CLPsych 2018 Shared Task: Predicting Current and Future Psychological Health from Childhood Essays

Veronica E. Lynn[1], Alissa Goodman[2], Kate Niederhoffer[3], Kate Loveys[4], Philip Resnik[5] and **H. Andrew Schwartz[1]**
[1]Stony Brook University, [2]University College London
[3]Circadia Labs, [4]Qntfy, [5]University of Maryland
{velynn, has}@cs.stonybrook.edu, alissa.goodman@ucl.ac.uk
kate@circadialabs.com, kate@qntfy.com, resnik@umd.edu

Abstract

We describe the shared task for the CLPsych 2018 workshop, which focused on predicting current and future psychological health from an essay authored in childhood. Language-based predictions of a person's current health have the potential to supplement traditional psychological assessment such as questionnaires, improving intake risk measurement and monitoring. Predictions of future psychological health can aid with both early detection and the development of preventative care. Research into the mental health trajectory of people, beginning from their childhood, has thus far been an area of little work within the NLP community. This shared task represents one of the first attempts to evaluate the use of early language to predict future health; this has the potential to support a wide variety of clinical health care tasks, from early assessment of lifetime risk for mental health problems, to optimal timing for targeted interventions aimed at both prevention and treatment.

1 Introduction

The ability to accurately predict current and future psychological health could be transformative in providing more personalized and efficient mental health care. Currently, the mental health care industry is strained and overworked, and many conditions are on the rise among certain populations. For example, suicide rates are climbing among veterans (USDVA, 2016) and youths (CDC, 2017).

Data-driven linguistic analysis offers a particularly attractive complement or alternative to traditional risk assessments, particularly in a clinical setting. Language analysis is often relatively fast and easy to conduct at scale. Further, whereas traditional risk assessments are typically limited to capturing one or a few psychological factors, language analysis has the advantage of being theoretically unlimited in what it can capture. By evaluating the relationship between linguistic markers and lifetime health outcomes, such research may provide benefits for intake assessment, monitoring, and preventative care.

Computational linguistics has now shown strong potential for aiding in mental health assessment and treatment. With few exceptions (e.g. De Choudhury et al. (2016), Sadeque et al. (2016)), work thus far from the NLP community has focused on predicting *current* mental health from language, and most exceptions have still only looked at the short-term future. While such research is valuable, predictions about the long-term future can aid with another class of applications: the understanding of early life markers and development of preventative care.

Here we describe the CLPsych 2018 shared task, the purpose of which is to evaluate multiple methods for analyzing linguistic markers as a signal for current and future psychological outcomes (i.e. risk assessment). We present three tasks centered around this goal: *Task A* focuses on cross-sectional psychological health at age 11, based on essays written at childhood. *Task B* uses these childhood essays to measure psychological distress across multiple life stages. Finally, the *Innovation Challenge* seeks to predict language used forty years in the future.

The data for this work comes from the National Child Development Study (Power and Elliott, 2006), a unique British study which follows a single, nationally-representative cohort of individuals over a sixty-year period starting at birth. The data available to shared task participants includes over ten thousand anonymized childhood essays,

Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pages 37–46
New Orleans, Louisiana, June 5, 2018.

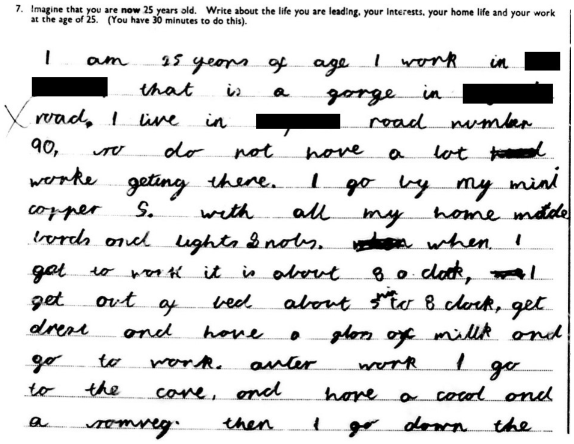
measures of psychological health taken at regular intervals, and adult writing at age 50, all collected as part of the NCDS study.

