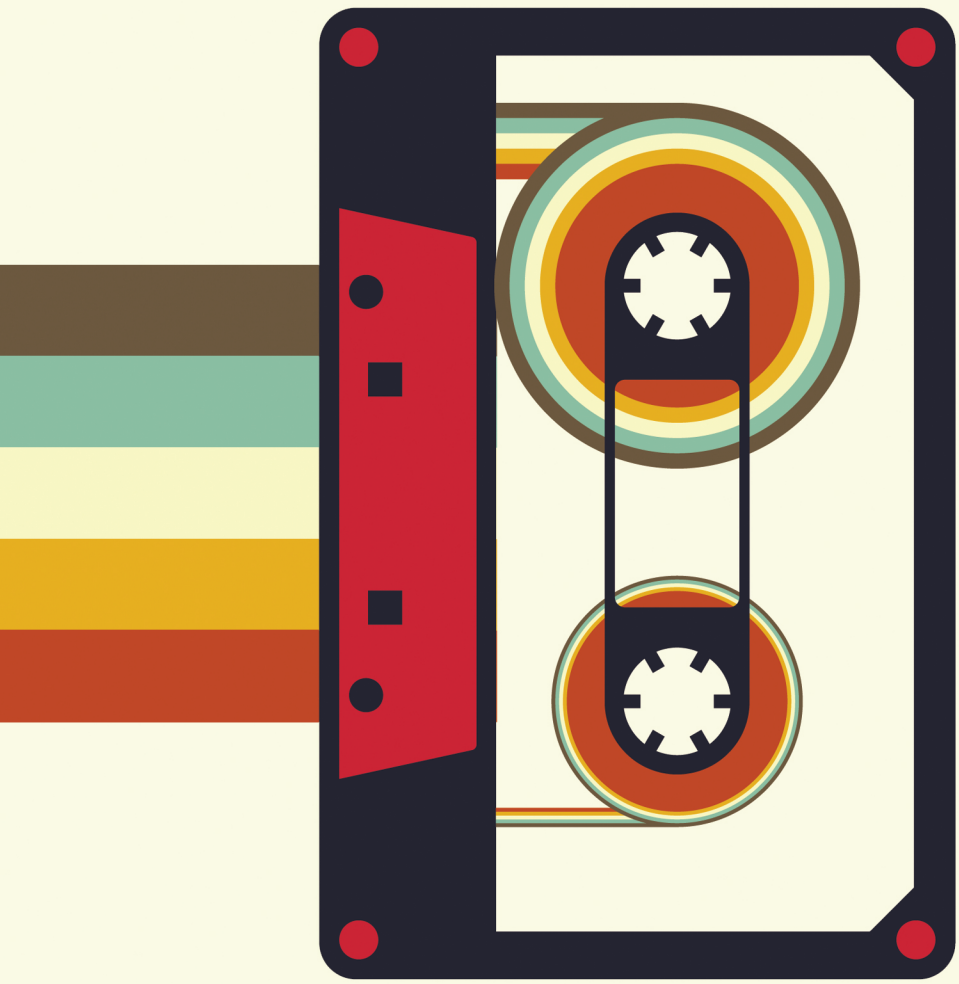
Reusing clinical data to improve health care: Challenge accepted!

S. Wiegersma



REUSING CLINICAL DATA TO IMPROVE HEALTH CARE: CHALLENGE ACCEPTED!

Sytske Wiegersma

REUSING CLINICAL DATA TO IMPROVE HEALTH CARE: CHALLENGE ACCEPTED!

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr.ir. A. Veldkamp,
on account of the decision of the Doctorate Board,
to be publicly defended
on Thursday the 17th of November 2022 at 14.45 hours

by

Sytske Wiegersma

born on the 16th of August 1988
in Smallingerland, The Netherlands

This dissertation has been approved by:

Supervisors: prof.dr.ir. B.P. Veldkamp
prof.dr. M. Olff

Cover design: Maaike Heitink
Printed by: Ipskamp printing, The Netherlands
Lay-out: Joost van Noije, K.P. Hart (TUD Dissertation template 2020)
ISBN: 978-90-365-5471-8
DOI: 10.3990/1.9789036554718

GRADUATION COMMITTEE:

