Verification and Validation Report: Natural Language Processing for Mental Health Risk Prediction

Team 13, The Cognitive Care Crew
Jessica Dawson
Michael Breau
Matthew Curtis
Benjamin Chinnery
Yaruo Tian

March 6, 2024

1 Revision History

Date	Version	Notes
Date 1	1.0	Notes
Date 2	1.1	Notes

2 Symbols, Abbreviations and Acronyms

symbol	description
T	Test

[symbols, abbreviations or acronyms – you can reference the SRS tables if needed —SS]

Contents

List of Tables

List of Figures

This document ...

3 Functional Requirements Evaluation

3.1 Task 3

1. **FRT-T3-1**

   Control: Manual

   Initial State: The task 3 model has been trained on a section of the training data provided by eRisk

   Input: A section of the training data that the model has not yet seen

   Output: The system will output predictions to a txt file where each line is of the format "{Username} {prediction for Q1} ... {prediction for Q28}"

   How test will be performed: A tester will run the system on the input data and verify that an output file is created and that it is of the correct format

   Result: Passed (due to privacy concerns the output text file can not be included here for verification)

2. **FRT-T3-2**

   Control: Manual

   Initial State: The task 3 model has been trained on a section of the training data provided by eRisk

   Input: A section of the training data that the model has not yet seen

   Output: The system will output predictions to a txt file where each prediction should be between 0 and 6

   How test will be performed: A tester will run the system on the input data and verify that none of the predictions in the output file are less than 0 or greater than 6

   Result: Passed (due to privacy concerns the output text file can not be included here for verification)

3. **FRT-T3-3**

   For test FRT-T3-3, FRT-T3-4, and FRT-T3-5 the system produced the following results:

	MAE	MZOE	RS	ECS	SCS	WCS	GED
zeros	3.58	0.88	3.63	2.93	4.04	3.64	3.45
sixes	2.42	**0.65**	3.13	3.68	2.91	2.91	3.02
average	**2.00**	0.90	**1.63**	1.59	2.04	**1.65**	**1.47**
model	2.60	0.82	2.45	**1.47**	**1.98**	1.79	1.59

   Control: Manual

   Initial State: The task 3 model has been trained on a section of the training data provided by eRisk and accuracy measures have been calculated for a baseline strategy where a score of 0 is guessed for every question

   Input: A section of the training data that the model has not yet seen

   Output: The system will print accuracy metrics for the model to the console where the model scores higher than the baseline strategy on at least one metric

   How test will be performed: A tester will run the system on the input data and verify that the model outperforms the baseline on at least one metric

   Result: Passed

4. **FRT-T3-4**

   Control: Manual

   Initial State: The task 3 model has been trained on a section of the training data provided by eRisk and accuracy measures have been calculated for a baseline strategy where a score of 6 is guessed for every question

   Input: A section of the training data that the model has not yet seen

   Output: The system will print accuracy metrics for the model to the console where the model scores higher than the baseline strategy on at least one metric

How test will be performed: A tester will run the system on the input data and verify that the model outperforms the baseline on at least one metric

Result: Passed

5. **FRT-T3-5**

   Control: Manual

   Initial State: The task 3 model has been trained on a section of the training data provided by eRisk and accuracy measures have been calculated for a baseline strategy where the average score for a question based on the training data is guessed

   Input: A section of the training data that the model has not yet seen

   Output: The system will print accuracy metrics for the model to the console where the model scores higher than the baseline strategy on at least one metric

   How test will be performed: A tester will run the system on the input data and verify that the model outperforms the baseline on at least one metric

   Result: Passed

4 Nonfunctional Requirements Evaluation

4.1 Usability

4.2 Performance

4.3 etc.

5 Comparison to Existing Implementation

5.1 Task 3

A comparison between our model and the best metric scores from last year's iteration of the task are compared. The best scores are the highest accuracy achieved on that metric accross all teams, not the metrics from a "best" model.