Related Work. Relatively little work has been done on *future* mental health predictions. De Choudhury et al. (2013) examine depression in individuals by analyzing social media signals up to a year in advance of its reported onset. Similarly, De Choudhury et al. (2016) aims to identify individuals who are likely to engage in suicidal ideation in the future. Sadeque et al. (2016) predict whether posters on a mental health forum will leave the forum within a particular (one, six, or twelve month) time frame. In addition to these cases, some have used temporal information within *cross-sectional* analyses. Zirikly et al. (2016), for example, use timestamp data to help classify the severity levels of posts to a mental health forum. Loveys et al. (2017) explore mental health within the context of *micropatterns*, or sequences of posts occurring within a small time frame. The goal of this shared task is to predict mental health not only at the time of writing, but years or decades into the future.

2 Data Set

This shared task seeks to use childhood language to predict aspects of mental health at ages 11, 23, 33, 42, and 50. The data for this task comes from the National Child Development Study (Power and Elliott, 2006) — also known as the 1958 British Birth Cohort Study — which follows a cohort of all children born in a single week in Great Britain, beginning in March 1958 and continuing until the present day. The study additionally includes a number of children who were born during the target week and who immigrated to Great Britain at or before age 16. This cohort has been followed since their birth and have been surveyed at various points in their life to monitor their progress across a wide range of life domains including their mental health.

Psychological health at age 11. The measure of psychological health at age 11 selected for this task was the Bristol Social Adjustment Guide (BSAG) (Stott, 1963; Ghodsian, 1977), as reported by the participants' teachers. The BSAG includes twelve subscales, plus a total score, that measure different aspects of childhood behavior. For example, teachers were asked if the students displayed characteristics such as "does not know what to do with himself, can never stick at anything long" or "miserable, depressed, seldom smiles". For the purposes of the shared task, we focused on the total BSAG score, as well as anxiety and depression subscales in order to mirror previous CLPsych tasks.


7. Imagine that you are now 25 years old. Write about the life you are leading, your interests, your home life and your work at the age of 25. (You have 30 minutes to do this).

I am 25 years of age I work in [redacted] that is a gorge in [redacted] road, I live in [redacted] road number 90, we do not have a lot worke geting there. I go by my mini copper S. with all my home madde words and lights & nobs. when I get to work it is about 8 o clock, I get out of bed about 5 to 8 clock, get drest and have a glass of millk and go to work. after work I go to the cafe, and have a cool and a romveg. then I go down the

Figure 1: Example of an essay written by an NCDS participant at age 11, imagining where they saw themself at age 25.

Psychological distress across the lifetime. The Malaise Inventory (Rutter et al., 1970) is a measure of psychological distress, used to measure mental health of the cohort participants as adults, at ages 23, 33, 42, and 50. These scores represent the total score on a 9-item scale, where a value at or above 4 is the commonly adopted cutoff indicative of depression. The 9 items are:

> *Do you feel tired most of the time?*
> *Do you often feel depressed?*
> *Do you often get worried about things?*
> *Do you often get into a violent rage?*
> *Do you suddenly become scared for no good reason?*
> *Are you easily upset or irritated?*
> *Are you constantly keyed up and jittery?*
> *Does your heart often race like mad?*
> *Does every little thing get on your nerves and wear you out?*

Essays. At age 11, participants were asked to write a short essay on where they saw themselves in the future according to the following prompt:

> *Imagine you are now 25 years old. Write about the life you are leading, your interests, your home life and your work at the age of 25.*

	# Train	# Test	Mean
Task A (Age 11)			
Anxiety	9146	993	0.47 (1.03)
Depression	9146	993	1.01 (1.47)
Total	9146	992	8.03 (8.49)
Task B (Across Lifespan)			
A23 Distress	7060	754	1.11 (1.47)
A33 Distress	6483	677	0.95 (1.50)
A42 Distress	6402	689	1.49 (1.79)
A50 Distress	—	586	1.42 (1.86)
Innovation Challenge (Age 50)			
Essays	4235	458	—
Total	9217	1000	—

Table 1: Number of training and testing instances available across all outcomes. Age 11 essays were provided for all instances. Mean (along with standard deviation) is based on the test data. Participants were not provided with training data for age 50 distress in order to measure out-of-sample performance.

These essays, in which childhood language captures the author's thoughts towards the future, are the primary focus for predicting lifetime mental health in this shared task. At age 50, participants were given a similar prompt to write about where they saw themselves at age 60; these were included as part of the Innovation Challenge described in Section 3.

