A Psychologically Informed Approach to CLPsych Shared Task 2018

Almog Simchon and **Michael Gilead**
Department of Psychology
Ben-Gurion University of the Negev, Israel
almogsi@post.bgu.ac.il mgilead@bgu.ac.il

Abstract

This paper describes our approach to the CLPsych 2018 Shared Task, in which we attempted to predict cross-sectional psychological health at age 11 and future psychological distress based on childhood essays. We attempted several modeling approaches and observed best cross-validated prediction accuracy with relatively simple models based on psychological theory. The models provided reasonable predictions in most outcomes. Notably, our model was especially successful in predicting out-of-sample psychological distress (across people and across time) at age 50.

1 Introduction

In recent years, technological advances have made it possible to extract psychological features from textual input in an automated manner (Boyd and Pennebaker, 2015; Pennebaker et al., 2003; Schwartz and Ungar, 2015).

In a recent review, Guntuku et al. (2017b) show promising evidence that depression and mental illness can be predicted from text provided in online environments at an encouraging range of moderate to high accuracy. Attempts for predicting other psychopathologies such as ADHD (Guntuku et al., 2017a), schizophrenia Mitchell et al. (2015) and suicidal tendencies (Robinson et al., 2016; Won et al., 2013) have also shown promise.

In the spirit of these cutting-edge developments, the Computational Linguistics and Clinical Psychology Workshop (CLPysch) have brought together linguists, psychologists and computer scientists to form a place for a multidisciplinary research, utilizing computational linguistics to the study of mental health. In former years, CLPysch launched a Shared Task, bringing together groups of researchers to tackle a single problem expressed in one dataset. Past events included depression and PTSD detection (Coppersmith et al., 2015) and crisis classification from online message boards (Milne et al., 2016; Milne, 2017). This year, the shared task focused on longitudinal data taken from the National Child Development Study (NCDS; UCL, 2018). Participating teams in the shared task were provided with essays of 11-year-old participants alongside with their corresponding gender and Socio-Economic Status (SES) and were requested to predict: (a) cross-sectional psychological health at age 11 measured by the total score in the Bristol Social Adjustment Guides (BSAG; Stott, 1963) and two sub-measures of depression and anxiety; (b) Future psychological health at ages 23, 33, 42 and 50 as measured by the participants' score of psychological distress in the Malaise Inventory (Rutter et al., 1970).

2 Methods

This study has undergone ethics review by the BGU Department of Psychology Ethics Committee and has been deemed approved.

Participants in the Shared Task were given a training set consisted of 9,217 observations, with some missing data (Table 1).

Task A			Task B			
total	depression	anxiety	age 23	age 33	age 42	age 50
9,146	9,146	9,146	7,060	6,483	6,402	Not provided

Table 1: Final number of observations in the training set for each dependent variable.

2.1 Features

Spelling Errors: Since the input text belonged to 11-year-olds, data cleansing was the first step. We used `spelling` (Ooms and Hester, 2017) library for R to detect spelling errors, and replaced all error with the first suggested correction by the `hunspell` library (Ooms, 2017). We counted spelling errors and computed spelling-error ratio

Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pages 113–118
New Orleans, Louisiana, June 5, 2018.

as a feature. All other features were based on the corrected text. The intuition behind using this as a measure of psychological well-being stems from a hypothesized relation between impulsivity/ADHD (Seymour et al., 2012), scholastic success (Desocio and Hootman, 2004), and psychological outcomes. Apart from the high comorbidity between ADHD and anxiety and mood disorders (Kessler et al., 2006), ADHD is associated with antisocial behaviors (Storeb and Simonsen, 2016), which is embedded in different subscores of the BSAG measure.

Physical vs. Intellectual Interests: based on a lay psychological theory according to which interest in physical rather than intellectual activity could reflect tendencies towards attention/hyperactivity, we included a measure of interest in sports and academia, by compiling dictionaries of sports and english premier league clubs, and University related words (i.e. `Oxford`, `Cambridge`, `University`). These were added using `LIWCalike` (Benoit, 2018).

Handwriting Comprehensibility: The original text file contained asterisks for marking misunderstandings by the text typist. The comprehensibility measure was defined as the sum of asterisks in the original text. Again, the idea being that individuals with disorganized handwriting are more likely to suffer from ADHD and lower scholastic success.

Affect Norms: We calculated mean value of the valence, arousal and dominance of the text using `ANEW` (Bradley and Lang, 1999). The three features correspond to the three-dimensional view of emotion (Russell and Mehrabian, 1977). The psychological intuition is that individuals who are prone to negative affect and high arousal will use language that reflects these characteristics.