Chair / secretary:	prof.dr. T. Bondarouk (University of Twente)
Supervisors:	prof.dr.ir. B.P. Veldkamp (University of Twente)
	prof.dr. M. Olff (University of Amsterdam)
Committee Members:	prof.dr. D.K.J. Heylen (University of Twente)
	dr. C.M. Hoeboer (University of Amsterdam)
	prof.dr. M.E. Iacob (University of Twente)
	prof.dr. J.A.M. Van der Palen (University of Twente)
	prof.dr. H.G.J.M. Vermetten (Leiden University)
	prof.dr. G.J. Westerhof (University of Twente)

Contents

1 General introduction **1**
1.1 Data sources . . . 2
1.1.1 Routine data . . . 3
1.1.2 Research data . . . 3
1.2 Structured versus unstructured data . . . 4
1.3 New technologies . . . 4
1.4 Critical elements and challenges for data reuse . . . 5
1.4.1 Data quality . . . 5
1.4.2 Completeness . . . 5
1.4.3 Privacy . . . 6
1.5 General aims and outline . . . 7
1.6 References . . . 9

2 Fitness for purpose of routinely recorded health data **13**
2.1 Introduction . . . 15
2.1.1 The diagnosis of Sjögren's syndrome . . . 15
2.1.2 Primary care data . . . 15
2.1.3 Secondary care data . . . 16
2.2 Methods . . . 17
2.2.1 Data sets . . . 17
2.2.2 Developing the algorithm . . . 18
2.2.3 Validating the algorithm . . . 20
2.3 Results . . . 22
2.3.1 Patient selection algorithm . . . 22
2.3.2 Algorithm validation . . . 25
2.4 Discussion . . . 30
2.5 References . . . 35
Appendix 2-A . . . 37

3 Automated supervised text cassification tool **39**
3.1 Introduction . . . 41
3.2 Methods . . . 43
3.2.1 Corpus: Letters from the Future . . . 43
3.2.2 ASTeCT . . . 44
3.2.3 Model development . . . 45
3.2.4 Text classification pipeline . . . 48
3.2.5 Model selection . . . 56
3.2.6 Model evaluation . . . 57
3.2.7 Save and apply final model . . . 57

3.3 Results . . . 58
3.3.1 Binary classifier . . . 58
3.3.2 Multiclass classifier . . . 60
3.3.3 English classifier . . . 64
3.4 Discussion . . . 64
3.5 References . . . 67

4 Improving web-based treatment intake for mental health disorders 73
4.1 Introduction . . . 75
4.2 Methods . . . 78
4.2.1 Data set . . . 78
4.2.2 Automated text screening model . . . 79
4.2.3 Analytical strategy . . . 83
4.2.4 Text classification tool . . . 84
4.3 Results . . . 84
4.3.1 Data set . . . 84
4.3.2 Screening model . . . 86
4.4 Discussion . . . 92
4.4.1 Principal results . . . 92
4.4.2 Theoretical and practical contributions . . . 93
4.4.3 Limitations . . . 95
4.4.4 Future research . . . 96
4.4.5 Conclusions . . . 97
4.5 References . . . 98
Appendix 4-A . . . 103

5 Recognizing hotspots in Brief Eclectic Psychotherapy for PTSD 107
5.1 Introduction . . . 109
5.2 Methods . . . 111
5.2.1 Sample and data set . . . 111
5.2.2 Operational constructs for automatic recognition . . . 114
5.2.3 Classification pipeline . . . 117
5.3 Results . . . 125
5.3.1 Sample characteristics . . . 125
5.3.2 Model comparison . . . 127
5.3.3 Final model . . . 128
5.4 Discussion . . . 130
5.5 References . . . 135
Appendix 5-A . . . 145
Appendix 5-B . . . 150
Appendix 5-C . . . 153

6 Exploring “Letters from the Future” by visualization 157
6.1 Introduction . . . 159
6.1.1 Letters from the Future . . . 160

6.2 Methods . . . 161
6.2.1 Data set . . . 161
6.2.2 Preprocessing . . . 161
6.2.3 Mathematical structure . . . 164
6.2.4 Analytical procedure . . . 166
6.2.5 Text visualization design . . . 166
6.3 Results . . . 167
6.3.1 Selected LIWC categories . . . 167
6.3.2 Visualizations . . . 169
6.4 Discussion . . . 174
6.5 References . . . 177