Figure 1 shows an example of the age 11 essays. Below is an excerpt from one of the digitally entered age 50 essays.

> *Hopefully I will still be in good health. I will have moved to a smaller property and will have paid off my mortgage. Making my financial position more comfortable. I anticipate I will still be working, probably still full time.*

Controls. Two non-linguistic variables were included as *controls* — variables known to be important for childhood language and also to relate to current and future mental health which, therefore, are desirable to out-predict. These included biological sex and childhood social class, according to the father's occupation (Elliott and Lawrence, 2014). The NCDS data is rich with other childhood variables (such as cognitive exams). However, as we ultimately hope this task motivates more and more development of language-based assessments, we decided not to start with a "high-bar" in terms of controls to out-predict, but rather controls that are almost always available in some form.

Table 1 shows the size of the training and test sets across all outcomes. This dataset was chosen such that all instances contained an age 11 essay with at least 50 words, but one or more mental health outcomes may be missing. The test set was selected randomly and was released to shared task participants approximately one month after the training set with one week to produce predictions.

Privacy Considerations. Every effort had been made in the original study to anonymize the data. However, even de-identified data used for research purposes must obtain human subjects review at one's home institution. In the US, many university ethics boards already specifically list the NCDS data as "exempt" under the revised common rule, but only an institutional review board (IRB) can make the final decision.[1] Within manuscript submissions, all participants were required to affirm that they have had an appropriate review completed at their home organization. Participants were provided with a Template Letter containing information about the dataset in order to make the IRB process smooth for those who had not previously done research involving human subjects review. The Stony Brook University Institutional Review Board found the research analyses conducted by the authors of this manuscript to qualify as exempt.

3 Task Definitions

The shared task consisted of two subtasks, Task A (Cross-Sectional Psychological Health) and Task B (Future Psychological Health), which were designed to target both latitudinal (i.e. at the same time, across individuals) and longitudinal (i.e. assessed in the future) mental health prediction. In addition, teams were given the option to participate in the *Innovation Challenge on Future Psychological Language Generation*. Participants could choose which tasks to submit to.

[1] https://www.hhs.gov/ohrp/regulations-and-policy/regulations/finalized-revisions-common-rule/index.html

3.1 Task A: Cross-Sectional Psychological Health

Task A involves an essay-based psychological assessment of a person's mental health at the time of writing, answering the question of what a person's language says about their current psychological health. For this task, participants were asked to predict the age 11 anxiety, depression, and total BSAG scores. They were provided with the age 11 essays and the socio-demographic controls (gender and social class), as well as the BSAG scores of the training set.

3.2 Task B: Future Psychological Health

Task B addresses the question of how well one predict, based on the childhood essays written at age 11, what a person's psychological health will be at different stages of life. Shared task participants were asked to predict the age 23, 33, 42, and 50 psychological distress scores. As in Task A, they were provided with the age 11 essays and socio-demographic controls. However, although they were given the training set psychological distress scores at ages 23, 33, and 42, the scores at age 50 were intentionally withheld. This was done in order to create an outcome that was out-of-sample across both people and time, roughly simulating a situation where one makes future predictions (i.e. forecasts) when the outcome has not yet happened. Participants were given the option of whether or not to submit age 50 predictions.

3.3 Innovation Challenge: Future Language Generation

One of the limitations of traditional psychological assessments is that they typically only capture one or a few psychological factors. In contrast, language has been shown to capture many aspects of an individual (Pennebaker, 2011; Coppersmith et al., 2014; Schwartz and Ungar, 2015; Kern et al., 2016), making language-based assessments an attractive compliment or alternative. Language generation tools for mental health could indicate whether an individual is likely to produce signs of mental distress in future, e.g. "I want to end my life." Should language generation tools be adequately reliable and valid indicators of future mental health states, these tools could serve as a means of identifying individuals who could be targeted for early intervention or preventative treatments. The Innovation Challenge is a difficult task intended to motivate methods that move the field towards using more open-vocabulary outputs in psychological predictions.

At age 50, the NCDS participants were asked to write a short essay on where they saw themselves ten years in the future — similar to the essays they wrote at age 11. The goal of the Innovation Challenge is to use the age 11 essays to generate the language used in the age 50 essays. Shared task participants were provided with the age 11 essays and controls of the training and testing instances, as well as the age 50 essays from the training set.

