Replicating "Extracting schemas from thought re cords using natural language processing"

Cameron Buster and Elijah Rockers[1]
cbuster6@gatech.edu, erockers3@gatech.edu

***Abstract*— As a final project for CSE 6250, we attempt to replicate the paper "Extracting schemas from thought records using natural language processing" by Burger et al [1]. We successfully construct and reproduce two of the four original hypotheses (H1 and H2) finding key model performance metrics differing slightly from the original paper. For H1, we find that schemas most schemas (Attachment, Competence, Global self-evaluation, Health, Power and Control, Hopelessness, Other's views on self) were able to be automatically extracted with kNN-C (k = 7), all schemas were able to be extracted from kNN-R (k = 6), most schemas were able to be extracted with SVC (Attachment, Competence, Global self-evaluation, Health, Hopelessness, and Other's views on self), all schemas were able to be extracted with SVR, and all schemas were able to be extracted with multi-label RNN. We are unable to reproduce H4 due to a bug in the per-schema RNN output.**

1 INTORDUCTION

In Sec. 1, we discuss the original paper and how it contributes to cognitive psychotherapy using machine learning and data techniques that rely on natural language processing (NLP).

1.1 High Level Project Overview

The original paper conducts a cognitive approach to psychotherapy which primarily aims to change a maladaptive schema (negative view) that a patient may have about themselves, the world, or the future.

To obtain maladaptive schemas, records of a patient's (N=320) thought process are obtained through invoking emotional responses via questionnaire. Each patient is asked five questions consisting of various "utterances" that reflect a cognitive process – attachment, competence, global self-evaluation, health, power and control, metacognition, other people, hopelessness, other's views on self. Each of these cognitive processes correspond to the response variables used in modeling.

[1] elijahrockers/cse6250_project (github.com)

Additionally, the original paper uses NLP software pretrained on all 2014 English Wikipedia articles where the representations of words and utterances are then mapped to schemas using k-nearest neighbors (kNN), support vector machines (SVC, SVR), and recurrent neural networks.

The result pipeline of the original paper found that for the more frequently occurring schemas, all ML were able to leverage linguistic patterns. For six of the nine schemas, two RNN trained per schema outperformed the other ML methods.

In total, 1600 thought records for 320 patients comprising 5747 utterances is presented as the baseline performance for psychotherapists and ML researchers interested in contributing to a cognitive approach to psychotherapy.

Lastly, utterances are gained for each patient using the downward arrow technique (DAT). To use DAT, the immediate and unreflected appraisal of a situation is called an automatic thought. These automatic thoughts are often determined by thought schemas which are the cognitive structures that make up one's world view. This core part of cognitive psychotherapy involves teaching patients to monitor thoughts. By stating the automatic thought that occurs regarding the schema being explored, the patient is asked why it would be upsetting or what would be the worst that could happen in regard to the automatic thought last mentioned. This is how utterances are obtained. DAT can be applied as many times are the patient will allow.

2 SCOPE OF REPRODUCIBILITY

In Sec. 1, we discuss the overview of the original paper including key vocabulary required to understand what topics from psychotherapy are being explored and how those ideas are being leveraged for results relevant to field using ML. In Sec. 2, we will discuss the scope of reproducibility including which of the four hypotheses from the original paper we intend to reproduce.

2.1 Original Hypotheses

The original paper tests four original hypotheses. The first, H1, is that schemas can be extract automatically from the utterances. The future intent of this hypothesis is to create a conversational agent that can conduct cognitive therapy utilizing positive feedback and DAT. This figure intent also necessitates the study of ways to improve automatic schema identification resulting in three additional hypotheses. H2, automatic predictions improve as the DAT progresses. H3, within individuals, similar situations will activate the same schemas. H4, across individuals, there is a relationship

between the active schemas and score on mental health scales. We intend to reproduce H1, H2, and H4.

For H1, we intend to reproduce the paper results by running kNN-C (k = 4), kNN-R (k = 5), SVC and SVR using the basis function kernel.

