Automatic Psychological Text Analysis

Classifying unstructured text into corresponding schemas using Support Vector Machine (SVM)

Jeongwoo Park (j.park-3@student.tudelft.nl)
Professor: Willem-Paul Brinkman
Supervisor: Merijn Bruijnes
CSE3000 01-07-2021
TU Delft

1 BACKGROUND

- **Schema is a emotional and cognitive pattern [1].**
- **SVM is a supervised machine learning technique and it is used for classification.**
- **Better Text Analysis algorithm required for the Chatbot!**
- **Data set: 67 Schema Mode Inventory Questionnaire per story**
  - **Questionnaire reflects 3 weeks of the patient**

2 QUESTION

How well can a schema be automatically classified from a text using SVM?

- Kernel function
- Text Transformation
- Comparison with kNN and RNN

3 METHODS

DATA PREPROCESSING

- Cleaning data
  Remove noninformative data, Remove stopwords , Lower case, Expand contractions, Tokenization, Labelling dataset
- Word embedding (fastText)

CLASSIFICATION

- SVM classification using Scikit-learn
- Binary classification model for each schema
  - ["*vulnerable*", "*angry*", "*impulsive*", "*happy*", "*detached*", "*punishing*", "*healthy*"]
- Kernel functions: RBF, Linear, Polynomial

EVALUATION

- Compare result between different kernels
- Compare result between different data preprocessing technique
- Compare result between different classifications

4 RESULT

Binary Classification

Output Label: Binary Label calculated based on Allaart's Criteria. (True/False) for each schema

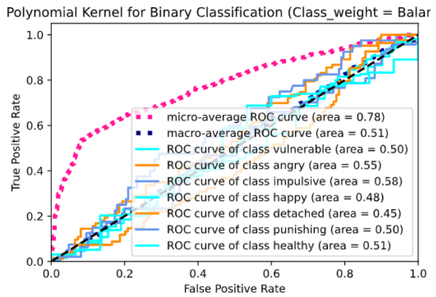
Result

- According to **F1-Score**, Polynomial is the best numerically (0.579)
- According to **Accuracy**, average of the accuracy per schema is 66.8%.
- According to **Classification Report**: Happy, Healthy are well classified.
- According to **ROC**, Impulsive has the highest AUC.

Classification Report of Polynomial

Class	Precision	Recall	F-score
vulnerable	0.27	0.12	0.17
angry	0.38	0.75	0.5
impulsive	0.07	0.03	0.04
happy	0.75	0.89	0.82
detached	0.18	0.07	0.1
punishing	0.2	0.09	0.12
healthy	0.93	0.98	0.95
micro avg	0.63	0.63	0.63
macro avg	0.4	0.42	0.39
weighted avg	0.57	0.63	0.58
samples avg	0.66	0.71	0.64

ROC curve of Polynomial



Ordinal Classification

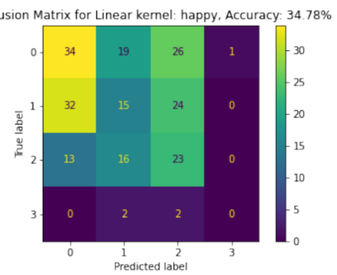
Output Label: Ordinal label calculated by mapping average of questionnaire to (0 - 3)

Result

- According to **Performance metric** [2] , Linear kernel gives the highest performance
- According to **Spearman Correlation**, Happy is the most well classified schema
- Low positive correlation

Performance metric

Kernel	Perf
RBF	0.005957
Linear	0.014437
Polynomial	0.004598



Spearman Correlation

Schema	Spearman
Vulnerable	0.078488
Angry	0.023733
Impulsive	0.003271
happy	0.123908
detached	-0.089540
Punishing	0.073704
Healthy	0.019707

Per Questionnaire Classification

Output Label: 67 questionnaire values

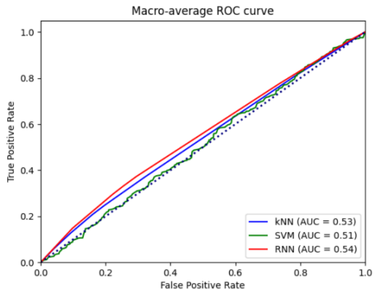
Result

- According to **F1-Score**, RBF is the highest.
- According to **Classification Report**,
  - Happy and Healthy are well-classified
  - Overall high Recall value

Classification Report of RBF

Class	Precision	Recall	F-score
vulnerable	0.34	0.83	0.48
angry	0.39	0.76	0.51
impulsive	0.2	0.62	0.3
happy	0.75	0.88	0.81
detached	0.28	0.75	0.41
punishing	0.19	0.53	0.28
healthy	0.92	0.93	0.93
micro avg	0.48	0.82	0.6
macro avg	0.44	0.76	0.53
weighted avg	0.6	0.82	0.67
samples avg	0.51	0.85	0.59

Comparison with RNN and kNN



Schema	SVM	kNN	RNN
Vulnerable	0.27	0.34	0.38
Angry	0.38	0.40	0.48
Impulsive	0.07	0.13	0.17
happy	0.75	0.80	0.77
detached	0.18	0.35	0.36
Punishing	0.2	0.22	0.34
healthy	0.93	0.96	0.95

Schema	SVM	kNN	RNN
Vulnerable	0.078	0.13	0.28
Angry	0.023	0.08	0.18
Impulsive	0.0033	0.12	0.042
happy	0.12	0.06	-0.057
detached	-0.090	0.08	0.24
Punishing	0.074	0.09	0.27
Healthy	0.020	0.06	0.09

5 DISCUSSION

- **Imbalanced data set**
- **Incorrect labelling due to SMI questionnaire which reflects 3 weeks of patient's mental state.**
- **Small size of the data set**

6 CONCLUSION

- **RNN is the best classifier in this experiment**
- **One specific best kernel does not exist.**
- **RBF was always the first or the second rank.**

Future work

- **Research on the current data set**
- **Specific labelling**
- **Collecting more data**

[1]] J. Lobbestael, M. v. Vreeswijk, P. Spinhoven, E. Schouten, and A. Arntz, "Reliability and Validity of the Short Schema Mode Inventory (SMI)," Behavioural and Cognitive Psychotherapy, vol. 38, no. 4, pp. 437– 458, 2010. Publisher: Cambridge University Press
[2] Burger Franziska. Natural language processing for cognitive therapy: extracting schemas from thought records. 2021.