4 Evaluation

In this section, we describe the official metrics used for evaluating the shared task. We also present the baseline systems developed by the shared task organizers against which to compare the participants.

4.1 Tasks A and B

For Tasks A and B, the official metric used for ranking submissions was a disattenuated correlation based on the Pearson Product-Moment Correlation Coefficient (Spearman, 1904) between the predicted and actual mental health outcomes. This metric, though isomorphic to a Pearson correlation, accounts for measurement error and therefore produces values with larger variance, making it easier to draw comparisons between system performances. We take the measurement error from literature on the reliability of the adult psychological distress measure ($r_{meas1} = 0.77$; Ploubidis et al. (2017)) and of similar, language-based prediction measures ($r_{meas2} = 0.70$; Park et al. (2015)). The metric is thus:

$$r_{dis} = \frac{r_{Pearson}}{\sqrt{r_{meas1} \cdot r_{meas2}}}$$

In addition to the disattenuated correlation, we also report the Mean Absolute Error (MAE) for all outcomes, as it is common to use methods that optimize error-based metrics. MAE provides another interpretation of accuracy — on average, how far were predictions off from the real predictions (see Table 1 for descriptives; a 9-point scale in the case of Task B).

For Task A, participants were asked to predict the age 11 anxiety, depression, and total BSAG scores. The disattenuated Pearson correlation of the total BSAG score was used for overall system

	Methods Used				Unique Attributes
	NN	*LR*	*SVR*	*E*	
Task A					
Çöltekin et al.		✓			Only word and character n-grams
Guntuku et al.		✓		✓	LDA topics
Liu et al.			✓		Data preprocessing
Simchon & Gilead		✓			Gaussian GLM
TTU		✓			Mixed effects w/ gender, social class intercepts
UGent – IDLab 1	✓	✓	✓	✓	RNN, boosting techniques
UGent – IDLab 2		✓	✓		
UKNLP 1	✓	✓		✓	Ensemble of CNNs + Ridge over n-grams + LIWC
UKNLP 2	✓			✓	Ensemble of CNNs + spectral loss over LIWC
Task B					
Çöltekin et al.		✓			Only word and character n-grams
Guntuku et al.		✓		✓	LDA topics
Liu et al.			✓		Data preprocessing
Radford et al. 1		✓			Spell-corrected words
Radford et al. 2		✓			Syntactic, entity, expert features
Simchon & Gilead		✓			Time series analysis for age 50 predictions
TTU		✓			Mixed effects w/ gender, social class intercepts
UKNLP 1	✓	✓		✓	CNN + N-Grams + LIWC
UKNLP 2	✓			✓	CNN + LIWC

Table 2: Attributes of participant systems for Tasks A and B. Overall, there were eighteen submissions from eight teams. Methods used are Neural Networks (NN), Regularized (i.e. Ridge, Lasso, ElasticNet) Linear Regression (LR), Support Vector Regression (SVR) and Ensemble Techniques (E).

rankings. For Task B, participants predicted the psychological distress scores at ages 23, 33, 42, and, optionally, 50. In order to rank participants, we took the mean of the disattenuated Pearson correlation across the age 23, 33, and 42 predictions.

4.2 Innovation Challenge

To evaluate the Innovation Challenge we compute the BLEU Score (Papineni et al., 2002), a measure commonly used for evaluating machine translation models, between the generated age 50 essay and the actual essay. We then report the average BLEU score across all documents. However, BLEU is not a perfect metric for this task. First, it was intended to be used to compare entire corpora, not individual documents as we do here. Second, this score was designed for machine translation, which our task is not. Instead, we are trying to predict a person's response to an open-ended prompt, based on their response to a similar prompt forty years prior.

For these reasons, we employ a second metric for evaluation based on the semantic similarity between the predicted and actual essays. Here, we represent each age 50 essay using document-level embeddings — computed as the average embedding for all words in the document — and measure the cosine similarity between the generated essay's embedding and that of the actual essay. The word-level embeddings are Word2Vec (Mikolov et al., 2013) embeddings learned from the age 50 essay training set; words that appeared less than ten times were replaced with an out-of-vocabulary token. This approach is similar to that of Garten et al. (2017), which uses embeddings to capture semantic similarity when applying psychological lexica. It's also similar in motivation to metrics like TERp (Snover et al., 2009) and METEOR (Denkowski and Lavie, 2014) which leverage semantic similarity for evaluating language generation. For this metric, we report the average cosine similarity across all essays.

