

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

1	Revision History	i
2	Symbols, Abbreviations and Acronyms	ii
3	Functional Requirements Evaluation	1
3.1	Task 1	1
3.2	Task 2	2
3.3	Task 3	3
4	Nonfunctional Requirements Evaluation	3
4.1	Usability Requirements	3
4.2	Safety and Security Requirements	4
4.3	Legal Requirements	5
5	Comparison to Existing Implementation	5
6	Unit Testing	5
7	Changes Due to Testing	5
8	Automated Testing	6
9	Trace to Requirements	6
10	Trace to Modules	6
11	Code Coverage Metrics	6

List of Tables

List of Figures

This document ...

3 Functional Requirements Evaluation

3.1 Task 1

1. FRT-T1-1

Control: Automatic

Initial State: No models have been trained by the task 1 system yet, input data as jsonl files are available

Input: The Task 1 Input data

Output: The system will extract required sentences from the jsonl files and output data that is appropriate to be fed into Task 1.

Test Case Derivation: The system will receive its input data, process then parse according to the Task's input format.

How test will be performed: The tester will run the parser with the given jsonl files as input and compare the parsed data.

Results: Passed

2. FRT-T1-2

Control: Automatic

Initial State: No models have been trained by the task 1 system yet, past years input data, training data, and golden truth values are available

Input: The Task 1 Training Data, Input data, and Golden Truth Values from a past year eRisk Competition

Output: The system will output sentence ranking and diagnostic metrics based on the input data.

Test Case Derivation: The system will receive its training data and input data corresponding to a prior year. It's analysis will output RP and NDCG accuracy metrics.

How test will be performed: The system should derive its training data and feature sets based on correlation to 21 depression symptoms of beck

depression index. The outputted metrics will be compared to golden truth data to ensure the accuracy is within the desired bounds and is improving on prior year implementation.

Results: Passed

3. **FRT-T1-3**

Control: Automatic

Initial State: No models have been trained by the task 1 system yet, input data is available

Input: The Task 1 Input data

Output: The system will output sentence ranking and diagnostic metrics based on the input data to a txt file.

Test Case Derivation: The system will make an analysis based off the provided input data and return each prediction to a test file in a specified formatting.

How test will be performed: The system's output txt file will be verified that each entry is on its own line in the format "{symptom_number}, Q0, {sentence-id}, {position_in_ranking}, {score}, {system_name}".

Results: Passed

3.2 **Task 2**

1. **FRT-T2-1**

Control: Automatic

Initial State: No models have been trained by the task 2 system yet, input data is available

Input: The Task 2 Input data consisting of user posts

Output: The system will output sentence ranking and diagnostic metrics based on the input data to a txt file.

Test Case Derivation: The system will make an analysis based off the provided input data and return each prediction to a test file in a specified formatting.

How test will be performed: The system's output txt file will be verified that each entry is on its own line in the format "{Username}{prediction}".

Results: Passed

2. **FRT-T2-2** Control: Automatic

Initial State: The task 2 system has received input data on a given individual and made a corresponding prediction.

Input: More data corresponding to the same user analyzed in the input state.

Output: The system will output an updated prediction taking into account all data provided.

Test Case Derivation: The system should create a new prediction for the chosen user taking into account all data provided.

How test will be performed: The system should provide a prediction based off the original data input, after receiving new data, the system should combine posts made by the same user to create a new prediction.

Results: Passed

3.3 Task 3

1.

4 Nonfunctional Requirements Evaluation

4.1 Usability Requirements

1. **NFRT-U-1**

Type: Dynamic, Manual

Initial State: System is completed and trained

Input/Condition: The system will be ran on a Windows desktop/laptop and macOS desktop/laptop device with the correct environment setup

Output/Result: The system should generate an expected result

How test will be performed: After setting up the environment on to both macOS and Windows machines, the system will be ran on both environment with the same command. Test will pass as long as an expected result is generated from both machines.

Results: Passed

4.2 Safety and Security Requirements

1. NFRT-SS-1

Type: Dynamic, Manual

Initial State: The system is completed and trained

Input/Condition: The system will be ran on the developer's computer

Output/Result: The generated result should not overuse the processing power of the developer's computer

How test will be performed: CPU, GPU, RAM and live power usage will be monitored while running the software. Usage of all aspects should not increase more than 20 percent.

Results: Passed

1. NFRT-SS-2

Type: Automatic, Dynamic

Initial State: The system has been trained and is awaiting data to form predictions on.

Input/Condition: A set of data that the system can predict on.

Output/Result: The resulting predictions, in a form where no sensitive data from the input is present in the output.

Test Case Derivation: Sensitive in this case refers to the post history of the subjects eRisk provides as training and test data. If these posts can be reconstructed from any part of our system outputs it is possible the identity of the individual could be discovered. This represents an unacceptable breach of privacy and can not happen.

How test will be performed: After the system has produced it's predictions a script will be run that takes all posts in the input data, forms a string out of all consecutive three word triplets (ie. "hello my

good friend” forms ”hello my good” and ”my good friend”), both with and without processing (removal of stopwords, punctuation, etc.), and scans the output to ensure that none of these triples are present.

Results: Passed

4.3 Legal Requirements

1. NFRT-L-1

Type: Static, Manual

Initial State: The source code and documentation is prepared

Input/Condition: The user reviews the entirety of the source code and related documentation

Output/Result: Copyright licenses, appropriate credits and/or MIT license must attached

How test will be performed: The testers will review the entirety of the project and check to see if appropriate copyright licenses and/or credits are included for resources that require them

Results: Passed

5 Comparison to Existing Implementation

This section will not be appropriate for every project.

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 Modules
- 11 Code Coverage Metrics
- References

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.)

When the team originally created the VnVPlan, our understanding of our project and the eRisk competition requirements were much more lacking at the time and as a result we have had to change a lot of our requirements as well as the tests for those requirements to fit our project. Originally our understanding of the fundamental structure of the NLP pipeline lacked depth and as such our tests did not make much sense after we completed In future projects