Predicting Psychological Health from Childhood Essays with Convolutional Neural Networks for the CLPsych 2018 Shared Task (Team UKNLP)

Anthony Rios[1], **Tung Tran**[1], and **Ramakanth Kavuluru**[1,2]
[1]Department of Computer Science
[2]Division of Biomedical Informatics, Department of Internal Medicine
University of Kentucky, Lexington, KY
ramakanth.kavuluru@uky.edu

Abstract

This paper describes the systems we developed for tasks A and B of the 2018 CLPsych shared task. The first task (task A) focuses on predicting behavioral health scores at age 11 using childhood essays. The second task (task B) asks participants to predict future psychological distress at ages 23, 33, 42, and 50 using the age 11 essays. We propose two convolutional neural network based methods that map each task to a regression problem. Among seven teams we ranked third on task A with disattenuated Pearson correlation (DPC) score of 0.5587. Likewise, we ranked third on task B with an average DPC score of 0.3062.

1 Introduction

The Fifth Annual Workshop on Computational Linguistics and Clinical Psychology (CLPsych) includes a shared task on predicting current and future psychological health from childhood essays. The organizers provided participants with a dataset of 9217 essays written by 11-year-olds and 4235 essays written at age 50 for training. 1000 age 11 essays are provided for testing. The data is from the National Child Development Study (NCDS) (Power and Elliott, 2005) which followed the lives of 17000 people born in England, Scotland, and Wales in 1958. There are three shared tasks using this dataset: (i) Task A involves predicting behavioral health scores at age 11 using childhood essays. Specifically, participants were asked to develop methods to score the anxiety and depression levels of a child given their essay. (ii) Task B asks participants to predict future psychological distress at ages 23, 33, 42, and 50 using the age 11 essays. Ground truth training scores are provided for ages 23, 33, and 42. Participants are not given age 50 distress scores and must infer them based on scores at the previous ages. (iii) The innovation challenge involves generating essays written at age 50 given the age 11 essays.

In this paper, we summarize our submission for the 2018 CLPsych shared tasks A and B. This paper is organized as follows: Section 2 describes our two submissions – models UKNLPA and UKNLPT. In Section 3, we present the official results and then discuss future directions in Section 4.

2 Methods

We submitted results from two different models, UKNLPA and UKNLPT, to tasks A and B. Both use the same convolutional neural network (CNN) architecture that has been shown to work well across a wide variety of tasks (Kim, 2014; Rios and Kavuluru, 2015, 2017; Tran and Kavuluru, 2017). After a brief overview of the CNN architecture in Section 2.1, we describe the UKNLPA model in Section 2.2 and the UKNLPT model in Section 2.3.

2.1 Convolutional Neural Networks

The basic CNN architecture for both UKNLPA an UKNLPT are shown in Figure 1. The CNN contains three main components. The first component is the input layer, which takes an essay x as input and represents it as a matrix $\mathbf{D}$, where each row is a word vector. The number of rows will depend on the number of words in the essay. The next component transforms $\mathbf{D}$ into a vector. Convolution filters transform every successive n-gram (n successive word vectors) into a real number. The convolution layer, applied to every successive n-

Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pages 107–112
New Orleans, Louisiana, June 5, 2018.