4.3 Baseline Systems

For Tasks A and B, we used a Ridge Regression model trained over unigrams extracted from the age 11 essays to predict each of the psychologi-

	Total		Anxiety		Depression	
	R-Dis	*MAE*	*R-Dis*	*MAE*	*R-Dis*	*MAE*
Baselines						
Gender	0.220	6.428	0.065	0.717	0.152	1.098
Social Class	0.195	6.398	-0.001	0.715	0.163	1.092
Gender+Soc. Class	0.291	6.278	0.011	0.714	0.214	1.086
Ridge-Unigrams	0.493	6.038	0.191	0.704	0.433	1.048
Participant Systems						
Çöltekin et al.	**0.579**	5.615	0.153	0.630	**0.467**	0.968
UGent – IDLab 1	0.567	5.691	0.195	0.476	0.454	1.004
UKNLP 1	0.559	5.695	0.222	0.526	0.433	0.951
UKNLP 2	0.521	5.839	0.092	0.516	0.340	0.944
Simchon & Gilead	0.521	5.677	0.111	0.475	0.390	0.947
UGent – IDLab 2	0.514	5.688	0.176	0.697	0.419	1.019
Liu et al.	0.475	5.803	0.076	0.819	0.361	1.036
TTU	0.461	6.050	0.142	0.704	0.330	1.055
Guntuku et al.	0.443	6.142	**0.235**	0.700	0.362	1.050

Table 3: Results for Task A, measured using both the Disattenuated Pearson R and the Mean Absolute Error. The Total Disattenuated R is the official ranking metric. Bold indicates the best result among participants for each column.

cal health outcomes. Unigrams were restricted to those used by at least 1% of users (roughly 1,000 unigrams) and encoded as both booleans and relative frequencies. The ridge penalty was tuned using cross validation over the entire training set. In addition to the unigrams baseline, we train Ridge Regression models using only the sociodemographic control variables. We produce gender, social class, and gender + social class baselines against which to compare. We encode social class both using a six-point scale and as one-hot features. To produce the age 50 baseline predictions, where no training data was provided, we used the average of the age 23, 33, and 42 predictions.

We used the OpenNMT-py library (Klein et al., 2017) to train a baseline model for the Innovation Challenge. This model, an LSTM Encoder/Decoder, used 2048-dimensional word embeddings and hidden states, but otherwise used the library default settings. 500 instances from the training set were held out for parameter tuning.

5 Participant Approaches and Results

This section summarizes the approaches taken by participants for each of the tasks, as well as the results obtained by each. Participants were allowed to submit up to two times per task. Overall, there were twenty submissions across eight teams.[2]

5.1 Task A

Seven teams participated in Task A, with two teams submitting twice, for a total of nine submissions. An overview of the approaches taken is provided in Table 2. Most teams used some form of regularized linear regression in their models, though using an ensemble of techniques was common. Neural networks were also tried, though typically in conjunction with linear models.

Table 3 shows the results of Task A. Despite the complexity of some of the submitted systems, the top performing team, Çöltekin et al., simply used regularized linear regression with character- and word-level n-gram features. From the participant system descriptions, we believe this was the only system to use character n-grams in addition to word n-grams.

The second place system, UGent – IDLab 1, is an ensemble of many different techniques: ridge regression, SVMs, boosting, and CNNs, RNNs, and feed-forward neural networks. They considered multiple feature types, including TF-IDF, number of spelling mistakes, average word length,

[2]Twenty teams signed up to participate but only 8 teams submitted predictions in the end. Some teams that did not submit cited the tight timeline and being dissatisfied with results as reasons for dropping out.

