Can adult mental health be predicted by childhood future-self narratives? Insights from the CLPsych 2018 Shared Task

Kylie Radford and **Louise Lavrencic** and **Ruth Peters** and **Kim M. Kiely**
Neuroscience Research Australia (NEURA)
Barker Street, Randwick
NSW 2031, Australia
initial.lastname@neura.edu.au

Ben Hachey
EdgeDown
Sydney, Australia
ben.hachey@gmail.com

Scott Nowson
EdgeDown IE
Dublin, Ireland
nowson@gmail.com

Will Radford
EdgeDown
Sydney, Australia
wejradford@gmail.com

Abstract

The CLPsych 2018 Shared Task B explores how childhood essays can predict psychological distress throughout the author's life. Our main aim was to build tools to help our psychologists understand the data, propose features and interpret predictions.

We submitted two linear regression models: MODELA uses simple demographic and word-count features, while MODELB uses linguistic, entity, typographic, expert-gazetteer, and readability features. Our models perform best at younger prediction ages, with our best unofficial score at 23 of 0.426 disattenuated Pearson correlation. This task is challenging and although predictive performance is limited, we propose that tight integration of expertise across computational linguistics and clinical psychology is a productive direction.

1 Introduction

Life course epidemiology can provide important insights into the prediction, pathogenesis and prevention of many physical and mental disorders, which can arise from a complex array of risk factors. The CLPsych 2018 Shared Task B used longitudinal data from the British 1958 Birth Cohort and aimed to predict psychological distress across four adult time points, 23, 33, 42 and 50 years using essays written by participants at age 11 years. Gender and a measure of the child's parental social class were the only other details available. As such, the task was focused on using natural language features of these childhood essays to develop a model.

Our goal was to utilise insights from a panel of psychology researchers with particular interest in the psychology of ageing, to determine 'expert' features for use in conjunction with more conventional natural language processing (NLP) approaches. Childhood psychological features which were expected to influence risk of psychological distress in adulthood included intelligence (Koenen et al., 2009; Wraw et al., 2016), adverse childhood experiences (ACEs) (Hammond et al., 2015), and personality (e.g. neuroticism) (Kotov et al., 2010). These factors might be reflected in the linguistic structure and content of childhood essays. Content features were anticipated to be closely linked to time and place (i.e. 1960's Britain), including social norms at that time (e.g. gender roles). Our 'expert' panel each reviewed a different subset of essays to glean specific thematic features that might generalise across essays and relate to psychological distress.

We submitted two linear regression models: our simple model, MODELA used gender, social class and the number of words in the essay as its only features, whereas MODELB added a number of stylistic, syntactic, readability and expert features developed by the panel. In the post-evaluation period, MODELB was best at age 23, with 0.426 disattenuated Pearson correlation, whereas the simpler MODELA was better at 33, 42 and 50 with 0.280, 0.177 and 0.248 respectively.

We present feature analysis to try and characterise *which* factors are important, as well as examine predictive fairness. Finally, we use a hypothetical deployment tightly integrated into clinician workflow to discuss the challenges and opportunities in clinincal deployment of NLP tools.

2 Background

Large-scale longitudinal studies that include linguistic and psychological variables are relatively rare, largely due to the high-complexity of such studies and challenges of participant attrition, long time-scales and significant investment. One key example is the seminal Nun Study which demonstrated an association between the linguistic features (specifically idea density features) of autobiographical essays written in early life and the onset of Alzheimer's disease in late life (Snowdon et al., 1996). Similar findings were later found in a different, less homogeneous sample (Engelman et al., 2010). The idea that linguistic data from early life could predict Alzheimer's disease 60 or more years later has influenced the way we understand this disease, particularly in terms of dementia prevention. Another approach to analysing such early life essays demonstrated a link between number and range of positive emotion words and lifespan (Danner et al., 2001). These studies suggest that it might be possible to predict aspects of late life health and longevity using features of early life texts. Key childhood factors that could influence poorer adult mental health outcomes include lower intelligence (Koenen et al., 2009; Wraw et al., 2016), ACEs (Hammond et al., 2015), and neuroticism (Kotov et al., 2010). It may be possible to detect these factors using linguistic features of early life essays (Snowdon et al., 1996; Danner et al., 2001; Rude et al., 2004; Pennebaker et al., 2003), as well as to identify novel features for predicting psychological distress across the life course.

A diverse set of linguistic features has been used to try and characterise attributes of the author or speaker. These include counts of words with positive and negative emotional valence, grammatical complexity, specific words or word categories, and speech particles. Idea density has itself been extended to incorporate more sophisticated syntactic features such as dependencies (Sirts et al., 2017). Predicting personality traits from text has popularised the LIWC sets of gazetteers (Pennebaker et al., 2015), categorised word lists tailored towards isolating specific aspects of personality.