Passive Voice: We extracted passive voice by calculating the percentage of passive auxiliary verbs in the text using `spaCy NLP` (Honnibal and Johnson, 2015) and its wrapper for R (Benoit and Matsuo, 2018). The theoretical impetus behind including this feature is work showing a relation between lack of sense of control and depression (Lachman and Weaver, 1998), and work within our lab showing the relation between passive voice and lack of sense of control (Simchon and Gilead, in preparation).

LIWC: The Linguistic Inquiry and Word Count (`LIWC`; Pennebaker et al., 2015) is a dictionary-based program for text analysis. LIWC holds dozens of dictionaries tapping into psychological and linguistic features. The program provides the word-use of each dictionary as output. These dictionaries supposedly provide a good coverage of themes that are important in individuals psychological makeup (e.g., family, motivation, affect, and so on).

Absolutist Words: In light of prior research showing that the use of absolutist words are related to mental health outcomes (Al-Mosaiwi and Johnstone, 2017).

Text Concreteness: Brysbaert et al. (2014) compiled a list of 40k English lemmas rated on a bipolar scale from abstract to concrete. We extracted the average concreteness ratings of the text. The motivation for extracting this feature lies in the idea that language abstractness often relates to cognitive performance (Fyfe et al., 2015; Vellutino and Scanlon, 1985), which is associated with mental health outcomes (Roca et al., 2015).

Unusualness of the Text: For each individual, we calculated sum of squared deviations from the average of each `LIWC` dimension across the entire sample, as a proxy for overall unusualness of the text. This was motivated by the lay psychological theory that individuals who are non-normative would also suffer from negative psychological implications due to such factors as social exclusion.

Unique Words: Number of unique words in the text. The idea is that linguistic richness may reflect high intellect, which is believed to be a resilience factor for mental health (Block and Kremen, 1996).

BSAG-Predictive Words: Scores of general distress, anxiety and depression related words were based on splitting the training set by the corresponding BSAG score into low and high subgroups, extracting the frequent words used by the two splits, subtracting the relative words use of the two parts and normalizing the score. For example, the score of the word husband is 25.95, which means it is positively associated with low score BSAG total, while the score of football is -13.08, which is positively associated with high BSAG total.

In addition to these features, gender, SES and number of unigrams in the text were provided and used in the model as well.

3 Results and Discussion

3.1 Task A

In task A, the goal was to predict the teachers evaluation of the Bristol Social Adjustment Guides (BSAG) at age eleven, based on the child's text. We attempted several different models (e.g., SV regression; random forest), and saw, perhaps surprisingly, that the linear model produced the best cross-validated accuracies. Moreover, given our background in theoretical psychology, we favored the added benefit of the interpretability of such a model.

We fitted a linear regression model comprised of the above mentioned features without interactions to predict the square root of the BSAG total, BSAG anxiety and BSAG depression. For the purpose of model estimation, we conducted a 10-fold cross-validation. The predicted values were converted back to the original scale and presented in Table 2 alongside with the true results. The main metric is Disattenuated Pearson correlation coefficient between the predicted results and the observed results, divided by the reliability of psychological distress questionnaires (0.77; Ploubidis et al., 2017) and of a recent assessment of related language-based measures (0.70; Park et al., 2015). In this metric, higher values represent better predictions.

$$r_{Disattenuated} = \frac{r_{Pearson}}{\sqrt{0.7 \cdot 0.77}}$$

Mean Absolute Error (MAE), which is the average of the absolute error term between the predicted and observed values is also reported. In this metric, lower values represent better predictions.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

	10-fold CV			Official Test Results		
	total	anxiety	depression	total	anxiety	depression
r_d	0.49	0.22	0.37	0.52	0.11	0.39
MAE	5.83	0.59	0.96	5.67	0.47	0.94

Table 2: 10-fold cross-validation and official test results of task A.

3.2 Task B

In this task, the goal was to predict psychological distress scores at ages 23, 33, 42 and 50 based on the Malaise Inventory (Rutter et al., 1970). Age 50 predictions were particularly challenging since not only were they out-of-sample across people, they were also across time (i.e. age 50 distress was never part of the training sample). To tackle this problem, we built a multivariate linear model that included the same features as in Task A. The model produced predictions for ages 23, 33 and 42. On these predicted values, we built a time series for each subject, comprised of the three predicted time points. We used `forecast` library for R (Hyndman, 2017; Hyndman and Khandakar, 2008) to predict the 4th value in the series which corresponds to age 50, using an automatic exponential smoothing. Results are shown in Table 3. Like in Task A, the main metric is Disattenuated Pearson correlation coefficient. Mean Absolute Error (MAE) is also reported.