	Avg. 23-42		Age 23		Age 33		Age 42		Age 50*	
	R-Dis	*MAE*	*R-Dis*	*MAE*	*R-Dis*	*MAE*	*R-Dis*	*MAE*	*R-Dis*	*MAE*
Baselines										
Gender	0.282	1.19	0.366	1.13	0.262	1.10	0.217	1.35	0.236	1.33
Social Class	0.088	1.22	0.168	1.17	0.126	1.10	-0.029	1.39	0.079	1.36
Gender+Soc. Class	0.293	1.18	0.404	1.11	0.284	1.09	0.192	1.35	0.247	1.33
Ridge-Unigrams	0.295	1.20	0.406	1.14	0.283	1.09	0.197	1.37	0.257	1.34
Participant Systems										
Çöltekin et al.	**0.319**	1.09	0.443	1.01	**0.318**	0.99	0.196	1.28	—	—
TTU	0.314	1.18	**0.457**	1.09	0.277	1.09	0.208	1.35	—	—
UKNLP 1	0.306	1.09	0.431	1.01	0.290	0.98	0.198	1.28	0.231	1.30
Guntuku et al.	0.290	1.12	0.387	1.06	0.271	1.01	**0.211**	1.28	0.008	1.42
Simchon & Gilead	0.276	1.08	0.454	0.99	0.246	0.95	0.128	1.31	**0.301**	1.29
Radford et al. 1	0.230	1.17	0.396	1.08	0.105	1.08	0.189	1.34	0.209	1.39
UKNLP 2	0.226	1.15	0.378	1.04	0.188	0.99	0.112	1.42	0.168	1.35
Liu et al.	0.202	1.39	0.227	1.45	0.233	1.18	0.146	1.55	—	—
Radford et al. 2	0.179	1.17	0.368	1.09	-0.040	1.10	0.210	1.33	0.214	1.37

Table 4: Results for Task B, measured using both the Disattenuated Pearson R and the Mean Absolute Error. The official ranking metric is the average Disattenuated R across ages 23, 33, and 42. Bold indicates the best result among participants for each column. *Participants were not required to submit predictions for age 50, for which no training data was provided to simulate a true prospective prediction.

and sentiment. Like this UGent – IDLab team, many of the top systems used ensemble techniques; their strong performance is likely due to using a combination of models that were able to pick up on different signals in the data.

The results for depression generally followed a similar ordering to the total scores, with teams that performed better at predicting the total BSAG scores also doing well at predicting the depression scores. However, this was not the case for anxiety. There, the performance was somewhat random across the teams, with the top performing system for anxiety, Guntuku et al., having the lowest performance for the total scores.

Out of the nine submissions, six systems beat our Ridge-Unigrams baseline for total BSAG, three for anxiety, and two for depression. The socio-demographic control baselines performed significantly worse than the language-based systems.

5.2 Task B

Task B received nine submissions from seven teams. An overview of the participant systems is shown in Table 2 and the results are in Table 4.

The top performing system was submitted by Çöltekin et al. As with Task A, they trained a linear regression model with L2 regularization over character and word n-gram features. Their system obtained the highest average disattenuated R for ages 23, 33, and 42, as well as the highest *R-Dis* for age 33 itself. Çöltekin et al. indicated that this model was actually intended to be their 'baseline system', but they found it to out-predict more sophisticated models such as Poisson regression and deep networks. This is also supported by the overall results in that submissions indicating use of neural nets (CNNs, RNNs, or FFNNs) came in lower positions but still mostly within the upper-half of rankings.

TTU had the highest *R-Dis* for age 23, as well as the second-best performing system overall. They used a linear mixed-effects regression model with intercepts based on the gender and social class controls. Their features included a number of lexica including LIWC (Pennebaker et al., 2015), the Moral Foundations Dictionary (Graham et al., 2009), and LDA-derived terms. Of significance, this was the only system that did not simply treat the controls as additional features. Instead, by using intercepts based on the controls, their model focused on using the essays to predict what was not accounted for by the controls.

Overall, our baselines were very strong,

	BLEU	W2V Sim.
Baselines		
LSTM	0.413	0.759
Participant Systems		
Liu et al. 1	**0.246**	**0.866**
Liu et al. 2	0.114	0.804

Table 5: Results for the Innovation Challenge, measured using the BLEU score and the cosine similarity between document word embeddings. Bold indicates the best result among participants for each column.

with Gender, Gender+Social Class, and Ridge-Unigrams all performing competitively with the participant systems. Surprisingly, the Gender baseline produced the single best result across all systems at age 42.

