Exploiting Textual and Numerical Data to Improve Predictive Models for Depression

Academy Assistants: Plamen Dimitrov, Reinier Kop
Supervisors: Mark Hoogendoorn, Michel Klein, Heleen Riper, Piek Vossen

A large body of text collected from an Internet-based treatment for adults with depressive symptoms is evaluated and assigned CES-D (depression) scores.

- SVM (support vector machines)
- Naïve Bayes
- Dutch stopwords and stemmers
- word presence vs. frequency

**Derive Concepts from Numerical Survey Data Set**

**Tune model parameters**
- Compare three methods
  1. Initial learning period: Train on pre-intervention week only
  2. Incorporate both weeks
  3. Nearest neighbour: single parameter set per subject

**Future Work**

1. Applying classifier to general text. Using frequency profiling to compare body of text to other corpora. Use improved text classifier to produce input for the mood prediction model.
2. Deeper investigation regarding individualizing parameters for this specific model.
3. Investigate if other/improved clustering algorithms make a difference.

**Acknowledgements**
- Numerical Data Set: Prof. Berking & Anna Radkovsky (University of Marburg)
- Textual Data Set: Lisanne Warmerdam (VU University Amsterdam)

**Evaluation and comparison of techniques**

- Some techniques tend to consistently outperform others w.r.t. accuracy: SVM produces better results than Naïve Bayes.
- Max. achieved accuracy is still under 80% (78%) - further experimentation with features and techniques is required.

**Mood prediction model**

**Validation...**

1. Incorporating post-intervention survey weeks from other historical patients is beneficial.
2. Individualizing parameter sets can be beneficial in some cases.

**Bibliography**