A key goal was to maximise the benefit of a multi-disciplinary team, and it is important to enable quick insights into the dataset, model and issues around feasibility of deployment. As noted in Kogan et al. (2009), regression is not widely used in NLP, however they had some success with interpretable models with interpretable feature sets. Nguyen et al. (2011) also use regression to predict author age, using $l1$ regularisation to induce a sparser model, selecting a subset of informative features to further analyse. Ethical considerations are also extremely important in clinical NLP (Suster et al., 2017), health research (Benton et al., 2017) and shared tasks in general (Parra Escartín et al., 2017). While (mercifully) much of the data access logistics for the CLPsych18 shared task were handled by the organisers, it is still critical to consider how raw data and interim results are distributed amongst the team, whether models perform unusually poorly for subsets of participants, potential dual use of any developed technology, and suitability of different deployment techniques.

3 Data

Our submission to the shared task focussed on Task B, to predict `pdistress` scores from the Malaise inventory (Rutter et al., 1970) at ages 23, 33, 42 and 50. Scores range from 0-9 on this measure, with a score ≥ 4 indicative of depression. Inputs are author gender, social class and their essay written at age 11, which asked them to imagine their life at age 25. A training set of 9,217 essays was provided, as well as social class and gender. The number of members with recorded scores declines over time to 7,060 at 23, 6,483 at 33, and 6,402 at 42.

3.1 Essay preprocessing

The essays were transcribed to digital form and accordingly have transcription (marked by "*") and anonymisation artefacts (e.g. "[female name]"). The essays vary substantially in topic and grammaticality, with some hardly intelligible. After tokenising and detecting sentence boundaries with spaCy[1], the document sizes range from 48 to 1,640 tokens with a median of 207.

We applied several preprocessing steps. First, we ran the shallow version of the spaCy model as mentioned above to identify tokens and sentence boundaries. We replaced "####" with "£" after examining the context in which it was used. Then, we used the pyenchant[2] spell-correction library to

[1] `https://spacy.io` Package version 2.0.11, Model en-core-web-sm version 2.0.0.

[2] `https://pypi.python.org/pypi/pyenchant` version 2.0.0

correct each token if it was composed of letters, not digits, and not a currency symbol. Applying spelling correction to spaCy output out of context does introduce errors, and we used a combination of hardcoded replacements and exceptions to try and mitigate this, fully-detailed in the Appendix Section A.

The noisily spell-corrected essays were processed a second time with spaCy as tokens were replaced with one or more corrected tokens. In addition to the shallow processing, the model also predicted part-of-speech and named entity tags.

3.2 Expert review

A feature of our team is that we are geographically-distributed, cross-disciplinary and had limited time to work on the submission. Accordingly, we felt it important to maximise the time we spent exploring the data. We built a static website that we could filter and sort the participant records by their demographic variables and `pdistress` outputs, and click through to read their essay. This let us accelerate the reviewing, and the four NeuRA researchers allocated themselves a block of 2,000 participants and used a range of different strategies (e.g. random sampling, `pdistress`-targeted sampling) to quickly read the essays, flag problems and build intuitions for what psychological factors might be useful to model. The researchers spent approximately 10 hours in total and wrote detailed notes, which we used to inform the preprocessing and feature modeling.

4 Model

We used linear regression optimised by stochastic gradient descent (SGD) from scikit-learn[3]. Our pipeline scaled all feature values before applying the SGDRegressor with elastic net regularisation. We optimised hyper-parameters using 10-fold cross-validation over the training using grid search over regularisation `alpha` (0.01, 0.1, 1), penalty balance `l1_ratio` (0.1, 0.15, 0.2; i.e. closer to $l2$ than the sparsity-inducing $l1$) and optimisation iterations `max_iter` (500, 1000, 2000), choosing combinations with the highest disattenuated Pearson correlation, the official metric.[4] The SGD optimisation can be unstable, however we found the fast experiment time critical to iterating quickly over feature ideas.

4.1 Features

Our approach relied on trying to identify groups of theoretically-motivated features to use in the linear regression above.

Demographics We used one variable for gender (male as 0) (`cntrl_gender`) and one-hot encoding for each of the social class variables (`cntrl_a11_social_class=$CATEGORY`).

Document statistics We extracted the number of tokens in the corrected doc (`stat_n_tokens`), the number of unique tokens (`stat_n_types`), the ratio between token and type count (`stat_p_type`), the number of sentences (`stat_n_sentences`) and the mean number of tokens per sentence (`stat_mean_sentence`).

