds of interviews. No external validation was conducted, which means our models cannot be generalized beyond the context within which they were tested—and this is further caveated by the fact that the specific question asked by the interviewer may contribute relevant information for prediction.

Sixth, the present project predicted defenses at the talk-turn-level, but in applied research ODF scores are calculated by aggregating at the transcript-level (e.g., Carlucci et al., 2022a, 2022b). Thus, the result that is most critical to applied research would be an algorithm's

performance at the transcript-level. Accordingly, we recommend that future research try making predictions at the transcript-level.

Conclusion

To our knowledge, this study was the first to apply state-of-the-art BERT-based language models to content coding in psychological transcripts. Unlike previous works, this method achieved adequate performance with only the context of one talk-turn pair, inherently lending itself to a wider range of applications. Further, this study demonstrated the potential utility of machine learning to identify instances of defensive functioning from psychological interviews. Although our current models cannot replace and/or fully automate human coding, the models may substantially reduce workload and increase accessibility to observer coding for both researchers and clinicians. This is especially important in the assessment of automatic or unconscious psychological processes (e.g., defensive functioning) that are often inaccessible to participants/patients yet play an important role in their symptomology and treatment progress. Machine learning also has the potential to improve clinical decision making and expedite training in domains that require complex judgements about unconscious processes. Overall, the current study highlighted the potential utility of machine learning in the study of defensive functioning and, more broadly, psychotherapy; however, more work is needed to validate the objective models, improve their accuracy, and eliminate potential bias.

Commented [JW1]: Not sure what this means… could we just remove this?

Declaration of interest statement: There is no conflict of interest.

Funding: The original studies were funded by the XX.

References

Aafjes-van Doorn, K., Kamsteeg, C., Bate, J., & Aafjes, M. (2021). A scoping review of machine learning in psychotherapy research. *Psychotherapy Research, 31*(1), 92-116. https://doi.org/10.1080/10503307.2020.1808729

Alvin, M. I., Kouides, R. W., & Shapiro, D. E. (1998). Estimating the accuracy of screening mammography: A meta-analysis. *American Journal of Preventive Medicine 14*(2), 143-153. https://doi.org/10.1016/S0749-3797(97)00019-6

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). American Psychiatric Association.

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). American Psychiatric Association.

Baghdasaryan, V, Davtyan, H., Sarikyan, A., & Navasardyan, Z. (2022). Improving tax audit efficiency using machine learning: The role of taxpayer's network data in fraud detection. Applied Artificial Intelligence, 36(1). https://doi.org/10.1080/08839514.2021.2012002

Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020). *TweetEval: Unified benchmark and comparative evaluation for tweet classification.* arXiv. https://doi.org/arXiv:2010.12421

Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. *Journal of Epidemiology.* https://doi.org/10.1093/aje/kwz189

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Carlucci, S. (2022). *Defensive functioning in adults with binge-eating disorder*. Doctoral dissertation, University of Ottawa.

Carlucci, S., Chyurlia, L., Presniak, M., Mcquaid, N., Wiebe, S., Hill, R., Wiley, J., Garceau, C., Baldwin, D., Slowikowski, C., Ivanova, I., Grenon, R., Balfour, L., Maxwell, H., & Tasca, G. A. (2021). Assessing defense mechanisms in binge-eating disorder: Preliminary validity and reliability of the Defense Mechanism Rating Scale (DMRS) coded from Adult Attachment Interviews. [Manuscript in preparation]. School of Psychology, University of Ottawa.

Carlucci, S., Chyurlia, L., Presniak, M., Mcquaid, N., Wiebe, S., Hill, R., Wiley, J. C., Garceau, C., Baldwin, D., Slowikowski, C., Ivanova, I., Grenon, R., Balfour, L., & Tasca, G. A. (2022a). Change in defensive functioning following group psychodynamic-interpersonal psychotherapy in women with binge-eating disorder. *International Journal of Group Psychotherapy*, *72*(2), 143-172. https://doi.org/10.1080/00207284.2022.2061980

Carlucci, S., Chyurlia, L., Presniak, M., Mcquaid, N., Wiley, J. C., Wiebe, S., Hill, R., Garceau, C., Baldwin, D., Slowikowski, C., Ivanova, I., Grenon, R., Balfour, L., & Tasca, G. A. (2022b). A group's level of defensive functioning affects individual outcomes in group psychodynamic-interpersonal psychotherapy. *Psychotherapy*, *59*(1), 57–62. https://doi.org/10.1037/pst0000423