The age 50 predictions were challenging, as they were out-of-sample across both *time* and *people*. A common technique was to simply reuse the age 42 predictions or, in the case of our baseline model, to take the average across the age 23, 33, and 42 predictions. In contrast, Simchon & Gilead used time series analysis to produce the age 50 predictions, which ultimately ended up significantly outperforming the other systems. As one might expect, the performance of all systems generally worsened the farther in the future they were asked to predict. However, the strong performance of Simchon & Gilead's approach suggests that this task is still doable.

5.3 Innovation Challenge

The results for the Innovation Challenge are shown in Table 5. There were two submissions, both by Liu et al. This was a very difficult task, both due to the very small training set size (by deep learning standards) and the difficulty of predicting the answer to an open-ended question forty years in the future.

The top submission, Liu et al. 1, generates the age 50 essays using an RNN. The generated essays are coherent, using full sentences and reasonable grammar. However, these outputs suffer from a common problem with deep learning approaches to language generation: the model has simply memorized the training set, rather than learning to produce novel text. A comparison between the generated essays from Liu et al. 1 and the training set shows that 99.6% of trigrams appearing in the generated essays also appear in the training set. In addition, 31.9% of the generated essays appear in their entirety in the training set.

The second submission, Liu et al. 2, uses both RNNs and LSTMs for generation. It's not surprising that this more complicated model would perform worse than the simpler Liu et al. 1, given that the overall training set size is quite small. Unlike the previous submission, the generated essays from this model are often nonsensical, with outputs such as:

> *still working in the same as i am still working and enjoying my children and enjoy my children and enjoy my children and enjoy my children...*

This repetition of words or phrases is another common problem in language generation, often stemming from a lack of training data.

Despite obtaining a decent BLEU score, our baseline system suffers from a similar repetition problem. The limitations of BLEU, as outlined in Section 4.2, are evidenced by the inflated score for the baseline system. The Embedding Similarity score more reasonably reflects the quality of the generated essays, based on our own observations.

Instead of attempting to generate the age 50 essays themselves, an alternative would be to predict the relative frequency of words deemed psychologically relevant according to literature (e.g. singular versus plural pronouns; 'excited', 'hate', 'friends'). This problem is likely simpler, as it can be approached using regression instead of generation, but would still capture meaningful aspects of language for further analysis. We also suggest future systems consider pretraining or creating embeddings using deep learning over a larger data set of childhood writing and then fitting such models to this specific data.

5.4 Discussion

Considering the results in relation to the clinical use of childhood essays to assess mental health, several points are of significance. First, we saw a gradual trend of psychological outcomes becoming more difficult to predict, with age 11 BSAG scores being easiest (though a different type of outcome) and age 42 psychological distress being the hardest. This suggests that, as one might expect, the difficulty of a mental health prediction increases as its temporal distance from the observed

language increases. Still, age 50 psychological distress was predicted better than 42.

Predominantly, essay-based predictions were more accurate than those from gender and social class alone. Thus, such assessments seem at least valuable in situations where mental health assessments are not easily available. They also suggest promise in situations where thorough mental health assessment is already available, but it is not clear if there is an incremental advantage at this point. For example, a Ridge Regression model trained on age 11 anxiety, depression, and total BSAG scores, along with the gender and social class controls, obtained an average *R-Dis* of .348 for predicting psychological distress at ages 23, 33, and 42, which slightly outperformed participating systems that were based only on the essays, gender, and social class. This result provides a good target for future researchers to work towards.

Based on the current results, essay-based assessments may be most valuable where administering detailed assessments is particularly costly or burdensome (relative to the cost or burden of collecting open text), or where a wider set of non-theory driven information is likely to be especially valuable. In the end, we see this consistent result, across all teams using a variety of approaches, as evidence for the strength of language-based assessments for current and future mental health.

We suggest next steps toward clinical use include: (1) continued improvement of model predictive accuracy, (2) further evaluation of the statistical and psychometric properties of such assessments in comparison to existing standards, and (3) careful trial deployment of language-based assessments in clinical practices — only seen by trained and experienced mental health professionals who would evaluate their utility and ultimately guide us toward a randomized controlled trial of language-based assessments within clinical treatment regimens.