The multi-label RNN trained in batches of 32 utterances with 100 epochs. The baseline RNN will consist of two hidden layers: the first an embedding layer (on all 2014 English Wikipedia articles) and a bidirectional long short-term memory later of 100 nodes. Dropout probability of 0.1, categorical cross-entropy loss, sigmoid activation function for each nine nodes of the output layer and mean absolute error as the key model evaluation metric. The per schema RNN will be setup the same as far as hyperparameters go but the modeling approach slightly differed in the original paper from the multi-label approach. Mainly, the per schema RNN have four output which correspond to one of four possible scores for each schema. 0, has absolutely nothing to do with the schema. 1, correspond a little bit with the schema. 2, corresponds largely with the schema. 3, corresponds completely with the schema. The activation function of the final layer on the per schema RNN is softmax which expresses "the likelihood with which a certain utterance has each of the scores" [1].

We will begin with these choosing of ML as a baseline performance then we will adjust hyperparameters in search of better key performance metrics.

For H2, we intend to calculate the mean correlation between the manually labeled schema scores in the training set versus the predicted schema scores on both the test and validation set. The original paper found that steps at a deeper DAT level could not score better than the best H1 model than steps at a shallower DAT level. The original paper uses the Spearman correlation metric which we intend to replicate measurement as well.

For H4, five linear regression models are used to explore the link between active schemas of a patient via thought records and the outcomes of five mental health categories. HDAS-Anxiety and two Cognitive Distortion scales correlated with two schemas - Global Self-Evaluation and Power and Control. HDAS-Anxiety is linked to Global Self-Evaluation, $p=0.003$. Cognitive Distortions-Relatedness, $p<0.001$. Power and Control schema also significantly correlates with Cognitive Distortion-Relatedness, $p=0.007$. With our trained ML, we will report our p-values to see if our ML is significantly like the trained ML in the original paper.

3 METHODOLOGY

In Sec. 2, we discuss the scope of reproducibility, hypotheses from the original paper, hypotheses that we hope to recreate, and the corresponding experiments required to recreate them. In Sec. 3, we will discuss the methodology required to recreate the models required to test the hypotheses stated in Sec. 2.

3.1 Model description

Five different algorithms are used – kNN-C, kNN-R, SVC, SVR, and RNN. The RNN differs in two ways – a per schema approach and a multilabel approach.

For kNN-C, we run k from 2 to 10 and use Spearman Correlation (SC) as the training objective. For kNN-R, we run k from 2 to 10 and use Spearman Correlation as the training objective. For SVC, default parameters were used. For SVR, default parameters were used. For the per schema RNN, we were unable to successfully run the model. For the multilabel RNN, a total of 253,809 parameters were discovered, 122,609 being trainable, 131,200 untrainable. The best multilabel RNN had a batch size of 64, dropout of 0.5, ran for 100 epochs, uses mean absolute error as the loss function, 100 LSTM units, and uses Adam as the primary optimizer.

For H2, we had to use the KNN-R model, because the RNN models weren't working, and the SVM model dropped rows which made it impossible to re-associate the predictions with the depth of the utterance.

3.2 Dataset Description

The source of the data[2] was obtained from the 4TU ResearchData repository directly accessible via the peer-reviewed publication itself.

The data was collected organically from 320 patients subjecting themselves to a cognitive approach to psychotherapy. The dataset, in total, is comprised of 1600 thought records and 5747 utterances each corresponding to a patient and a psychotherapeutic schema.

The dataset, after preprocessing, is 5016 records in total. 80% training, 20% testing with an additional 734 record holdout validation set. There are nine labels. Due to the multi label distribution of the labels, the distribution is quite hard to describe in one succinct

[2] Dataset and Analyses for Extracting Schemas from Thought Records using Natural Language Processing (4tu.nl)

image. Records can be associated with multiple labels so without multidimensional figures or mapping the labels in a binary fashion so that the labels can be displayed in two dimensions (thus, only serving as an approximation), showing the label distribution wouldn't be very academic.