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv. https://doi.org/arXiv:1412.3555

Crangle, C. E., Wang, R., Perreau-Guimaraes, M., Nguyen, M. U., Nguyen, D. T., & Suppes, P. (2019). Machine learning for the recognition of emotion in speech of couples in psychotherapy using the Stanford Suppes Brain Lab Psychotherapy Dataset. arXiv. https://doi.org/arXiv:1901.04110

Cummins, R., Ewbank, M. P., Martin, A., Tablan, V., Catarino, A., & Blackwell, A. D. (2019). TIM: A tool for gaining insights into psychotherapy. In *Proceedings of the World Wide Web Conference*. https://doi.org/10.1145/3308558.3314128

Dalal, M. K., & Zaveri, M. A. (2011). Automatic text classification: A technical review. *International Journal of Computer Applications, 28*(2), 37-40. https://doi/org/10.5120/3358-4633

Davidson, K., & MacGregor, M. W. (1998). A critical appraisal of self-report defense mechanism measures. *Journal of Personality, 66*(6), 965-992. https://doi.org/10.1111/1467-6494.00039

Devlin, J., Chang, M. W., Lee, L., Toutanova, K. (2019). *BERT: Pre-training of deep directional transformers for language understanding.* arXiv. https://doi.org/arXiv:1810.04805

Di Giuseppe, M., Perry, J. C., Lucchesi, M., Michelini, M., Vitiello, S., Piantanida, A., Fabiani, M., Maffei, S., & Conversano, C. (2020). Preliminary reliability and validity of the DMRS-SR-30, a novel self-report measure based on the Defense Mechanisms Ratings Scales. *Frontiers in Psychiatry, 11*, 870. https://doi.org/10.3389/fpsyt.2020.00870

Ewbank, M. P., Cummins, R., Tablan, V., Bateup, S., Catarino, A., Martin, A. J., & Blackwell, A. D. (2020). Quantifying the association between psychotherapy content and clinical outcomes using deep learning. *JAMA Psychiatry, 77*(1), 35-43. https://doi.org/10.1001/jamapsychiatry.2019.2664

Ewbank, M. P., Cummins, R., Tablan, V., Catarino, A., Buchholz, S., & Blackwell, A. D. (2021). Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic

coding of session transcripts. *Psychotherapy Research, 31*(3), 300-312. https://doi.org/10.1080/10503307.2020.1788740

Gaut, G., Steyvers, M., Imel, Z. E., Atkins, D. C., & Smyth, P. (2017). Content coding of psychotherapy transcripts using labeled topic models. *IEEE Journal of Biomedical and Health Informatics, 21*(2), 476-487. https://doi.org/10.1109/JBHI.2015.2503985

George, C., Kaplan, N., & Main, M. (1985). *Adult Attachment Interview*. Unpublished manuscript. University of California, Berkeley.

Hawa, S., Akella, S., Kaushik, S., Joshi, V., & Kalbande, D. (2020). Analysis of therapy transcripts using natural language processing. *International Journal of Engineering and Advanced Technology, 9*(6), 489-494. https://doi.org/10.35940/ijeat.F1598.089620

Høglend, C.P., & Perry, J.C. (1998). Defensive functioning predicts improvement in major depressive episodes. *The Journal of Nervous & Mental Disease, 186*, 238-243. https://doi.org/10.1097/00005053-199804000-00006

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). Unintended machine learning biases as social barriers for persons with disabilities. Accessibility and Computing, 125, 1. https://doi.org/10.1145/3386296.3386305

Idalski Carcone, A., Hasan, M., Alexander, G. L., Dong, M., Eggly, S., Brogan Hartlieb, K., Naar, S., MacDonell, K., & Kotov, A. (2019). Developing machine learning models for behavioral coding. *Journal of Pediatric Psychology, 44*(3), 289-299. https://doi.org/10.1093/jpepsy/jsy113

Johnson, C. (1985). Initial consultation for patients with bulimia and anorexia nervosa. In D.M. Garner & P.E. Garfinkel (Eds.), *Handbook of psychotherapy for anorexia nervosa and bulimia*. Guilford Press.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260. https://doi.org/10.1126/science.aaa8415

Kamath, U., Liu, J., & Whitaker, J. (2019). *Deep learning for NLP and speech recognition.* Springer.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Kutner, M. H., & Neter, J. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill Irwin.