7 General discussion **181**
7.1 Main findings . . . 182
7.1.1 Fitness for purpose of routinely recorded health data . . . 182
7.1.2 Automated supervised text cassification tool . . . 183
7.1.3 Improving treatment intake for mental health disorders . . . 183
7.1.4 Recognizing hotspots in Brief Eclectic Psychotherapy . . . 184
7.1.5 Exploring “Letters from the Future” by visualization . . . 185
7.2 Interpretation of findings . . . 185
7.3 Limitations . . . 186
7.4 Future perspectives . . . 187
7.5 Conclusion . . . 188
7.6 References . . . 190

Summary **191**

Samenvatting **195**

Dankwoord **199**

Curriculum Vitæ **203**

List of Publications **205**

1

General introduction

1

Traditionally, scientific research is based on primary data specifically collected for the aim of the study through standard, validated techniques such as randomized controlled trials, experiments, questionnaires, or interviews. Primary data sets consist of information that was systematically collected from the source, to answer pre-defined research questions or test hypotheses (Hox & Boeije, 2005). According to the First Law of Medical Informatics (van der Lei, 1991), data should only be used for the purpose for which they were originally collected. However, the increasing availability of clinical data, the widespread adoption of new technologies such as data mining and machine learning, and the possibility to link and combine individual patient data from different sources, have made data reuse a rapidly growing area with high potential for clinical practice and scientific research (de Lusignan & van Weel, 2006; Kassam-Adams & Olff, 2020; Olff, 2020). Data reuse, or secondary data use, refers to the use of data for purposes other than primarily intended (Safran, 2017). Clinical data reuse is believed to potentially improve health care quality and management, reduce health care costs, and enable more effective clinical research and population health management (Meystre et al., 2017).

Clinical data collection (e.g., through trials) is time consuming and costs have been rising for decades (Berndt & Cockburn, 2013; Collier, 2009; Peek & Rodrigues, 2018). This has led researchers and informaticians to explore alternatives such as real-world, routine data available from, for example, electronic health records (EHRs) and health insurance claims databases. Such sources contain rich, often longitudinal, data that are routinely recorded at the point of care (Sherman et al., 2016). Thanks to this, routine data may also lead to results that generalize better to the general population than results based on trial data, as clinical trial conditions and participants may differ significantly from those in practice (Peek & Rodrigues, 2018; Sherman et al., 2016; Weinfurt et al., 2017).

From an academical point of view, data reuse is expected to lead to more efficient research processes, faster and more generalizable clinical evidence, improved patient identification, and more precise epidemiological estimates (Coorevits et al., 2013). The idea of data reuse is further strengthened by the active call for sustainable research. In order to promote maximum use of resources, researchers are encouraged to make use of existing data available from, for example, open patient cohorts and research data depositories and to make their own data (openly) available for reuse and for the validation and reproducibility of results. To stimulate scientific data reuse, the FAIR Data Principles were formulated and an action plan was published by the European Commission Expert Group on FAIR data in 2018 (European Commission Expert Group on FAIR Data, 2018). FAIR stands for Findability, Accessibility, Interoperability, and Reusability, and the principles emphasize the machine-driven discovery and reuse of data, as computers are increasingly used to deal with the large volumes of data becoming available.

1.1. Data sources

Major sources of clinical data are routine data recorded as part of ongoing patient care and research data specifically collected for scientific purposes.

1.1.1. Routine data

The amount of available routinely collected care data is vast and ever-growing, thanks to the use of EHRs and the availability of (linkable) administrative and health insurance claims databases (Safran, 2017). Deeny and Steventon (2015) divide routine data into administrative data, clinically generated data, and patient-generated data. Administrative data, such as diagnosis related groups (DRG) coded insurance claims data, are collected during the routine administration of delivered care. Clinically generated data, such as coded diseases and symptoms or laboratory test results, are collected by health care workers as part of the diagnosis or treatment process. Patient-generated data (either clinically or individually directed) can be patient-reported (e.g., derived from patient-reported outcome measures or patient narratives) or non-traditional self-measurement data (e.g., derived from e-health apps or wearables). The main benefits of routine data are their wide coverage, longitudinality, low self-report bias, and the simple fact that they are routinely available (van Dalen et al., 2014).

In hospitals alone, 97% of all the data produced each year remains unused (Cornell University et al., 2019). The general consensus is that this wealth of information, sometimes referred to as a "by-product of care", can and should be put to better use. In clinical practice, integrating available care data and analytics as part of the care trajectory may support medical professionals by providing insights enabling predictive, individualized, and efficient care (Cornell University et al., 2019). Clinical data can be used in the development of decision support systems, for example for the early detection of diseases or for referring patients to the most suitable treatment. Pakhomov et al. (2007) demonstrated, for example, that textual input from EHRs could be used to effectively identify patients with heart failure, whereas He et al. (2012) used narratives extracted from the open questions of an e-health survey to screen trauma survivors for PTSD. Clinical data can also be deployed to provide patient-specific, or personalized, care, for example by allowing for case or peer comparison and by gaining insight in the effectiveness of possible treatment or prescription options for similar patients or patient groups (Meystre et al., 2017; Safran, 2017). Towards improving health care quality, patterns and deviations in routine care can lead to the identification of opportunities for care improvement, for example to prevent medication errors (Spencer et al., 2014).

1.1.2. Research data

Thanks to initiatives such as Research Data Netherlands (https://researchdata.nl/, a Dutch coalition of data archives promoting the sustainable archiving and reuse of research data) and Narcis (https://www.narcis.nl/, the research data portal from Dutch universities), Dutch researchers are encouraged to secure, share, and reuse research data. Global examples are DataCite (https://datacite.org/), a global non-profit organization providing DOIs (persistent identifiers) for research data, and DataCite's service Re3data, listing over 2,000 global research data depositories.

Health research data are generally collected through health surveys, clinical trials, cohort studies, or patient-reported outcomes or experience measures. These data are often well-defined and highly structured. In mental health research, pa-

tient narratives or treatment records form an additional, rich source of information. However, such data are more difficult to reuse and share due to their unstructured format and privacy sensitive contents.

1.2. Structured versus unstructured data

Structured data are considered the most convenient for secondary purposes in direct care, such as offering decision-support (Meystre et al., 2017) or assessing compliance to medical guidelines (Vuokko et al., 2015). Care data can be structured using classification systems or standardized forms, thanks to which structured patient data offer more uniform documentation and higher quality, more complete, information (Vuokko et al., 2015). Commonly used classification systems in medicine are, for example, the International Classification of Diseases (ICD; World Health Organization, 2004), the Diagnosis-Related Groups classification system for hospital reimbursements (DRG; Hasaart, 2011), the International Classification of Primary Care (ICPC; Lamberts & Wood, 1987), and the Diagnostic and Statistical Manual for Mental Disorders (DSM; American Psychiatric Association, 2013). Such controlled terminologies allow for the exchange and comparison of health care data across different care settings and systems (Meystre et al., 2017).

Despite the advantages of structured data, the majority of clinical data consists of unstructured text. A study on US hospital EHR data revealed that only a third of the data was stored in a structured format, versus two-thirds of unstructured text (Cannon & Lucci, 2010). Apart from the formal classification systems, all kinds of useful information and medical concepts can be encoded in and extracted from unstructured data. For example, annotating the occurrence and frequency of key-moments or breakthroughs in psychiatric treatments (e.g., Nijdam et al., 2013), or distinguishing different types of narrative processing (e.g., Sools et al., 2015), results in new, structured data sets based on which new knowledge and insights can be generated. In order to structure and process the large volumes of data becoming available and convert those into useful information, new technologies such as artificial intelligence (AI) are increasingly used (Cornell University et al., 2019).

1.3. New technologies

To efficiently extract information stored in free text, unstructured text data can be processed automatically using natural language processing (NLP) and text mining (TM), converting the data to a more structured format (Meystre et al., 2017; Vuokko et al., 2015). NLP is typically conducted as a first step to clean and preprocess the data and convert it to structured information. NLP applications can be as simple as searching text data for a list of pre-specified key words and calculating the frequency or proportion of occurrence of each key word. Speech data can be processed on the level of textual content or using speech features that capture one's manure of speaking. To be able to process the contents, speech is first converted to text, ideally using automatic speech recognition (ASR).

The features extracted from text and speech data can then be used as input

for (text) data mining and machine learning (ML) algorithms (see Chapter 3 for an extensive description of NLP, TM, and ML). A popular ML approach to organize data is supervised classification, which involves assigning objects to a set of predefined class labels using a classification model trained on existing, labeled data (Bird et al., 2009). The use of an automated classifier not only lowers the cost of manual annotation (Sebastiani, 2002) but was also found to result in a more reliable and precise extraction of clinical information than manual classification (Friedman et al., 2004). Moreover, modern techniques enable the extraction of information not easily processed or noticed by human encoders, and certainly not on a large scale, such as speech characteristics (e.g., speech rate, pitch, or quality, see more in Chapter 5).

1.4. Critical elements and challenges for data reuse

All in all, data reuse is expected to lead to improved efficiency, quality, and effectiveness of both clinical practice and health care research, and to assist in the discovery of new knowledge (Safran, 2017). However, when reusing (routine) data for research, one should keep in mind that the data were often not originally collected, organized, and optimized for the intended research purposes. This may lead to uncertainties and challenges regarding data quality, completeness, and privacy (Sherman et al., 2016).

1.4.1. Data quality

Data quality is a critical element in secondary data use. Much has been written about the quality of routine data. Overhage and Overhage (2013) described a range of intentional and unintentional issues occurring in routine clinical data, such as inaccuracies due to efforts to maximize reimbursement (also termed 'upcoding'; Verheij et al., 2018), due to the requirement of specific diagnostic codes to motivate the use of certain tests or procedures, or due to simple data entry errors. Consequently, diagnostic and procedure codes are often biased, underspecified, and lack the detailed and accurate information needed by, for example, scientists, policymakers, and clinicians (Meystre et al., 2017).

Data quality can also be problematic in the literal sense, for example the quality of audio and video recordings initially made for research or monitoring purposes. Especially when reusing old, analog recordings, sound or image quality may be poor due to the use of basic recording equipment or may diminish over the years due to the transitory nature of analog material. Recording quality may have been of less importance for the primary purpose, while essential for the secondary purpose. In such cases, new technologies such as ASR might not be applicable to convert the data, forcing data reusers to return to traditional methods such as manual transcription. This is illustrated in Chapter 5.

1.4.2. Completeness

Clinical information is documented by health professionals for clinical use (e.g., to track the patient's conditions and to inform each other) or billing purposes

(Meystre et al., 2017). Consequently, data are mainly recorded from the clinician's perspective instead of the patient's perspective (Deeny & Steventon, 2015) and seem to lack both the detail and the outcome measures needed for effective research (Meystre et al., 2017; Safran, 2017). Research data, for example collected through randomized controlled trials, give a more complete and balanced overview of a patient's characteristics (Overhage & Overhage, 2013; Peek & Rodrigues, 2018). However, the scope of such trials is often limited, focusing on a small selection of patients treated in specific care settings, for a confined time period.

As stated by Vuokko et al. (2015), successful secondary use requires complete and interoperable patient records. A way to increase the scope and completeness of patient information is to combine routine or research data with supplementary care data available from other sources within and outside the care system. For example, in-depth EHR data can increasingly be linked to more superficial longitudinal claims databases, which provide additional information regarding diagnoses, specialized treatment, or clinical outcomes (Lin & Schneeweiss, 2016; Safran, 2017). Similarly, survey or clinical trial data may be supplemented with EHR data to gain more insight in the care trajectories or diagnostic profiles of patients, and to study differences between patient groups. As such, linking data helps to fill the gaps and increase the scope of the original data set (Weber et al., 2014), as is shown in Chapter 2.

1.4.3. Privacy

Finally, privacy and security are dominant aspects when reusing health data (Safran, 2017). Especially when data derived from multiple sources are linked, de-identification or anonymization becomes increasingly complex (Weber et al., 2014). This is even more challenging when working with text data or audio or video recordings, which are difficulty to de-identify without removing or distorting valuable information such as voices, facial expressions, or body language.

Modern techniques based on NLP and ASR are successfully applied in text and audio de-identification tools (e.g., Kayaalp et al., 2015; Cohn et al., 2019). However, de-identification is not the same as anonymization, and with text or audio data the risk and consequences of re-identification are high (Meystre et al., 2017). When working with sensitive clinical information, such as patient narratives or clinical notes, it therefore could be sensible to do a 'blind' analysis, i.e., analyze the data on location without having insight in the actual contents, as is done in Chapter 4. A new initiative which promotes local data reuse is the Personal Health Train (PHT; Deist et al., 2020) by the Dutch Health Research Infrastructure (HealthRI, https://www.health-ri.nl/). The PHT enables researchers to work on sensitive health data from various sources without the need to transport and centralize the data into one database. Instead, data analysis takes place locally at the source, which makes reuse of privacy sensitive data safer. This is especially useful when working with data that are hard to anonymize, such as the mental health data used in this thesis.

1.5. General aims and outline

All in all, expectations of data reuse and its effect on scientific research and clinical practice are high, and much has been written on the topic. However, most research on clinical data reuse is found to report on how reuse is supposed to impact health care (e.g., care processes, quality, and outcomes) instead of truly realizing and demonstrating data reuse and its (dis)advantages in practice (Vuokko et al., 2015). Data quality, completeness, and privacy are critical and often challenging elements when reusing data. New technologies such as AI may provide solutions for these challenges, for example in extracting and encoding information from unstructured data sets, enriching limited data sets by linking them to data from other sources, and processing privacy sensitive narratives. However, successful application and adoption of such technologies depends on their availability and usability for care providers and researchers without a technical background. A major goal to be achieved is to provide clinicians with efficient, intuitive tools to support their own clinical research and to do research in real care settings (Meystre et al., 2017).

The overall aim of this thesis was to investigate how new technologies such as AI can contribute to the successful reuse of clinical data towards improving (mental) health care practice and research. This thesis gives a broad overview of clinical data reuse in practice, illustrating the challenges encountered and solutions available when reusing existing care data sets for secondary purposes. Data reuse is demonstrated using both routine care and research data sets. The data sets originate from different care settings and different phases in the care process, ranging from routinely recorded general practitioner (GP) visits, online intake questionnaires for patient referral in mental health care, face-to-face specialized patient treatment sessions, and administrative hospital claims records. Data formats range from coded diagnoses to patient narratives and audio recordings of therapy sessions. The challenges encountered when reusing these data include poor data quality, lack of sufficiently detailed information, missing outcome scores, and patient privacy. We made use of new (AI) technologies such as machine learning, text and audio mining, and data linkage to deal with these challenges. In other cases, we returned to traditional methods such as the manual transcription of low-quality audio recordings instead of using ASR.

The first study presented in this thesis (Chapter 2) examines the usability (fitness for purpose) of routinely recorded care data to identify patients with complex diseases and to estimate the prevalence of such diseases in the general population. This study describes the development and validation of a patient selection algorithm using ICPC coded GP contacts and disease episodes, combined with a simple keyword search in the disease episode titles included in primary care EHRs. As the primary care data lacked the diagnostic outcome data needed to validate the selection algorithm, the data set was enriched with outcome data from secondary care by linking the primary care data to a nation-wide hospital claims database covering DRG coded and ICD-10 coded diagnoses.

Chapter 3 provides an elaborate description of supervised text classification and the value of this popular text mining technique for (psychological) research. To make this method available for researchers and care professionals with little to no

experience in computer science, statistical modeling, or programming, an Automated Supervised Text Classification Tool (ASTeCT) was developed and tested by reusing narrative data collected through a health promotion instrument previously used for a psychological study. This tool enables users to easily and safely apply supervised text classification directly to their own text data set and generate their own classification models.

In Chapter 4 we use the supervised text classification tool introduced in Chapter 3 to automatically screen for multiple mental and substance use disorders using the textual responses on a patient intake questionnaire. Outcomes could be used to support care providers in the intake process and refer patients to the most suitable treatment. To deal with privacy issues associated with patient-written narratives, the tool was run blind, in the local environment of the data owner offering web-based treatment.

Chapter 5 illustrates the development of a multimodal (text and audio) supervised classification model based on existing, hand-coded, clinical trial data. The aim of this study was to develop a model to automatically recognize hotspots (key elements of the used exposure therapy) based on text and speech features, which might be an efficient way to track patient progress and predict treatment efficacy. This study shows the challenges and lessons learned when reusing audio data of poor recording quality.

In Chapter 6, we explore how visualization of outcomes of the standardized NLP tool LIWC (Linguistic Inquiry and Word Count) can be used to explore differences in narrative styles, reusing the same psychological research data as in Chapter 3. This study shows how visualizations can lead to additional information on existing data sets and provide directions for future research in both the visualization of narrative structure and the field of narrative psychology.

Finally, Chapter 7 provides a summary and integrated discussion of the findings of this thesis.

1.6. References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*.

Berndt, E. R., & Cockburn, I. M. (2013). *Price indexes for clinical trial research: A feasibility study* (tech. rep.). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w18918

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'reilly Media, Inc.

Cannon, J., & Lucci, S. (2010). Transcription and EHRs: Benefits of a Blended Approach. *The American Health Information Management Association*.

Cohn, I., Laish, I., Beryozkin, G., Li, G., Shafran, I., Szpektor, I., Hartman, T., Hassidim, A., & Matias, Y. (2019). Audio de-identification: A new entity recognition task. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *2*, 197–204. https://doi.org/10.18653/v1/n19-2025

Collier, R. (2009). Rapidly rising clinical trial costs worry researchers. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, *180*(3), 277–278. https://doi.org/10.1503/cmaj.082041

Coorevits, P., Sundgren, M., Klein, G. O., Bahr, A., Claerhout, B., Daniel, C., Dugas, M., Dupont, D., Schmidt, A., Singleton, P., De Moor, G., & Kalra, D. (2013). Electronic health records: New opportunities for clinical research. *Journal of Internal Medicine*, *274*(6), 547–560. https://doi.org/10.1111/joim.12119

Cornell University, INSEAD, & WIPO. (2019). *Global Innovation Index 2019: Creating Healthy Lives - The Future of Medical Innovation*. WIPO.

Deeny, S. R., & Steventon, A. (2015). Making sense of the shadows: Priorities for creating a learning healthcare system based on routinely collected data. *BMJ Quality and Safety*, *24*(8), 505–515. https://doi.org/10.1136/bmjqs-2015-004278

Deist, T. M., Dankers, F. J., Ojha, P., Scott Marshall, M., Janssen, T., Faivre-Finn, C., Masciocchi, C., Valentini, V., Wang, J., Chen, J., Zhang, Z., Spezi, E., Button, M., Jan Nuyttens, J., Vernhout, R., van Soest, J., Jochems, A., Monshouwer, R., Bussink, J., … Dekker, A. (2020). Distributed learning on 20 000+ lung cancer patients – The Personal Health Train. *Radiotherapy and Oncology*, *144*, 189–200. https://doi.org/10.1016/j.radonc.2019.11.019

de Lusignan, S., & van Weel, C. (2006). The use of routinely collected computer data for research in primary care: Opportunities and challenges. *Family Practice*, *23*(2), 253–263. https://doi.org/10.1093/fampra/cmi106

European Commission Expert Group on FAIR Data. (2018). *Turning FAIR into reality*. https://doi.org/10.2777/1524

Friedman, C., Shagina, L., Lussier, Y., & Hripcsak, G. (2004). Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, *11*(5), 392–402. https://doi.org/10.1197/jamia.M1552

Hasaart, F. (2011). *Incentives in the diagnosis treatment combination payment system for specialist medical care: A study about the behavioral responses of medical specialists and hospitals in the netherlands*. Maastricht University. https://doi.org/10.26481/dis.20111104fh

He, Q., Veldkamp, B., & De Vries, T. (2012). Screening for posttraumatic stress disorder using verbal features in self narratives: A text mining approach. *Psychiatry Research*, *198*(3), 441–447. https://doi.org/10.1016/j.psychres.2012.01.032

Hox, J. J., & Boeije, H. R. (2005). Data Collection, Primary vs. Secondary. https://doi.org/10.1016/B0-12-369398-5/00041-4

Kassam-Adams, N., & Olff, M. (2020). Embracing data preservation, sharing, and reuse in traumatic stress research. *European Journal of Psychotraumatology*, *11*(1). https://doi.org/10.1080/20008198.2020.1739885

Kayaalp, M., Browne, A. C., Dodd, Z. A., Sagan, P., & McDonald, C. J. (2015). An Easy-to-Use Clinical Text De-identification Tool for Clinical Scientists: NLM Scrubber. *AMIA 2015 Annual Symposium*, 1522.

Lamberts, H., & Wood, M. (1987). *ICPC, international classification of primary care*. