Tübingen-Oslo system: Linear regression works the best at Predicting Current and Future Psychological Health from Childhood Essays in the CLPsych 2018 Shared Task

Çağrı Çöltekin
Department of Linguistics
University of Tübingen, Germany
ccoltekin@sfs.uni-tuebingen.de

Taraka Rama
Department of Informatics
University of Oslo, Norway
tarakark@ifi.uio.no

Abstract

This paper describes our efforts in predicting current and future psychological health from childhood essays within the scope of the CLPsych-2018 Shared Task. We experimented with a number of different models, including recurrent and convolutional networks, Poisson regression, support vector regression, and L_1 and L_2 regularized linear regression. We obtained the best results on the training/development data with L_2 regularized linear regression (ridge regression) which also got the best scores on main metrics in the official testing for task A (predicting psychological health from essays written at the age of 11 years) and task B (predicting later psychological health from essays written at the age of 11).

1 Introduction

The words we use reflect who we are and how we are. There have been many successful demonstrations of predicting personal properties and mental state from language use of individuals from many forms of linguistic output including social media text. Examples include predicting basic personal features like gender or age (Barbieri, 2008; Peersman et al., 2011; Burger et al., 2011; Nguyen et al., 2014), predicting personality traits (Luyckx and Daelemans, 2008; Celli et al., 2013; Plank and Hovy, 2015), predicting sentiment towards the topic of a text (Pang et al., 2008), and predicting mental state or health (Ramirez-Esparza et al., 2008; Coppersmith et al., 2014). Most of these studies are based on social media text, and often obtain the target variables (e.g., mental health) through best-effort approaches based on reports by the subjects, or the venues where the texts appear.

The Task A of CLPsych-2018 Shared Task *Predicting Current and Future Psychological Health from Childhood Essays* has a similar aim. The task is to predict the mental health of 11 year old children from the essays they have written. The variables to be predicted are depression and anxiety levels as well as a 'total' value indicating mental health of the child. The Task B, takes a similar but a more ambitious aim, predicting future mental health from the same essays. More precisely, task B comprises of predicting mental health at ages 23, 33, 42, and 50 given the essays from the age of 11. The task B also includes a surprise component where the mental health at 50 years age is not given in the training set. In Task B, outcome variable is the psychological distress score, based on responses to a questionnaire. Both tasks are based on a well-known longitudinal study (Power and Elliott, 2005), where the psychological health (among other variables) were assessed approximately every ten years after the essays were written.

The problems tackled by both tasks, assessing current and/or future mental health from linguistic output, are clearly relevant to monitoring public health, as well as having possible applications to monitoring or diagnosing mental health of individuals. The methods presented can be used to complement well-established traditional methods, as well as providing an alternative where traditional methods are not possible or difficult to employ. This all may be possible, of course, if one can achieve reasonable accuracies in these task.

As part of our submission to the CLPsych-2018 Shared Task, we experimented with a number of different models, both traditional and relatively new, for predicting the outcome of both tasks. However, we did not attempt to predict age-50 mental health in Task B. Although, we only submitted the results from ridge regression models, we also describe a few other promising models, such as Poisson regression and recurrent neural networks (RNNs) that we experimented with.

2 Models

In this section, we describe the models we experimented with. Essentially, we experimented with a number of 'traditional' regression models, and also a few neural network architectures, which also differ in the way the features are presented to the systems. Although we were able to experiment with the models discussed here extensively we discuss each model briefly and do not report results with all of these models at this paper.

2.1 Regression (non-neural) models

The non-neural models are trained with bag of n-grams as features which are obtained through concatenating the word and character n-grams of different order and weighted through a global sublinear TF-IDF scaling applied to all the word and character n-grams. We experimented with combinations of different orders of both word and character n-grams. Besides n-gram order, we also experimented with case normalization, and feature selection based on document frequency.

We also explored combination of the control variables (gender and social class) with the textual bag-of-n-gram features. For the experiments where control variables were included, gender was coded as a binary input variable, while we used one-hot (or one-of-k) representation for the social class.