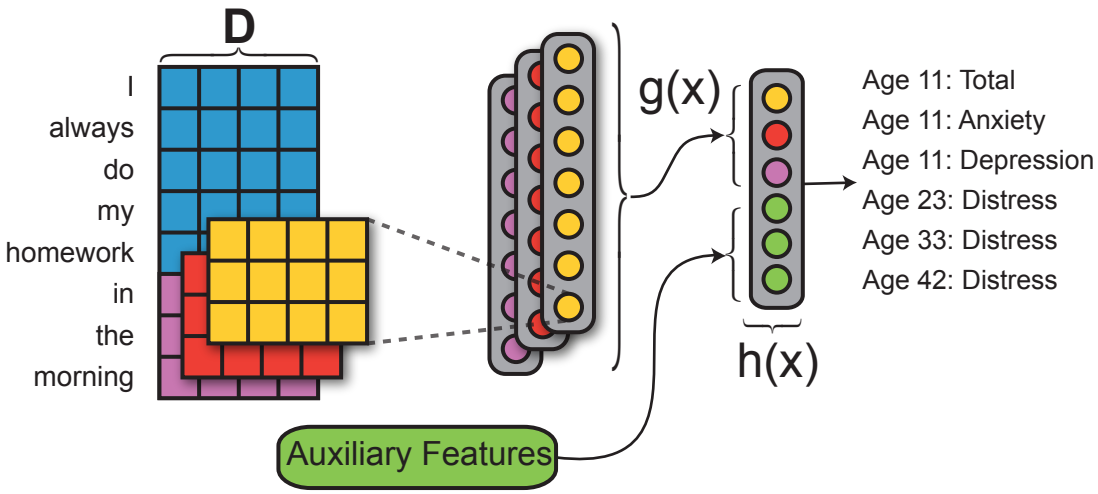

Figure 1: The CNN model layout. We append auxiliary features to the max-pooled CNN features then pass it to an affine output layer. For UKNLPA, the auxiliary features are the 59 LIWC features and the gender. UKNLPT uses LIWC, gender, and social class auxiliary features.

gram in the essay, will produce a vector representation (feature map) of the essay. The length of the feature map will depend on the number of words in the essay. Multiple convolution filters produce multiple feature maps. To form a fixed-size vector representation of the essay, we use max-over-time pooling across each feature map. These max values are combined to form the final fixed-size vector representation of the essay. In the remainder of this paper we refer to the fixed size vector as $g(x)$. Finally, we refer to prior work (Kim, 2014) for more details about the architecture.

2.2 UKNLPA

For our first model, we represent each essay x as

$$h(x) = g(x) \parallel l(x) \parallel s(x)$$

where $h(x)$ is the concatenation of the CNN feature vector $g(x)$ with 59 Linguistic Inquiry and Word Count (LIWC) features (Pennebaker et al., 2001) $l(x)$ and a binary feature $s(x)$, representing the gender of the child.

Let m represent the six psychological health scores for both task A and B: age 11 total, age 11 anxiety, age 11 depression, and distress values at ages 23, 33, and 42. To predict these six scores, we pass $h(x)$ through an affine output layer

$$\hat{y} = W\, h(x) + b$$

where $\hat{y} \in \mathbb{R}^m$, $W \in \mathbb{R}^{m \times p}$ and $b \in \mathbb{R}^m$.

Training Procedure To train our model we use the huber loss as our training objective. The huber loss combines both the mean squared error (MSE) loss with the mean absolute error (MAE) loss. We define the huber loss as

$$L_\delta(y', \hat{y}) = \begin{cases} \frac{1}{2}(y' - \hat{y})^2 & \text{for } |y' - \hat{y}| \leq \delta \\ \delta|y' - \hat{y}| - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases}$$

where δ is a hyperparameter that weights the difference between between MSE and MAE and y' is the ground truth encoding for one of the six psychological health factors. For small errors, the huber loss is equivalent to MSE and a weighted MAE is used for large errors. Therefore, the huber loss is less sensitive to outliers compared to MSE.

During preliminary experiments, we tried training all outputs jointly and separately. We found our model performs best across all psychological health factors when trained jointly except for age 11 total. Thus, we trained two models. One with a multi-task loss

$$\ell_{\hat{y}} = \sum_{j=1}^{m} L_\delta(y'_j, \hat{y}_j)$$

optimized across all six heath factors and one model trained only on age 11 total. We mask the loss for missing values of a particular outcome variable. Finally, because age 50 ground truth scores were not given for training, we output the age 42 predictions directly as the scores for age 50.

Linear Model We train a ridge regression model with three sets of features: term frequency–inverse document frequency (TFIDF) weighted unigrams and bigrams, 59 LIWC features, and a binary feature representing gender.

Ensemble Our final UKNLPA model is an ensemble of multiple CNNs and the linear model. Specifically, we average the predictions of five CNNs trained on different 80/20 splits of the training datasets with the predictions from the linear model, where all models are weighted equally.