		M 23-42	age 23	age 33	age 42	age 50
10-fold CV	r_d	0.26	0.37	0.23	0.19	NA
	MAE	1.18	1.17	1.03	1.33	NA
Official Test Results	r_d	0.27	0.45	0.25	0.13	0.30
	MAE	1.084	0.99	0.95	1.31	1.29

Table 3: 10-fold cross-validation and official test results of task B.

In This task, the main evaluation was based on average prediction of ages 23-42. The model provided reasonable predictions in general, but in age 50 predictions it produced the highest result out of all other competing CLPsych 2018 participants. As described, our models favored a simple approach building upon relatively straightforward linear models and psychologically-informed feature selection. This may provide some evidence in favor of simple models when out-of-sample across-people and across-time predictions are needed.

One of the benefits of using classic methods such as linear regression, is model interpretability. In Tables 4 and 5 we list the relevant features used in the our models that passed a significance threshold of $p < .05$ in the training and test sets.

4 Conclusions

We approached the Shared Task by building simple models comprised of various psychology-informed features. Although our models were not the most successful in the shared task, they did show some successful predictions on some of the outcome measures. Specifically, in predicting out-of-sample across-people and across-time, our model produced the best result out of CLPsych

total		anxiety		depression	
training	*test*	*training*	*test*	*training*	*test*
cntrl_gender	cntrl_gender	**arousal**	function	a11_total1grams	Sixltr
a11_total1grams	Sixltr	Sixltr	quant	**arousal**	affect
arousal	discrep	Dic	negemo	WC	posemo
WC	**spelling**	AllPunc	anx	Clout	health
Clout	**unique_words**	**pred_total**	sad	social	**spelling**
Sixltr	**pred_dep**	**pred_anx**	focusfuture	family	**unique_words**
Dic		**pred_dep**	swear	female	**pred_dep**
social			**spelling**	insight	
family				swear	
female				**spelling**	
insight				**misund**	
differ				**unique_words**	
swear				**pred_anx**	
spelling				SES	
spelling_ratio					
misund					
unique_words					
pred_dep					
SES					

Table 4: Significant features in Task A. Features in **bold** were not incorporated in LIWC or in the original dataset.

age 23		age 33		age 42	
training	*test*	*training*	*test*	*training*	*test*
cntrl_gender	cntrl_gender	cntrl_gender		cntrl_gender	tentat
a11_total1grams	Analytic	a11_total1grams		WC	relativ
WC	social	WC		affect	motion
friend	motion	**absu**		posemo	space
study	space	**pred_dep**		negemo	time
SES	**passive_aux**	SES		power	
				unique_words	

Table 5: Significant features in Task B. Features in **bold** were not incorporated in LIWC or in the original dataset.

2018 participating teams. That said, there is still much room for model improvements and feature extraction. Despite the performance advantages afforded by novel statistical approaches (e.g., neural networks, support vector regression, random forest regression and so forth), the linear models may still have some practical use in prediction problems, given their low complexity and variance. Furthermore, they produce the benefit of higher interpretability, which can facilitate gradual accumulation of knowledge regarding relevant features. Our findings also suggest that some potentially unexpected features (e.g., spelling mistakes, incomprehensibility of written text) can be derived from psychological theory, and augment prediction of meaningful outcomes.

Acknowledgments

The authors wish to thank Dr. Jonathan Rosenblatt and the CLPsych 2018 Shared Task organizers. We are also grateful to The Centre for Longitudinal Studies, UCL Institute of Education for the use of these data and to the UK Data Archive and UK Data Service for making them available. However, they bear no responsibility for the analysis or interpretation of these data.

References

Mohammed Al-Mosaiwi and Tom Johnstone. 2017. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, page 2167702617747074.

Kenneth Benoit. 2018. *LIWCalike: Text analysis similar to the Linguistic Inquiry and Word Count (LIWC)*. R package version 0.3.2.

Kenneth Benoit and Akitaka Matsuo. 2018. *spacyr: Wrapper to the 'spaCy' 'NLP' Library*. R package version 0.9.6.

Jack Block and Adam M. Kremen. 1996. IQ and ego-resiliency: Conceptual and empirical connections and separateness. *Journal of Personality and Social Psychology*, 70(2):349–361.