Noise We extracted the proportion of tokens that were mistranscribed using an "*" (`noise_p_asttoks`), the proportion of anonymised tokens with a "[" (`noise_p_left_bracket`), and the proportion of tokens which were replaced during spelling correction (`noise_p_replacement_tokens`).

Shallow syntax We extracted the proportion of tokens labelled with each part-of-speech label (`syn_p_pos-$POS`) and the ratio of nouns to adjectives (`syn_r_ADJ_NOUN`).

Readability We extracted a number of readability metrics from the essays using the `readability` package.[5] These fall into the broad categories of existing grades (`read_grades_$GRADE`), sentence information (`read_sentence_$METRIC`), and syntactic features for word usage (`read_word_$CATEGORY`) and sentence beginnings (`read_beginnings_$CATEGORY`).

Gazetteers We extracted proportions of matches against LIWC gazetteers[6] (`LIWC_p_$CATEGORY`), one-hot features if no terms were found (`LIWC_zero_$CATEGORY`).

[3] http://scikit-learn.org version 0.19.1.

[4] http://clpsych.org/shared-task-2018/384-2

[5] https://pypi.python.org/pypi/readability version 0.2.

[6] http://clpsych.org/shared-task-2018/384-2

Dataset	Label	Age 23		Age 33		Age 42		Age 50	
		disR	#	disR	#	disR	#	disR	#
Test	CLPSYCH18	0.406	-	0.283	-	0.197	-	0.257	-
	MODELA∗	0.396	5/9	0.105	8/9	0.189	6/9	0.209	4/6
	MODELB∗	0.368	8/9	-0.040	9/9	0.210	2/9	0.214	3/6
Test	MODELA	0.401	-	0.280	-	0.190	-	0.248	-
	MODELB	0.426	-	0.279	-	0.177	-	0.202	-

Table 1: Test data results showing disattenuated Pearson correlation and rank. Submissions marked ∗ include the rounding bug, and we show the fixed results in the row below.

Dataset	Label	Age 23		Age 33		Age 42	
		Mean	Std	Mean	Std	Mean	Std
Train	CLPSYCH18	0.326	-	0.227	-	0.196	-
	MODELA	0.376	0.028	0.251	0.032	0.239	0.058
	MODELB	0.401	0.038	0.268	0.033	0.233	0.064

Table 2: Training data results, showing the mean and standard deviation of the disattenuated Pearson correlation over the 10 folds.

We extracted the same features for *expert* gazetteers from the process described above (`EXPERT_p_$CATEGORY`, `EXPERT_zero_$CATEGORY`). The categories included: interpersonal relationships, nature, pets, occupations, positive affect, negative affect, wealth, travel, hobbies, sport, possessions, housing, time, uncertainty, trauma, affection, religiosity, grandiosity, physical appearance and sleep. See Appendix Section B for gazetteers.

Entity We extracted the ratio of named entities found to the number of words (`ents_p`) as well as the ratio of entities to tokens for each type found (`ents_p_$TYPE`).

4.2 Submitted systems

We learned independent models for `pdistress` at ages 23, 33 and 42. Predicting at age 50 is challenging as there was no training data available. We chose a simple heuristic, which was to return our prediction at age 42.

In our official submission, we rounded the `pdistress` predictions to integers, which caused better scores in some models and worse in others. Overall, rounding was detrimental, and we indicate in results below where it was used, otherwise reported results are unrounded.

We submitted MODELA, with both demographic features and `stat_n_tokens`. MODELB used all features.[7]

[7]Source code and notebooks are available at `TODO`.

5 Results

Predicting the `pdistress` outcomes are challenging, and our models tended to work best at younger prediction ages, with simpler models working better than complex at older ages. Table 1 shows the results of MODELA and MODELB at each age. We report the system rank and official metric, disattenuated Pearson correlation, which incorporates measurement error, but is "not suited to statistical hypothesis testing" (Muchinsky, 1996). We compare to an official baseline (CLPSYCH18) that used token unigram features as regression features. As noted above, our official submissions rounded the `pdistress` outputs, which had a negative impact on age 23 and 33 scores. The submitted results were disappointingly all below that of the CLPSYCH18, except for MODELB at age 42, which ranked #2 at 0.210.

After submission, we found and fixed the rounding bug, and re-evaluated our predictions, which we show in the bottom half of Table 1. At age 23, MODELB with access to all features performs better than MODELA and CLPSYCH18. At the older ages, the simpler MODELA increasingly performs better than MODELB, and is competitive with CLPSYCH18, but less so at the older ages (-0.003, -0.007 and -0.009 at ages 33, 42 and 50).