3.3 Computational Implementation

We use Jupyter notebooks to replicate the code. From Scikit-Learn, we import shallow learning techniques kNN-C, kNN-R, SVC, and SVR (from SVM). The RNN is constructed via the Keras-Tensorflow API wrapper. The hardware used is an Intel Core i3 7100 @ 3.90GHz, 40GB RAM, running on the course-provided Docker container.

The shallow learning techniques use default hyperparameters. The RNN utilizes (None, 25, 50) Embedding, (None, 200) Bidirectional Long Short-Term memory nodes, (None, 200) dropout rate, (None, 9) Dense. Each run ran 100 epochs with either binary cross entropy loss or categorical cross entropy less with 50 or 100 LSTM units and either RMSprop or Adam as the model optimizer.

3.4 Code

The code was slightly tweaked by us to run and collect measurements however, most of the project was treated as a replication in that we attempted to replicate the experiment exactly rather than rebuild it from the ground up. Given the amount of time given, we felt this was a more practical approach.

4 RESULTS

In Sec. 3, we discuss the methodology relating to the modeling architecture, training objectives, dataset descriptions, and computational implementation. In Sec. 4, we will discuss the results for all experiments that were run to test hypotheses and compare these results with the results from the original paper.

4.1 Scores and Key Metrics

H1

For kNN-C, we find the following results – Attachment, SC = 0.556. Competence, SC = 0.664. Global self-evaluation, SC = 0.362. Health, SC = 0.771. Power and Control, SC = 0.112. Meta-cognition, SC = NaN. Other people, SC = 0.322. Hopelessness, SC = 0.498. Other's views on self, SC = 0.480. This compares to the original paper which found the

following results - Attachment, SC = 0.55. Competence, SC = 0.69. Global self-evaluation, SC = 0.40. Health, SC = 0.74. Power and Control, SC = 0.11. Meta-cognition, SC = NaN. Other people, SC = 0.28. Hopelessness, SC = 0.48. Other's views on self, SC = 0.45

For the kNN-R, we find the following results - Attachment, SC = 0.583. Competence, SC = 0.635. Global self-evaluation, SC = 0.430. Health, SC = 0.536. Power and Control, SC = 0.256. Meta-cognition, SC = 0.046. Other people, SC = 0.185. Hopelessness, SC = 0.426. Other's views on self, SC = 0.466. This compares to the original paper which found the following results - Attachment, SC = 0.63. Competence, SC = 0.66. Global self-evaluation, SC = 0.41. Health, SC = 0.53. Power and Control, SC = 0.23. Meta-cognition, SC = 0.10. Other people, SC = 0.24. Hopelessness, SC = 0.51. Other's views on self, SC = 0.46

For the SVC, we find the following results - Attachment, SC = 0.599 [0.548, 0.644]. Competence, SC = 0.666 [0.628, 0.706]. Global self-evaluation, SC = 0.360 [0.282, 0.407]. Health, SC = 0.650 [0.570, 0.732]. Power and Control, SC = NaN [0.000, 1.000]. Meta-cognition, SC = NaN [0.000, 1.000]. Other people, SC = NaN [0.000, 1.000]. Hopelessness, SC = 0.415 0.351, 0.470]. Other's views on self, SC = 0.422 [0.366, 0.475]. This compares to the original paper which found the following results - Attachment, SC = 0.65. Competence, SC = 0.68. Global self-evaluation, SC = 0.36. Health, SC = 0.73. Power and Control, SC = NaN. Meta-cognition, SC = NaN. Other people, SC = NaN. Hopelessness, SC = 0.54. Other's views on self, SC = 0.48

For the SVR, we find the following results - Attachment, SC = 0.653 [0.625, 0.679]. Competence, SC = 0.619 [0.589, 0.649]. Global self-evaluation, SC = 0.431 [0.402, 0.475]. Health, SC = 0.335 [0.297, 0.371]. Power and Control, SC = 0.236 [0.183, 0.283]. Meta-cognition, SC = 0.069 [0.016, 0.113]. Other people, SC = 0.143 [0.099, 0.187]. Hopelessness, SC = 0.506 [0.462, 0.536]. Other's views on self, SC = 0.479 [0.440, 0.518]. This compares to the original paper which found the following results - Attachment, SC = 0.68. Competence, SC = 0.64. Global self-evaluation, SC = 0.49. Health, SC = 0.35. Power and Control, SC = 0.31. Meta-cognition, SC = 0.11. Other people, SC = 0.19. Hopelessness, SC = 0.54. Other's views on self, SC = 0.52.