	MAE	MZOE	RS	ECS	SCS	WCS	GED
our model	2.60	0.82	2.45	**1.47**	1.98	**1.79**	**1.59**
2023's best	**2.19**	**0.67**	**1.59**	1.92	**1.90**	2.00	2.00

5.1.1 Discussion

The model does manage to outperform previous years on the ECS, WCS, and GED metrics. While we fail to outperform MZOE and RS these metrics were achieved by baselines like guessing all zero, sixes, or the average last year, and given we failed to outperform these baselines on these metrics again this year it is possible this has more to do with the nature of the data and the task itself than our model under performing compared to previous year's models. We only failed to outperform previous year's models on MAE and SCS.

6 Unit Testing

7 Changes Due to Testing

[This section should highlight how feedback from the users and from the supervisor (when one exists) shaped the final product. In particular the feedback from the Rev 0 demo to the supervisor (or to potential users) should be highlighted. —SS]

8 Automated Testing

9 Trace to Requirements

10 Trace to Requirements

10.1 Traceability Between Test Cases and Requirements

Requirement traceability from S2 of SRS to testing in this document.

Table 1: Traceability Between Functional Test Cases and
Functional Requirements, T1FR-1 to T3FR-3

Test IDs	Functional Requirement IDs									
	T1FR-1	T1FR-2	T1FR-3	T2FR-1	T2FR-2	T2FR-3	T2FR-4	T3FR-1	T3FR-2	T3FR-3
FRT-INP-1	X			X				X		
FRT-INP-2	X			X				X		
FRT-T1-1		X								
FRT-T1-2			X							
FRT-T2-1					X		X			
FRT-T2-2						X				
FRT-T2-1									X	X
FRT-T3-2									X	X
FRT-T3-3										
FRT-T3-4										
FRT-T3-5										

Table 2: Traceability Between Functional Test Cases and
Functional Requirements, T3FR-4 to T3FR-6

Test IDs	Functional Requirement IDs		
	T3FR-4	T3FR-5	T3FR-6
FRT-INP-1			
FRT-INP-2			
FRT-T1-1			
FRT-T1-2			
FRT-T1-3			
FRT-T2-1			
FRT-T2-2			
FRT-T2-1			
FRT-T3-2			
FRT-T3-3	X		
FRT-T3-4		X	
FRT-T3-5			X

Table 3: Traceability Between Non-Functional Test
Cases and Non-Functional Requirements

Test IDs	Non-Functional Requirement IDs

	GR1	GR2	SR1	SR2
NFRT-U-1	X			
NFRT-SS-1			X	
NFRT-SS-2				X
NFRT-L-1		X		

11   Traceability of Test Cases and Modules

Test IDs	Module IDs								
	M2	M3	EDM1	EDM2	EDM3	EDM4	EDM5	EDM6	EDM7
FRT-INP-1	X	X	X	X					
FRT-INP-2	X	X	X	X					
FRT-T1-1	X								
FRT-T1-2	X								
FRT-T1-3	X								
FRT-T2-1		X							
FRT-T2-2		X							
FRT-T3-1			X	X					
FRT-T3-2			X	X					
FRT-T3-3					X	X	X	X	X
FRT-T3-4					X	X	X	X	X
FRT-T3-5					X	X	X	X	X

12   Code Coverage Metrics

Appendix — Reflection

The information in this section will be used to evaluate the team members on the graduate attribute of Reflection. Please answer the following question:

1. In what ways was the Verification and Validation (VnV) Plan different from the activities that were actually conducted for VnV? If there were differences, what changes required the modification in the plan? Why did these changes occur? Would you be able to anticipate these changes in future projects? If there weren't any differences, how was your team able to clearly predict a feasible amount of effort and the right tasks needed to build the evidence that demonstrates the required quality? (It is expected that most teams will have had to deviate from their original VnV Plan.)