Table 2 shows how the models performed on the training data, using 10-fold cross-validation. In contrast to the test data, both MODELA and MODELB scored consistently higher than than the

23		33		42	
0.385	gender	0.262	gender	0.292	gender
0.091	class=Unskilled	0.095	class=Unskilled	0.089	class=Unskilled
0.072	class=Partly skilled	0.060	class=Skilled manual	0.008	class=Skilled manual
0.032	class=Skilled manual	0.038	class=Partly skilled	-0.013	class=Professional
-0.029	class=Skilled non-manual	-0.067	class=Managerial	-0.040	class=Skilled non-manual
-0.062	class=Professional	-0.077	class=Skilled non-manual	-0.073	class=Managerial
-0.122	class=Managerial	-0.095	class=Professional		

Table 3: Feature weights of MODELA. The best hyperparameters at age 42 had a higher regularisation alpha and a lower `l1_ratio`, leading to a sparser model without the `class=Partly skilled` feature.

CLPSYCH18 model on the training data, with the full set of features in MODELB giving the best results at ages 23 and 33. Performance is higher and more stable at younger ages and scores decline, and inter-fold standard deviation increases with prediction age.

6 Analysis

While linear models may lack complexity and modelling power found in other methods, the are relatively interpretable. We are able to extract the weights for each feature optimised during training and use them to understand the relative importance of different features.

6.1 What did the model learn?

Table 3 shows the feature weights learned in MODELA for the different prediction ages. The gender feature dominates the weight for each of the models and indicates that female gender is strongly associated with higher `pdistress` scores. The higher-skilled-occupation social class variables (i.e. professional, managerial and skilled non-manual) are associated with lower `pdistress` scores at all ages, which may indicate a protective role against psychological distress (in contrast to lower social class groups, especially `UNSKILLED`).

MODELB included many more features as shown in Tables 4, 5 and 6. Gender was still highly weighted at all prediction ages. Stylistically, essays with more sentences, more misspelled words, higher use of determiners (e.g. "the", "a") and fewer unique words (i.e. `stat_n_types` was negatively weighted) were associated with higher scores at age 23, but document statistics were "selected-out" of the age 33 and 42 models. Standard readability metrics like Kincaid et al. (1975) were not highly weighted, perhaps due to the noisy text, but usage of long words was associated with low `pdistress` scores. Few LIWC categories were weighted: *certainty* (e.g. "always", "never") and lack of *affect* matches were associated with high scores at age 23 and 42 respectively, perhaps indicating unmet expectations and dampened emotional expression.

23	
0.394	gender
0.094	read_beginnings_conjunction
0.088	EXPERT_zero_sport
0.075	noise_p_replacement_tokens
0.054	LIWC_p_Certain
0.049	syn_p_pos-CCONJ
0.049	stat_mean_sentence
0.037	EXPERT_p_wealth
0.022	EXPERT_p_interpersonal-second
0.021	EXPERT_p_interpersonal-first
0.019	read_sentence_words_per_sentence
0.017	syn_p_pos-PRON
0.016	EXPERT_zero_occupation-study
0.002	class=Partly skilled
0.001	EXPERT_zero_timeframe
0.001	read_grades_Kincaid
-0.003	class=Skilled non-manual
-0.010	read_grades_Coleman-Liau
-0.010	syn_p_pos-DET
-0.012	read_word_nominalization
-0.031	stat_n_types
-0.033	read_sentence_wordtypes
-0.038	class=Professional
-0.041	EXPERT_p_travel
-0.054	class=Managerial
-0.108	read_sentence_characters_per_word

Table 4: Feature weights of MODELB at age 23

Several expert categories received weight: the lack of discussion about sport was generally associated with higher `pdistress` scores, suggesting that social or physical benefits of sport may be protective. High proportions of tokens discussing wealth may indicate household financial pressures and adverse childhood events at the time of writing, and were associated with high scores. Discussion of sleep was mildly associated with higher scores at ages 33 and 42, and may be related to lower vitality or motivation. Higher incidence of travel terms was associated with lower scores, which may indicate affluence or psycho-

logical openness. Entity features performed well in some models: more mentions of dates and ordinal numbers were associated with lower scores at age 33 and 42, and mentioning people with higher scores at age 33 (similar to `expert` interpersonal features, which were predictive at age 23).

Table 7 shows the results of a feature ablation study detailing how much performance changes when we omit groups of features. It is difficult to ascertain a threshold for statistical significance for the δ values, so these are really only indicative of broad category trends. Gender and social class are overwhelmingly the most important features, with document statistics and noise features providing some benefit, whereas gazetteer features are only sometimes useful. The final row shows an orthogonal experiment where spell-correction was not used and this also degrades performance, underlining the importance of the noise and expert matching feature groups.