6 Conclusion

The CLPsych 2018 shared task sought to examine the power of childhood essays as a predictor of lifetime mental health. Task A took a cross-sectional approach, using essays written at age 11 to predict mental well-being outcomes from the same age. Looking towards the long term, Task B used the age 11 essays to estimate psychological distress across multiple life stages. The Innovation Challenge, which tasked participants with generating language forty years in the future, was intended to motivate a more open-vocabulary approach to psychological health predictions. The unique data for this task, following a nationally representative cohort of over 10,000 children over their lifetimes, is made available via the UK Data Service for further research use,[3] thus providing a resource for making further advances towards effective clinical use of computational linguistics.

Acknowledgments

We thank the UK ESRC for their support of essay transcriptions and initial analyses. We also thank Deven Shah, Anvesh Myla, Youngseo Son, Kiranmayi Kasarapu, Keshav Gupta, Neelaabh Gupta, and Deepak Gupta for their assistance with establishing baselines.

References

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. pages 51–60.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the AAAI Conference on Weblogs and Social Media*.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, pages 2098–2110.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. pages 376–380.

Jane Elliott and Jon Lawrence. 2014. *Refining Childhood Social Class Measures in the 1958 British Birth Cohort Study*. Centre for Longitudinal Studies.

Centers for Disease Control and Prevention. 2017. Quickstats: Suicide rates for teens aged 15–19 years, by sex — United States, 1975–2015. *Morbidity and Mortality Weekly Report* .

[3] www.ukdataservice.ac.uk, www.cls.ioe.ac.uk/page.aspx?&sitesectionid=724

Justin Garten, Joe Hoover, Kate M. Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Dehghani. 2017. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior Research Methods* .

M. Ghodsian. 1977. Children's behaviour and the bsag: Some theoretical and statistical considerations. *British Journal of Clinical Psychology* 16(1):23–28.

Jesse Graham, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology* 96(5):1029.

Margaret L. Kern, Gregory Park, Johannes C. Eichstaedt, H. Andrew Schwartz, Maarten Sap, Laura K. Smith, and Lyle H. Ungar. 2016. Gaining insights from social media language: Methodologies and challenges. *Psychological Methods* 21(4):507.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL*.

Kate Loveys, Patrick Crutchley, Emily Wyatt, and Glen Coppersmith. 2017. Small but mighty: Affective micropatterns for quantifying mental health from social media language. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. pages 85–95.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*.

U.S. Department of Veterans Affairs. Office of Mental Health and Suicide Prevention. 2016. Suicide among veterans and other Americans 2001–2014 .

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics*.

Gregory Park, H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Michal Kosinski, David J. Stillwell, Lyle H. Ungar, and Martin E. P. Seligman. 2015. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology* 108(6):934.

J. W. Pennebaker, R. J. Booth, R. L. Boyd, and M. E. Francis. 2015. Linguistic inquiry and word count: LIWC2015 .

James Pennebaker. 2011. *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury Press.

G. B. Ploubidis, A. Sullivan, M. Brown, and A. Goodman. 2017. Psychological distress in mid-life: Evidence from the 1958 and 1970 British birth cohorts. *Psychological Medicine* 47(2):291–303.

C. Power and J. Elliott. 2006. Cohort profile: 1958 British birth cohort (National Child Development Study). *International Journal of Epidemiology* 35(1):34–41.

Michael Rutter, Philip Graham, and William Yule. 1970. *A Neuropsychiatric Study in Childhood*. Clinics in Developmental Medicine. Heinemann Medical Books.

Farig Sadeque, Ted Pedersen, Thamar Solorio, Prasha Shrestha, Nicolas Rey-Villamizar, and Steven Bethard. 2016. Why do they leave: Modeling participation in online depression forums. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*. pages 14–19.

H. Andrew Schwartz and Lyle H. Ungar. 2015. Data-driven content analysis of social media: a systematic overview of automated methods. *The ANNALS of the American Academy of Political and Social Science* 659(1):78–94.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER?: Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*. pages 259–268.

Charles Spearman. 1904. The proof and measurement of association between two things. *American Journal of Psychology* 15(1):72–101.

D. H. Stott. 1963. *The Social Adjustment of Children: Manual of the Bristol Social Adjustment Guides*. University of London Press.

Ayah Zirikly, Varun Kumar, and Philip Resnik. 2016. The GW/UMD CLPsych 2016 shared task system. In *Proceedings of the Workshop on Computational Lingusitics and Clinical Psychology*. pages 166–170.