For the multilabel RNN, we find the following results - Attachment, SC = 0.139. Competence, SC = 0.209. Global self-evaluation, SC = 0.151. Health, SC = 0.066. Power and Control, SC = 0.083. Meta-cognition, SC = 0.053. Other people, SC = 0.057. Hopelessness, SC = 0.087. Other's views on self, SC = 0.195. This compares to the original paper which

found the following results - Attachment, SC = 0.73. Competence, SC = 0.76. Global self-evaluation, SC = 0.58. Health, SC = 0.75. Power and Control, SC = 0.28. Meta-cognition, SC = -0.01. Other people, SC = 0.22. Hopelessness, SC = 0.63. Other's views on self, SC = 0.58.

H2

Average spearman correlations for each depth level of KNN-R modeled utterances were as follows, from depth 1 through 12 respectively:

[0.59, 0.52, 0.55, 0.52, 0.67, 0.58, 0.72, 0.50, 0.49, 0.67, 0.65, 0.60]

As expected, performing a linear regression on these results yielded a non-significant r-value.

4.2 Original Paper Comparison

We were able to successfully replicate the kNN-C, kNN-R, SVC, and SVR quite closely albeit with slightly different hyperparameters for kNN-C (k = 7 instead of k = 4) and kNN-R (k = 6 instead of k = 5). SVC and SVR are exactly replicated.

RNN proved to be the most challenging part of this replication. While we were able to run the multilabel RNN, the results are quite poor compared to the other shallow learning techniques. The original paper found that the multilabel RNN performed the best for detecting the Power and Control schema as well as the Meta-cognition schema.

Finally, the original paper found that the per-schema RNN was the highest performing ML algorithm on all schemas except for the Power and Control and Meta-cognition schemas. We were unable to replicate the paper in this way due to a bug that prevented us from running the per-schema RNN successfully at this time.

As expected, the DAT was not found to converge when considering depth of thought records.

5 DISCUSSION

In Sec. 4, we discussed the results for all experiments that were run to test hypotheses and compare these results with the results from the original paper. In Sec. 5, we will discuss the overall reproducibility of the paper, describe our successes and difficulties during the reproduction, and make suggestions to the author or other reproducers on how to improve the reproducibility.

5.1 Assessing Reproducibility

We ran into issues running the RNN models, primarily due to the 16 hour computational time required to attempt debugging efforts. Additionally, it appears that H2 was given little attention, and H4 none at all in the provided code. That said, the authors did an exceptional job at making all the data available.

5.2 Successes and Difficulties

It was quite easy to test H1 and H2 with shallow learning techniques. It was the deep learning section and in particular the per-schema RNN that we were not able to replicate due to NaN output for each schema. We were able to run the multilabel RNN, but the results were vastly different than what the original paper found for the output for that model. One of the main issues contributing to this was the 16 hour computational time that was often needed to attempt to debug these models.

5.3 Suggestions

We would recommend the authors provide explicit code for reproducing results for each hypothesis. For reproducers, it would also be wise to consider possible optimizations if problems arise with the long computational times necessary.

6 REFERENCE

1. Burger, Franziska, Neerincx, Mark A., Brinkman, Willem-Paul. October 18, 2021. Natural language processing for cognitive therapy: Extracting schemas from thought records. *PlosOne*. Retrieved from Natural language processing for cognitive therapy: Extracting schemas from thought records (plos.org).

YouTube Video Presentation:

https://youtu.be/kuVp6EdyTxk