6.2 What did we hope would work?

We report here techniques that did not work well during the task. This is likely due to a combination of problems in implementation, hyperparameter selection, and modelling choices. We focussed our effort elsewhere, but these *may* be beneficial given more time.

Support vector regression This technique offered a principled way to generate feature interactions and handle noise and class imbalance. Unfortunately, early results were uninspiring ($\sim$0.150 in training at age 23).

Embedding features We had hoped to use low-dimensional document representations as features (e.g. the value of the d^{th} dimension). We optimised some pre-trained fastText (Bojanowski et al., 2017) embeddings on the training data, and these were selected by the model, but with lower scores ($\sim$0.350 in training at age 23). Embedding features could ultimately be useful, but they are difficult to interpret, and averaging token-wise embeddings may well obscure useful signal.

Longitudinal trajectories The structure of the task suggests that one approach might be to make a sequence of classifications, or a joint or repeated measures one that took `pdistress` at different ages into account. We spent some time analysing the score trajectories, but chose independent regression models for simplicity. Another consideration is that attrition in this longitudinal dataset is likely to be systematically associated with the `pdistress` outcome (Kelly-Irving et al., 2013; Hughes et al., 2017). Cases with missing outcome scores were excluded from our training models but appropriate imputation of missing data may have enhanced our predictions, particularly at older ages.

33	
0.223	gender
0.040	noise_p_replacement_tokens
0.033	syn_p_pos-SPACE
0.027	ents_p_PERSON
0.017	EXPERT_p_sleep
0.007	EXPERT_zero_sport
0.002	class=Skilled manual
-0.002	EXPERT_p_interpersonal-not
-0.007	ents_p_QUANTITY
-0.020	class=Skilled non-manual
-0.021	EXPERT_p_uncertainty
-0.027	ents_p_DATE
-0.036	syn_p_pos-DET
-0.037	EXPERT_p_travel
-0.050	read_sentence_characters_per_word

Table 5: Feature weights of MODELB at age 33

42	
0.254	gender
0.077	LIWC_zero_LIWC_Affect
0.068	EXPERT_zero_sport
0.053	EXPERT_zero_occupation-military
0.031	noise_p_left_bracket
0.025	stat_mean_sentence
0.025	EXPERT_p_sleep
0.023	EXPERT_p_affect-positive
0.023	EXPERT_p_wealth
-0.016	EXPERT_p_sport
-0.021	class=Managerial
-0.023	read_sentence_type_token_ratio
-0.030	class=Skilled non-manual
-0.031	EXPERT_p_travel
-0.032	EXPERT_p_interpersonal-not
-0.061	ents_p_ORDINAL

Table 6: Feature weights of MODELB at age 42

6.3 How fair are the predictions?

Ensuring that no one subset of your population is adversely served by your models is an important consideration when choosing which system to deploy. We joined the test data with the demographic variables to study this question in more detail, by selecting subsets of the population by gender and social class and re-running the evaluations for comparisons. All else being equal, we propose that a better model is one that shows relatively similar performance for different groups.

Table 8 and Table 9 show prediction correla-

Experiment	Age 23	δ	Age 33	δ	Age 42	δ
MODELB	0.401	-	0.268	-	0.233	-
-stat	0.393	-0.008	0.262	-0.006	0.230	-0.003
-noise	0.394	-0.007	0.265	-0.003	0.231	-0.002
-syn	0.404	+0.003	0.269	+0.001	0.229	-0.004
-read	0.399	-0.002	0.270	+0.002	0.232	-0.001
-liwc	0.402	+0.001	0.262	-0.006	0.228	-0.005
-expert	0.395	-0.006	0.271	+0.003	0.228	-0.005
-ents	0.395	-0.006	0.275	+0.007	0.235	+0.002
-cntrl	0.246	-0.155	0.195	-0.073	0.154	-0.079
-spell-correction	0.393	-0.008	0.264	-0.004	0.228	-0.005

Table 7: Ablation analysis over cross-validated training data using attenuated Pearson correlation. The first row shows the performance of MODELB. The middle set of rows show the impact of removing each feature group. The final row shows the impact of not correcting essay spelling.

tions split by gender and social class. For example, when trying to choose between MODELA and MODELB for age 23 according to this definition of fairness, we might prefer the latter as it has more balanced prediction across genders, while the former depends substantially on the gender feature and has uneven performance. However, Table 9's scores are substantially better for the lower social class groups at all ages and models. This was despite these categories having little or no weight as features in MODELB (see Tables 4, 5 & 6) and suggests the text features were more discriminative within low social class compared to high social class groups. Further analysis of essays focused on high social class groups could identify additional linguistic features to improve the fairness, and overall accuracy, of our model.