Ryan L. Boyd and James W. Pennebaker. 2015. A way with words: Using language for psychological science in the modern era. *Consumer Psychology in a Social Media World*, (October):222–236.

Margaret M Bradley and Peter J Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Citeseer.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39.

Janiece Desocio and Janis Hootman. 2004. Children ' s Mental Health and School Success. *The Journal of School Nursing*, 20(4):189–196.

Emily R Fyfe, Nicole M McNeil, and Bethany Rittle-Johnson. 2015. Easy as abcabc: Abstract language facilitates performance on a concrete patterning task. *Child development*, 86(3):927–935.

Sharath Chandra Guntuku, J. Russell Ramsay, Raina M. Merchant, and Lyle H. Ungar. 2017a. Language of ADHD in Adults on Social Media. *Journal of Attention Disorders*.

Sharath Chandra Guntuku, David B. Yaden, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt. 2017b. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378.

Rob J Hyndman. 2017. *forecast: Forecasting functions for time series and linear models*. R package version 8.2.

Rob J Hyndman and Yeasmin Khandakar. 2008. Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 26(3):1–22.

Ronald C Kessler, Lenard Adler, Russell Barkley, Joseph Biederman, C Keith Conners, Olga Demler, Stephen V Faraone, Laurence L Greenhill, Mary J Howes, Kristina Secnik, et al. 2006. The prevalence and correlates of adult adhd in the united states: results from the national comorbidity survey replication. *American Journal of psychiatry*, 163(4):716–723.

Margie E Lachman and Suzanne L Weaver. 1998. The sense of control as a moderator of social class differences in health and well-being. *Journal of personality and social psychology*, 74(3):763.

David N Milne. 2017. Triaging content in online peer-support: an overview of the 2017 CLPsych shared task.

David N Milne, Glen Pink, Ben Hachey, and Rafael A Calvo. 2016. CLPsych 2016 Shared Task: Triaging content in online peer-support forums. *3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–127.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the Language of Schizophrenia in Social Media. *Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies (NAACL HLT 2015)*, pages 11–20.

Jeroen Ooms. 2017. *hunspell: High-Performance Stemmer, Tokenizer, and Spell Checker*. R package version 2.9.

Jeroen Ooms and Jim Hester. 2017. *spelling: Tools for Spell Checking in R*. R package version 1.1.

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

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Review Literature And Arts Of The Americas*.

GB 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.

Jo Robinson, Georgina Cox, Eleanor Bailey, Sarah Hetrick, Maria Rodrigues, Steve Fisher, and Helen Herrman. 2016. Social media and suicide prevention: a systematic review. *Early intervention in psychiatry*, 10(2):103–121.

Miquel Roca, Saray Monzón, Margalida Vives, Emilio López-Navarro, Mauro Garcia-Toro, Caterina Vicens, Javier Garcia-Campayo, John Harrison, and Margalida Gili. 2015. Cognitive function after clinical remission in patients with melancholic and non-melancholic depression: a 6 month follow-up study. *Journal of affective disorders*, 171:85–92.

James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294.

Michael Rutter, Jack Tizard, and Kingsley Whitmore. 1970. *Education, health and behaviour*. Longman Publishing Group.

H. Andrew Schwartz and Lyle H. Ungar. 2015. Data-Driven Content Analysis of Social Media: A Systematic Overview of Automated Methods. *Annals of the American Academy of Political and Social Science*, 659(1):78–94.

Karen E. Seymour, Andrea Chronis-Tuscano, Thorhildur Halldorsdottir, Brandi Stupica, Kristian Owens, and Talia Sacks. 2012. Emotion regulation mediates the relationship between ADHD and depressive symptoms in youth. *Journal of Abnormal Child Psychology*, 40(4):595–606.

Ole Jakob Storeb and Erik Simonsen. 2016. The association between adhd and antisocial personality disorder (aspd): A review. *Journal of Attention Disorders*, 20(10):815–824. PMID: 24284138.

Denis Herbert Stott. 1963. *Bristol Social-adjustment Guides*. University of London Press.

Frank R Vellutino and Donna M Scanlon. 1985. Free recall of concrete and abstract words in poor and normal readers. *Journal of experimental Child psychology*, 39(2):363–380.

Hong Hee Won, Woojae Myung, Gil Young Song, Won Hee Lee, Jong Won Kim, Bernard J. Carroll, and Doh Kwan Kim. 2013. Predicting National Suicide Numbers with Social Media Data. *PLoS ONE*, 8(4):1–6.