Our submitted system is based on ridge regression, where the weights were trained to minimize the L_2 regularized sum of squared error. We log transformed the outcome variable, which was shifted linearly with a constant (to avoid $\log 0$) before transformation. We applied the inverse transformation at prediction time. In the case of both task A and task B, we treat all the three target variables as outcomes in a single linear regression model. The regularization parameter α_j was tuned separately for each target variable ($1 <= j <= 3$) through cross-validation by searching for the best combination of n-gram orders and α parameter.

In addition to linear regression, we also experimented with Poisson regression for task A. The intuition behind the experiments with Poisson Regression is that the target variable in task A is similar to count data. Therefore, we assume that a target variable such as 'total' variable is drawn from an independent Poisson random variable with mean λ. For a document feature vector x_i, the mean λ_i is equal to $\exp(\theta^T \mathbf{x_i})$. Subsequently, the probability of observing the target variable under a Poisson variable with mean λ_i is given by Poisson distribution. The Poisson regression model is a special case of Generalized Linear Model (Nelder and McCullagh, 1989). The model is trained through applying Stochastic Gradient Descent with RMSProp algorithm (Tran et al., 2015). However, we found that the results of the Poisson regression model were not better than the Ridge regression model at task A.

Besides being 'count data', another property of the outcome variables in this task is that the outcome variables is zero for many data points. This type of data is modeled better through so-called zero-inflated models (Lambert, 1992), which in essence a two-stage model where the data points are classified as zero and non-zero, and a regression model, e.g., Poisson regression, is applied only to data points predicted to be non-zero. We also tried a two-stage model along these lines. However, the initial results were rather discouraging and we did not experiment with this model thoroughly.

We also explored a support vector regression model for task A and task B which is trained in a similar fashion to SVM with TF-IDF features. The model's performance was close but lower than the L_2 regularized linear regression and therefore, we do not report the results of the model.

We also experimented with random forest regression and Bayesian regression both of which have been shown to be useful in a number of similar tasks. However, the implementations we have access to were not scalable due to the dimensionality of TF-IDF vectors and the computation time required. Hence, we do not report any results with these two methods either. The models discussed in this section were implemented with Scikit-learn package (Pedregosa et al., 2011) using liblinear back end (Fan et al., 2008). The Poisson regression model was implemented using Numpy.

2.2 Neural networks

In this paper, we experimented with a neural model consisting of bidirectional Gated Recurrent Units with character and word embeddings trained for the task. The first layer of the model consists of a separate embedding layers built on characters and words. The concatenated output vectors from character and word embeddings are then supplied as input to a Gated Recurrent Network (Cho et al.,

2014). The length of sequence was fixed at 1500 characters for training character embeddings and at 400 words for training word embeddings. The documents are lowercased for training words and characters embeddings. The number of GRU units was also fixed to reflect the sequence length in the case of GRU units. The output of the GRU network had a dimension of 256 and is followed by a fully connected layer with a single output that outputs a real number. The network is trained using Adam optimizer with mean squared error as the objective. All the neural models are implemented using Keras (Chollet et al., 2015) with Tensorflow as the backend (Abadi et al., 2015).

3 Experiments and results

In this section, we describe the dataset, methods, experiments, and results.

3.1 Data

The data for the shared task comes from National Child Development Study (NCDS), which is a longitudinal study following 17,416 babies born in Britain in 1958, (Power and Elliott, 2005). The part of the study relevant to the present task is the essays written by a subset of children at age 11. The shared task training data includes 9217 essays. We used only the training data for which none of the variables were missing. After removing the training instances with missing target variables, we were left with 9146 instances for Task A and 4938 instances for Task B. The training documents used for Task B is a subset of the training documents used for Task A. The length of the documents used for Task A have mean of 964.56 characters and 227.19 words. The document length exhibits quite some variability with standard deviations of 503.07 and 116.52 respectively.

Besides essays, the data includes two background control variables: gender, and the social class at age 11. We also used these variables as inputs to our models. The outcome variables for Task A are scores (number of underlined sentences) indicating anxiety, depression and total score (number of sentences underlined). The outcome variables for Task B are the number of questions indicating psychological distress. An important observation for all outcome variables is that the distributions are heavily skewed with many zero values.

3.2 Evaluation metrics

The predictions of the models are evaluated using mean absolute error and disattentuated R.[1] The systems are ranked using disattentuated R which is a modification to Pearson's R, to account for the correlation between the outcome variables.

3.3 Experimental procedure and hyperparameter tuning

We did not do extensive parameter tuning with all the models section 2. After some initial experiments, ridge regression and support vector regression seemed to yield promising performance scores.[2] Hence, we run a random search through the following hyperparameter values.

`c_ngmax` Maximum character n-gram order: 1–8

`w_ngmax` Maximum word n-gram order: 1–5

`min_df` Document frequency cutoff for feature selection: 1–5

α Regularization constant (α): 0.5–20.0 (we used $1/\alpha$ for SVR margin parameter C)

`lowercase` Case normalization: character n-grams, word n-grams, both or none

`ctrl_weight` Weight of control variables: 0.0–1.0

`a11_weight` Weight of Age-11 predictions: 0.0–1.0 (Task B only)

We used 5-fold cross validation on the training set for determining the best hyperparameter configuration. We trained the model with the best hyperparameters on the complete training set before producing the final predictions.

3.3.1 Task A

For Task A, we obtained best results on training set using 5-fold cross validation using the hyperparameter configuration reported in table 1. We obtained best disattentuated R scores of 0.5778, 0.2315 and 0.4678 for total, anxiety and depression respectively on training set with the parameter values in table 1. Similar performance scores were obtained using other (rather diverse) parameter settings. From a quick inspection, we did not observe any clear trends regarding usefulness

[1] https://www.rasch.org/rmt/rmt101g.htm

[2] To put it another way, the others, e.g., neural networks, we initially expected to perform better did not yield expected results.

Hyperparameter	Ridge	SVR
c_ngmax	5	6
w_ngmax	3	2
min_df	2	1
lowercase	word	word
$\alpha_{\texttt{total}}$	5.0	5.0
$\alpha_{\texttt{anxiety}}$	5.0	10.0
$\alpha_{\texttt{depression}}$	5.0	20.0
ctrl_weight	0.5	0.5

Table 1: Best hyperparameter values for ridge regression (Ridge) and support vector regression (SVR) models for Task A. The values are obtained through a random search from approximately 400 random parameter settings.

of parameter values, except more features (higher c_ngmax and w_ngmax values seem to help).

This model yielded disattentuated R scores of 0.5788, 0.153 and 0.4669 for total, anxiety and depression on the training set respectively. This model also obtained the top rank (based on total score) in Task A among other shared task participants.

3.3.2 Task B

The best parameter settings for both ridge regression and support vector regression models for Task B are reported in table 2. The ridge regression model with the hyperparameter settings reported in table 2 obtained psychological distress correlations (disattentuated) of 0.4118, 0.2919 and 0.2527 for ages 23, 33 and 42, respectively. The inclusion of age-11 predictions as predictors in Task B was useful if gold-standard scores were used. However, we did not observe any benefits if predicted age-11 outcomes were used. As a result, we used the same model, except we did not use the predicted age-11 scores as predictors in our final model. The model obtained disattentuated R scores of 0.443, 0.3175 and 0.1961 for ages 23, 33 and 42, respectively, on the official evaluation. With an average of 0.3189, it ranked best among other participating models.

Similar to Task A, the support vector regression model yielded similar but again slightly lower results. We obtained disattentuated R scores of 0.4092, 0.2779 and 0.215 on the training set with the parameter values presented in table 2.

Hyperparameter	Ridge	SVR
c_ngmax	4	7
w_ngmax	5	5
min_df	1	1
lowercase	word	word
$\alpha_{\texttt{total}}$	3.0	8.0
$\alpha_{\texttt{anxiety}}$	8.0	20.0
$\alpha_{\texttt{depression}}$	10.0	20.0
ctrl_weight	1.0	0.1
a11_weight	0.5	0.1

Table 2: Best hyperparameter values for ridge regression (Ridge) and support vector regression (SVR) models for Task B. The values are obtained through a random search from approximately 400 random parameter settings. a11_weight is based on predicted age-11 outcomes.

4 Discussion

In this paper we described the models we experimented with in our participation of CLPsych-2018 Shared Task, and reported our results. For our submission, we used a ridge regression model with bag-of-n-gram features as our final model. The model we used is simple linear regression with L_2 regularization, except the outcome variable was log-transformed. The model obtained best results on both tasks at official evaluation. We have also experimented with a range of other models, including Poisson regression and neural models. However, based on (somewhat) limited tuning efforts, none of these systems achieved scores close to ridge regression and support vector regression models (where the SVR model was close to the results of ridge regression model).

The promise of this line of work, namely, predicting mental health from language samples is interesting scientifically and it may also have important applications in monitoring public and personal health. Predicting future mental health is even more interesting as it may allow the clinicians to identify preventive interventions. The method is even more relevant and easily applicable due to the increase in the longitudinal collection of language output in the last few decades., These prospects are only attainable if we can predict present or future mental health from language samples with reasonable accuracy. Our results show that there is in fact a rather strong signal in the language samples for detecting mental health. As expected, the

predictions are less reliable as time gap between the language output and the prediction is wider. However, there is small but reliable correlation between language output at age 11 and psychological health at age 42.

The correlation results of both this paper and other participants clearly show that the linguistic data contains cues that can predict current and future mental health. One of the interesting questions is whether we can get more information from this data. To this end, better modeling of the data may allow us to get more information, i.e., detect more of the signal within the text at hand. With models involving Poisson regression, zero-inflated models, and neural networks, we tried to use models that are more suitable for the data at hand. However, we obtained best results with relatively simple linear models applied to log transformed data. The success of the simple linear models over more complex (e.g., neural networks) is in line with our experiences in a diverse set of text classification tasks (Çöltekin and Rama, 2016; Rama and Çöltekin, 2017; Çöltekin and Rama, 2018). The results indicating superiority of simple linear models, however, should not be considered as conclusive since exploration of more complex models was not thoroughly performed due to time limitations.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Federica Barbieri. 2008. Patterns of age-based linguistic variation in american english. *Journal of sociolinguistics*, 12(1):58–88.

John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1301–1309. Association for Computational Linguistics.

Fabio Celli, Fabio Pianesi, David Stillwell, Michal Kosinski, et al. 2013. Workshop on computational personality recognition (shared task). In *Proceedings of the Workshop on Computational Personality Recognition*.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111.

François Chollet et al. 2015. Keras. `https://github.com/keras-team/keras`.

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.

Çağrı Çöltekin and Taraka Rama. 2016. Discriminating similar languages with linear SVMs and neural networks. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 15–24, Osaka, Japan.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Diane Lambert. 1992. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Kim Luyckx and Walter Daelemans. 2008. Personae: a corpus for author and personality prediction from text. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

J. A. Nelder and Peter McCullagh. 1989. *Generalized Linear Models*, second edition. CRC Press, Boca Raton, FL.

Dong Nguyen, Dolf Trieschnigg, A Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska De Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1950–1961.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning

in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.

Barbara Plank and Dirk Hovy. 2015. Personality traits on twitter—or—how to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98.

Chris Power and Jane Elliott. 2005. Cohort profile: 1958 british birth cohort (national child development study). *International journal of epidemiology*, 35(1):34–41.

Taraka Rama and Çağrı Çöltekin. 2017. Fewer features perform well at native language identification task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 255–260, Copenhagen, Denmark. Association for Computational Linguistics.

Nairan Ramirez-Esparza, Cindy K Chung, Ewa Kacewicz, and James W Pennebaker. 2008. The psychology of word use in depression forums in english and in spanish: Texting two text analytic appr oaches. In *International Conference on Weblogs and Social Media*, pages 102–108.

Dustin Tran, Panos Toulis, and Edoardo M Airoldi. 2015. Stochastic gradient descent methods for estimation with large data sets. *arXiv preprint arXiv:1509.06459*.

Çağrı Çöltekin and Taraka Rama. 2018. Tübingen-Oslo at SemEval-2018 task 2: SVMs perform better than RNNs at emoji prediction. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, United States.