Age	Model	M	F
23	MODELA	0.021	0.307
	MODELB	0.250	0.231
33	MODELA	0.049	0.177
	MODELB	0.211	0.019
42	MODELA	-0.115	0.053
	MODELB	-0.016	0.049

Table 8: Prediction correlations on gendered subsets of the test data.

7 Discussion

The CLPsych call for papers asks "whether NLP solutions are ready to deploy in the clinical world, and what that deployment could look like." The shared task, especially Task B, is a bold approach to this question. We can imagine less ambitious

Age	Model	LOW	HIGH
23	MODELA	0.435	0.213
	MODELB	0.466	0.234
33	MODELA	0.251	0.228
	MODELB	0.295	0.189
42	MODELA	0.243	0.109
	MODELB	0.213	0.094

Table 9: Prediction correlations on social class subsets of the test data.

ways of approaching the question than predicting an observed variable 12-39 years into the future from a short essay. For instance, using a writing sample at any age to assess distress and offer assistance at that same age seems useful, especially if the assessment could be made using incidental data like school, social or professional writing.

To provide some analysis and discussion, we reframe the original question in these terms. Specifically, we ask: would it be possible to re-allocate resources based on predicted distress in a way that improves future distress?

7.1 Scenario: optimising clinician workflow

We focus on optimising clinician workflow to reduce the incidence of depression at age 23. To do this, we first binarise gold labels with values 4 or higher as True and others as False. The remaining analysis evaluates the model's ability to predict future depression, and the hypothetical impact this could have on optimising clinician workflow.

Figure 6.3 contains a receiver operating characteristic (ROC) curve to illustrate the diagnostic ability of submitted models without rounding at

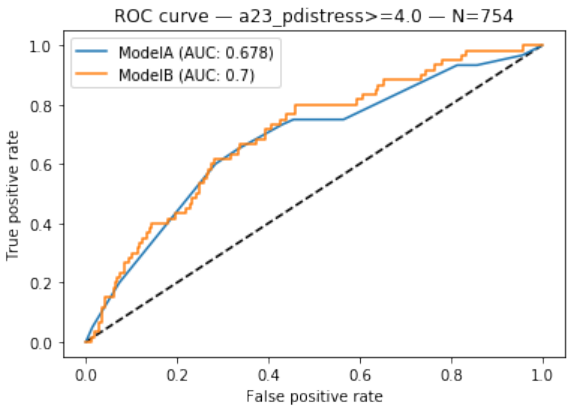

Figure 1: ROC curve for predicting depression at 23.

different thresholds. The true positive rate (TPR) on the y axis is the number of true positive predictions divided by the number of positives in the gold labels. The false positive rate (FPR) on the x axis is the number of false positive predictions divided by the number of negatives in the gold labels. The ROC curve suggests that our MODELA and MODELB are very similar, with the text features in the latter possibly providing an advantage at lower thresholds (towards the right).

We also calculate area under ROC curves across prediction tasks at different ages. These suggest more success predicting distress at lower ages for both MODELA (23: 0.678, 33: 0.622, 42: 0.577, 50: 0.579) and MODELB (23: 0.700, 33: 0.604, 42: 0.598, 50: 0.568). This is perhaps not surprising since it involves less intervening time, and we expect accumulation of life events to become a stronger factor relative to childhood experience over time. Interestingly, the relative performance of models and ages differs from the official disattenuated Pearson correlation score, suggesting it may not be the best for model optimisation or evaluation in a screening scenario.

We return to our scenario to select an operating point on the ROC curve. Imagine we work for an agency with a budget to assess and treat approximately 30% of a population. At a threshold of 1.456, MODELB (with text features) has a 0.617 TPR and a 0.282 FPR. Concretely, at this threshold, we would treat 61.7% of sufferers and we'd also treat 28.2% of non-sufferers. Note that non-sufferers outnumber sufferers 12:1 in our data, so this FPR accounts for most of our budget. At a similar threshold of 1.672, the MODELA (without text features) achieves a slightly lower 0.600 TPR and a similar 0.284 FPR. We use these as operating points for the rest of this discussion.

Let's say our agency is responsible for a population of 1 million individuals. We cannot assess this entire population, let alone treat each individual. We assume that: (1) without treatment, the prevalence of depression ($\texttt{pdistress} \geq 4$) at age 23 is 7.5%; (2) treatment at age 11 can reduce distress at age 23 in all cases; (3) the agency can intervene with 300,000 individuals (30% of the population); (4) we have access to incidental text, gender and socio-economic data at age 11.