Landis, J. R., & Gary, G. K. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. https://doi.org/10.2307/2529310

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyvanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* arXiv. https://doi.org/arXiv:1907.11692

Loshchilov, I., & Hutter, F. (2019). *Decoupled weight decay regularization.* arVix. https://doi.org/1711.05101

MacLaren Chorney, J., McMurtry, C. M., Chambers, C. T., & Bakeman, R. (2015). Developing and modifying behavioral coding schemes in pediatric psychology: A practical guide. *Journal of Pediatric Psychology, 40*(1), 154-164. https://doi.org/10.1093/jpepsy/jsu099

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology, 5*(9), 1315-1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

Maxwell, H., Tasca, G. A., Grenon, R., Faye, M., Ritchie, K., Bissada, H., & Balfour, L. (2017). The role of coherence of mind and reflective functioning in understanding binge-eating

disorder and co-morbid overweight. Attachment and Human Development, 19(4), 407-424. https://doi.org/10.1080/14616734.2017.1318934

Pace, B., Tanana, M., Xiao, B., Dembe, A., Soma, C., Steyvers, M., & Imel, Z. E. (2016). What about the words? Natural language processing in psychotherapy. *Psychotherapy Bulletin, 51*, 14-18.

Park, J., Kotzias, D., Kuo, P., Logan IV, R. L., Merced, K., Singh, S., Tanana, M., Karra Taniskidous, E., Lafata, J. E., Atkins, D. C., Tai-Seale, M., Imel, Z. E., & Smyth, P. (2019). Detecting conversation topics in primary care office visits from transcripts of patient-provider interactions. Journal of the American Medical Informatics Association, 26(12), 1493-1504. https://doi.org/10.1093/jamia/ocz140

Perry, J. C. (1990). *The Defense Mechanism Rating Scales* (5th ed.). Cambridge Hospital.

Perry, J. C., & Henry, M. (2004). Chapter 9: Studying defense mechanisms in psychotherapy using the defense mechanism rating scales. In *Advances in psychology* (Vol. 136, pp. 165-192). Elsevier Science & Technology. https://doi.org/10.1016/S0166-4115(04)80034-7

Perry, J. C., Kardos, M. E., Pagano, C. J. (1993). The study of defenses in psychotherapy using the Defense Mechanism Rating Scales (DMRS). In U. Hentschel, G. J. W. Smith, W. Ehlers, & J. G. Draguns (Eds.), *The concept of defense mechanisms in contemporary psychology*. Springer. https://doi.org/10.1007/978-1-4613-8303-1_8

Perry, J. C., Metzfer, J., & Sigal, J. J. (2015). Defensive functioning among women with breast cancer and matched community controls. *Psychiatry, 78*, 156-169. https://doi.org/10.1080/00332747.2015.1051445

Perry, J. C., Presniak, M., & Olson, T. R. (2013). Defense mechanisms in schizotypal, borderline, antisocial, and narcissistic personality disorders. *Psychiatry, 76*, 35-52. https://doi.org/10.1521/psyc.2013.76.1.32

Raschka, S., & Mirjalili, V. (2017). *Python machine learning*. Pakt.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Sammut, C., & Webb, G. I. (2017). Encyclopedia of machine learning and data mining. Springer. https://doi.org/10.1007/978-0-387- 30164-8

Tasca, G. A., Ritchie, K., Demidenko, N., Balfour, L., Krysanski, V., Weekes, K., Barber, A., Keating, L., & Bissada, H. (2013). Matching women with binge eating disorder to group treatment based on attachment anxiety: Outcomes and moderating effects. Psychotherapy Research, 23(3), 301-314. https://doi.org/10.1080/10503307.2012.717309

Weusthoff, S., Gaut, G., Steyvers, M., Atkins, D. C., Hahlweg, K., Hogan, J., Zimmermann, T., Fischer, M. S., Baucon, D. H., Georgiou, P., Narayanan, S., & Baucom, B. R. (2018). The language of interpersonal interaction: An interdisciplinary approach to assessing and processing vocal and speech data. *The European Journal of Counselling Psychology*, *7*(1), 69-85. https://doi.org/10.5964/ejcop.v7i1.82

Yager, J., Kay, J., & Kelsay, K. (2021). Clinicians' cognitive and affective biases and the practice of psychotherapy. *American Journal of Psychotherapy*, *74*(3), 119-126. https://doi.org/10.1176/appi.psychotherapy.20200025