Model Configuration and Preprocessing We preprocess each essay by lowercasing all words. Next, we replace each newline character with a special NEWLINE token and replace all illegible words with the token ILLEGIBLE. Likewise, all words that appear less than five times in the training dataset are replaced with the token UNK. For tokenization, we use a simple regex (\w\w+). We train the UKNLPA model with the Adam optimizer (Kingma and Ba, 2015) using a learning rate of 0.001. We initialize the word vectors of our model with 300 dimensional pre-trained 6B glove embeddings[1] (Pennington et al., 2014). The CNN is trained with windows that span 3, 4, and 5 words with 300 filters per window. Hence, the final neural vector representation of each essay $h(x)$ has 960 dimensions. Our model is regularized using both dropout and L2 regularization. We apply dropout to the embedding layer and to the CNN output $g(x)$ with a dropout probability of 0.2. The L2 regularization parameter is set to 0.001 and the huber loss parameter δ is set to 0.1. We train for a total of 25 epochs with a mini-batch size of 50 and checkpoint after each epoch. The best checkpoint based on a held-out validation dataset is used at test time. For the linear model, we set L2 regularization parameter to 0.1. Finally, we want to note that the social class was not used for UKNLPA. Preliminary experiments showed that it either did not improve or negatively impacted our validation results.

2.3 UKNLPT

The architecture of our second model shares the CNN design introduced in Section 2.2. The final feature vector for this model, $h \in \mathbb{R}^p$, is defined as

$$h(x) = g(x) \parallel l(x) \parallel s'(x)$$

where $g(x)$ is the CNN-based feature vector composition, $l(x)$ is a feature vector encoding LIWC scores, and $s'(x)$ is a feature vector encoding gender and social class for some input essay x. For each example, we emit two sub-outputs: one for linear regression and one for binary classification, the latter serving as as "switch" mechanism which determines whether the regression sub-output is passed to the final output. The regression output denoted by $\hat{y} \in \mathbb{R}^m$, for m output variables, defined such that

$$\hat{y} = W_1 h(x) + b_1$$

where $W_1 \in \mathbb{R}^{m \times p}$ and $b_1 \in \mathbb{R}^m$ are parameters of the network. The sub-output $\bar{y} \in \mathbb{R}^m$ serving as the "switch" is defined as

$$\bar{y} = \sigma(W_2 h(x) + b_2)$$

where $W_2 \in \mathbb{R}^{m \times p}$ and $b_2 \in \mathbb{R}^m$ are parameters of the network and σ is the sigmoid function. The final output $y \in \mathbb{R}^m$ of the network is defined as

$$y_i = \begin{cases} \max(0, \hat{y}_i) & \text{if } \bar{y}_i \geq 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

The idea is to recreate the distribution of the count-based scores by jointly learning to discriminate between the *zero* and *non-zero case*, the former of which occurs frequently in the ground truth. For this model, the age 50 predictions are made based on averaging the age 33 and 42 predictions.

Model Configuration The model is trained with a learning rate of 0.001 using the Adam optimizer (Kingma and Ba, 2015). The input text is lowercased and tokenized on contiguous spans of alphabetic characters using the same regex expression introduced in Section 2.2. The word embeddings are of length 200, randomly initialized without pre-training. The window sizes are 3, 4, and 5 with 200 filters per window size. The CNN-composed feature vector is therefore 600 in length. The LIWC features consist of 59 LIWC scores that have been normalized such that values are in the range $[-1, 1]$. The gender and social class designations are encoded as one-hot vectors and concatenated into a single vector of length 8. Therefore, the length of the final feature vector h is $p = 667$. Moreover, we apply a dropout rate of 50% at the CNN layer and L2 regularization with a λ-weight of 0.1. The model is trained with a mini-batch size of 16 for a maximum of 20 epochs.

Training Procedure Training this model involves optimizing on two separate loss objectives, one for each of the sub-outputs $\hat{y}$ and $\bar{y}$. Suppose

[1] https://nlp.stanford.edu/projects/glove/

Team Name	Total		Anxiety		Depression	
	MAE	DPC	MAE	DPC	MAE	DPC
Coltekin et al.	**5.615**	**0.5788**	0.630	0.1530	0.968	**0.4669**
UGent - IDLab 1	5.691	0.5667	0.476	0.1946	1.004	0.4536
Simchon & Gilead	5.677	0.5205	**0.475**	0.1105	0.947	0.3902
UGent - IDLab 2	5.688	0.514	0.697	0.1760	1.019	0.4192
Liu et al.	5.803	0.4748	0.819	0.0764	1.036	0.3608
TTU	6.050	0.4605	0.704	0.1417	1.055	0.3299
WWBP	6.142	0.4429	0.700	**0.2352**	1.050	0.3616
CLPsych Baseline	6.038	0.4931	0.704	0.1909	1.048	0.4334
uk_ens2 †	5.673	0.5677	0.592	0.1917	0.973	0.4479
uk_cnn †	5.756	0.5483	0.495	0.2214	**0.944**	0.4215
uk_linear †	5.916	0.5421	0.692	0.1419	1.032	0.4314
UKNLPA *	5.695	0.5587	0.526	0.2219	0.951	0.4333
UKNLPT *	5.839	0.5211	0.516	0.0916	**0.944**	0.3395

Table 1: Official task A results. Models we submitted for the competition are marked with *. Our models that were not official submissions for the competition are marked with †.

the ground truth is encoded as a vector $y' \in \mathbb{R}^m$, where m is the number of target variables to be predicted, then the mean squared error loss $\ell_{\hat{y}}$ for a single example is defined as

$$\ell_{\hat{y}} = \sum_{j=1}^{m} (y'_j - \hat{y}_j)^2$$

where $y'_j, \hat{y}_j$ denotes the jth value of $y', \hat{y}$ respectively. For the *switch* output, $\bar{y}$, the example-based binary cross entropy loss is defined as

$$\ell_{\bar{y}} = -\sum_{j=1}^{m} \gamma_j \log(\bar{y}_j) + (1 - \gamma_j) \log(1 - \bar{y}_j)$$

where $\gamma_j = \min(y'_j, 1)$. Each example-based loss is mean-averaged over the batch dimension to obtain a *mini-batch* loss. The learning objectives are trained in alternation for each mini-batch. We check-point at each epoch; the epoch with the best score (based on averaging the DPC measure over the m prediction variables) on the held-out development set of 500 examples is kept for test-time predictions.

We train two separate "instances" of the aforementioned model, one to learn on the a11_bsag_total variable and one to learn on the remaining five variables which share a similar range and distribution jointly: a11_bsag_anxiety, a11_bsag_depression, a23_pdistress, a33_pdistress, and a42_pdistress. Each "instance" is an ensemble of 3 models each trained with a different random parameter initialization and training/development set split.

3 Experiments

In this section, we compare our methods on the official test set. The competition reports two evaluation metrics: mean absolute error (MAE) and disattenuated pearson correlation (DPC)[2]. Final rankings for task A are based on the Total DPC. The average of the age 23 to 42 distress DPC scores are used to rank participants on task B.

Besides our two submissions, UKNLPA and UKNLPT, we also report the results for three variants of UKNLPA:

- uk_linear – the ridge regression model introduced in Section 2.2.
- uk_cnn – an ensemble consisting of five CNNs trained on different 80/20 splits of the training dataset.
- uk_ens2 – an ensemble of uk_linear and uk_cnn. Compared to the method described in Section 2.2, uk_ens2 gives more weight to the linear model.

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

	Avg. Ages 23-42		Age 23		Age 33		Age 42		Age 50	
Team Name	Avg. MAE	Avg. DPC	MAE	DPC	MAE	DPC	MAE	DPC	MAE	DPC
Coltekin et al.	1.091	**0.3189**	1.012	0.443	0.987	**0.3175**	**1.275**	0.1961	–	–
TTU	1.176	0.3141	1.087	**0.457**	1.092	0.277	1.350	0.2084	–	–
WWBP	1.117	0.2896	1.061	0.3868	1.008	0.2708	1.283	**0.2113**	1.421	0.0082
Simchon & Gilead	**1.084**	0.2761	**0.991**	0.4542	**0.954**	0.2463	1.308	0.1277	**1.288**	**0.3010**
Radford et al. 1	1.166	0.2300	1.079	0.3957	1.078	0.1054	1.341	0.1890	1.388	0.2087
Liu et al.	1.394	0.2021	1.453	0.2267	1.179	0.2333	1.549	0.1463	–	–
Radford et al. 2	1.172	0.1791	1.093	0.3676	1.098	-0.0403	1.325	0.2100	1.373	0.2137
CLPsych Baseline	1.199	0.2951	1.139	0.4056	1.087	0.283	1.372	0.1967	1.344	0.2569
uk_ens2 †	1.106	0.3095	1.039	0.4246	0.993	0.2935	1.285	0.2104	1.313	0.2558
uk_cnn †	1.082	0.3021	0.998	0.4317	0.969	0.2839	1.279	0.1909	1.291	0.2187
uk_linear †	1.154	0.277	1.113	0.3755	1.017	0.2552	1.331	0.2020	1.370	0.2692
UKNLPA *	1.088	0.3062	1.008	0.4307	0.977	0.2898	1.278	0.1981	1.295	0.2310
UKNLPT *	1.149	0.2259	1.040	0.3781	0.989	0.1878	1.417	0.1117	1.353	0.1675

Table 2: Official task B results. Models we submitted for the competition are marked with *. Our models that were not official submissions for the competition are marked with †.

3.1 Results

The task A results are shown in Table 1. Officially, UKNLPA placed third and UKNLPT placed fourth based on the Total DPC score. We observe that no single method is best across all three categories. uk_ens2 outperforms UKNLPA for the Total category. However, uk_ens2 underperformed UKNLPA and uk_cnn for anxiety. For both Total DPC and Depression DPC, uk_linear performs comparably to uk_cnn. Given that uk_cnn is an ensemble, this suggests that simple linear models are strong baselines for this task. Furthermore, the best performer based on MAE does not necessarily perform best on DPC measures. For example, both UKNLPT and uk_cnn achieve an MAE of 0.944 even though there is a 10% difference between their DPC depression scores. Because each of the psychological health aspects follow a zero-inflated probability distribution (many of the observed ground truth values are zero), MAE favors models that predict zero more often. DPC favors models that are more linearly correlated with the ground truth rather than predicting the exact psychological scores compared to uk_cnn.

Table 2 shows the official results for task B. UKNLPA ranked third, while UKNLPT ranked seventh. Similar to task A, we find that on average uk_ens2 slightly outperforms UKNLPA. Furthermore, we find that no single method performs best across all ages. We observe that uk_linear outperforms the CNN ensemble uk_cnn for ages 42 and 50 distress DPC metrics. However, uk_cnn outperforms uk_linear for age 23 and 33. For all methods except UKNLPT, we use the age 42 predictions to predict age 50 distress because ground truth age 50 distress scores was not provided for the training dataset. Because uk_linear performed better on age 42 compared to uk_cnn, it also performs best on age 50. Likewise, because uk_cnn performs poorly on age_50, when we ensemble it with uk_linear it performs worse compared to only using uk_linear.

4 Conclusion

In this paper, we describe our submissions to the 2018 CLPsych shared tasks A and B. Overall, our method UKNLPA ranked third on both tasks and UKNLPT ranked fourth on task A. We identify two avenues for future work.

- The childhood essays contain certain common characteristics. For example, many essays contain illegible words and spelling mistakes. If a word is misspelled, then we may ignore it because it occurs infrequently. So we hypothesize that data cleaning techniques such as using a spell checker to correct spelling issues may improve our results.

- For both tasks A and B, we observe that no single method performs best across all psychological health categories. Therefore, it would be beneficial to use different methods for each category depending on what performs best. Furthermore, if we combine the CNN and linear models with more sophisticated ensemble approaches, we may improve our overall results.

Office of Research Integrity Review

This study has undergone ethics review by the University of Kentucky IRB and has been deemed exempt given it does not involve human subjects, more specifically because, (1). the data analyzed is de-identified, (2). we (the participants) do not have access to a code to re-identify subjects, and (3). there is no collaborator listed on our protocol who has access to identifiers.

Acknowledgments

We thank anonymous reviewers for their constructive comments despite the short turnaround. This research is supported by the U.S. National Library of Medicine through grant R21LM012274. We also gratefully acknowledge the support of the NVIDIA Corporation for its donation of the Titan X Pascal GPU used for this research.

References

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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

Anthony Rios and Ramakanth Kavuluru. 2015. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. In *6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 258–267.

Anthony Rios and Ramakanth Kavuluru. 2017. Ordinal convolutional neural networks for predicting rdoc positive valence psychiatric symptom severity scores. *Journal of biomedical informatics*, 75:S85–S93.

Tung Tran and Ramakanth Kavuluru. 2017. Predicting mental conditions based on history of present illness in psychiatric notes with deep neural networks. *Journal of biomedical informatics*, 75:S138–S148.