Given these assumptions, we compare several scenarios:

- with no intervention, we expect 75,000 individuals to suffer depression at age 23;
- randomly sampling individuals for treatment, we expect 22,500 successful treatments leaving 52,500 sufferers;
- sampling using MODELA, we expect 45,000 successful treatments and 30,000 sufferers;
- sampling using MODELB, we expect 46,275 successful treatments and 28,725 sufferers.

Using MODELA (based on gender and socioeconomic level) reduces incidence of depression by 33% with respect to random sampling. Using MODELB (adding selected text features) reduces incidence by a further 5.7%. This suggests that NLP may indeed be a useful complement to other indicators in a hypothetical workflow optimisation scenario, but most of the predictive power comes from the baseline non-text features.

8 Conclusion

Our shared task submission allowed us to take on this very challenging task. While the prediction accuracy is underwhelming, there are further avenues for exploration. Linguistic features seem to vary across demographics. For example, essays from high social class participants tended to be grammatical and coherent, and spelling error features are not as discriminative as they are in other populations. This suggests that creating compound features that can model patterns that hold within groups could be promising. We hope to see further cross-disciplinary work to find useful ways for psychology to help inform how NLP researchers build tools for humans, and how we can build and deploy practical and useful tools to further support clinicians.

Acknowledgments

This study was approved by the University of New South Wales Human Research Ethics Advisory Panel (ref. HC180171). We thank the CLPsych reviewers for their thoughtful comments. KMK is funded by the Australian National Health and Medical Research Council (NHMRC) fellowship #1088313. KR is supported by the ARC-NHMRC Dementia Research Development Fellowship #1103312. LL is supported by the Serpentine Foundation Postdoctoral Fellowship. RP is supported by the Dementia Collaborative Research Centre.

References

A. Benton, G. Coppersmith, and M. Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain. ACL.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

D. D. Danner, D. A. Snowdon, and W. V. Friesen. 2001. Positive emotions in early life and longevity: findings from the nun study. *J Pers Soc Psychol*, 80(5):804–13.

M. Engelman, E. M. Agree, L. A. Meoni, and M. J. Klag. 2010. Propositional density and cognitive function in later life: findings from the precursors study. *J Gerontol B Psychol Sci Soc Sci*, 65(6):706–11.

K. W. Hammond, A. Y. Ben-Ari, R. J. Laundry, E. J. Boyko, and M. H. Samore. 2015. The feasibility of using large-scale text mining to detect adverse childhood experiences in a va-treated population. *J Trauma Stress*, 28(6):505–14.

K. Hughes, M. A. Bellis, K. A. Hardcastle, D. Sethi, A. Butchart, C. Mikton, L. Jones, and M. P. Dunne. 2017. The effect of multiple adverse childhood experiences on health: a systematic review and meta-analysis. *Lancet Public Health*, 2(8):e356–e366.

M. Kelly-Irving, B. Lepage, D. Dedieu, M. Bartley, D. Blane, P. Grosclaude, T. Lang, and C. Delpierre. 2013. Adverse childhood experiences and premature all-cause mortality. *Eur J Epidemiol*, 28(9):721–34.

J. Kincaid, R. Fishburne, R. Rodgers, and B. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Institute for Simulation and Training, University of Central Florida.

K. C. Koenen, T. E. Moffitt, A. L. Roberts, L. T. Martin, L. Kubzansky, H. Harrington, R. Poulton, and A. Caspi. 2009. Childhood iq and adult mental disorders: a test of the cognitive reserve hypothesis. *Am J Psychiatry*, 166(1):50–7.

S. Kogan, D. Levin, B. R. Routledge, J. S. Sagi, and N. A. Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the NAACL*, pages 272–280, Boulder, Colorado. Association for Computational Linguistics.

R. Kotov, W. Gamez, F. Schmidt, and D. Watson. 2010. Linking "big" personality traits to anxiety, depressive, and substance use disorders: a meta-analysis. *Psychol Bull*, 136(5):768–821.

P. M. Muchinsky. 1996. The correction for attenuation. *Educational and Psychological Measurement*, 56(1):63–75.

D. Nguyen, N. A. Smith, and C. P. Rosé. 2011. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 115–123, Portland, OR, USA. Association for Computational Linguistics.

C. Parra Escartín, W. Reijers, T. Lynn, J. Moorkens, A. Way, and C. Liu. 2017. Ethical considerations in nlp shared tasks. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 66–73, Valencia, Spain. Association for Computational Linguistics.

J. W. Pennebaker, R. Boyd, K. Jordan, and K. Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer. 2003. Psychological aspects of natural language. use: our words, our selves. *Annu Rev Psychol*, 54:547–77.

S. Rude, E. Gortner, and J. W. Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, 18(8):1121–1133.

M. Rutter, J. Tizard, and W. Kingsley. 1970. *Education, health and behaviour*. Longman.

K. Sirts, O. Piguet, and M. Johnson. 2017. Idea density for predicting alzheimer's disease from transcribed speech. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 322–332, Vancouver, Canada. ACL.

D. A. Snowdon, S. J. Kemper, J. A. Mortimer, L. H. Greiner, D. R. Wekstein, and W. R. Markesbery. 1996. Linguistic ability in early life and cognitive function and alzheimer's disease in late life. findings from the nun study. *JAMA*, 275(7):528–32.

S. Suster, S. Tulkens, and W. Daelemans. 2017. A short review of ethical challenges in clinical natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 80–87, Valencia, Spain. ACL.

C. Wraw, I. J. Deary, G. Der, and C. R. Gale. 2016. Intelligence in youth and mental health at age 50. *Intelligence*, 58:69–79.

A Spelling correction

We prevented correction of the following tokens due to interaction with spaCy's tokenisation model: "I", "NT", "nt", "alot", "oclock", "etc", "T.V.", "ve". We hardcoded some common incorrect replacements after manually reviewing the results over 200 essays, as follows.

before/after "n't"/"'nt","Iam"/"I'm", "thay"/"they", "wen"/"when", "wud"/"would", "hav"/"have", "moter"/"motor", "vist"/"visit", "wat"/"what", "haf"/"have", "ther"/"there", "worke"/"work"

B Expert gazetteers

interpersonal-first wife, husband, child/ren, son, daughter, twins, baby, babies, married, marriage, friend/s

interpersonal-second mother, father, grandmother, grandfather, mum, dad, mummy, daddy, ma, pa, granny, grandpa, aunt, uncle, brother/s, sister/s, parent/s

interpersonal-not alone, not married, bachelor, unmarried

natural-world tree/s, bird/s, flowers, garden, outdoors, park, camping, river, sea, ocean, beach, woods, forest, snow, animals

natural-pet dog/s, cat/s, pet/s, horse/s, pony, ponies

occupation-military military, airforce, army, RAF, navy, air force

occupation-vocation hairdresser, hairdressing, typist, nurse, nursing, teacher, teaching, chef, pilot, secretary, office, hotel, factory, job, work, doctor, vet, astronomer, footballer, accountant, bank, archaeologist, geologist, gas works, ambulance driver, shop work, office work, housewife, police, fireman, farmer, farm, computer, housework

occupation-study study, university, training, studying, college, degree

affect-positive like, enjoy, happy, I like, great, good, easy, good life, easy life, happy life, enjoy my life

affect-negative boring, bored, stuffy, sad, upset, lonely, don't have good fun, hard work, don't like, very hard, unhappy, hopeless

wealth rich, wages, wealth, wealthy, pay packet, pay, earn, money, pounds, paid

travel trip, travel, holiday, holidays, break, vacation, plane, caravan, boat, train, travel abroad, overseas, seaside, countryside, country, drive

hobbies reading, music, instrument, collecting, stamp collection, coin collecting, coin collection, reading, model building, art, artist, knitting

sport football, fishing, mountain climbing, horse riding, climbing, riding, horses, skiing, sailing, motorsport, racing, swimming, cycling, hunt, hunting

possessions car, cars, TV, television, new

house bedroom, bedrooms, rooms, house, home, flat, carpet, curtains, walls, wall, chair, furniture, kitchen,table

timeframe monday, tuesday, wednesday, thursday, friday, morning, afternoon, evening, night, lunch, tea, dinner, breakfast, weekend

uncertainty i don't know, might, not sure, unsure, maybe, perhaps

trauma flights, fight, fighting, death, dead, die, died, accident, accidents, hurt, injured, injury, shot, gun, crash, kill, killed, murder, murdered, murderer, bullet, knife

affection helping, help, caring, care, kissing, kiss, love, gentle, careful

religosity church, chapel, christmas, easter, religious, religion, spirituality, jesus, god, christening, pray, praying

grandiose best, perfect, mansion

physical tall, short, large, small, height, weight, hair, face, body, eyes, ears, skin, slim, slender, thin, fat, clothes, dress, shirt, suit, dressed, wear, wearing

sleep sleep, bed, tired, have to get up early, don't like waking early, waking, early, sleepy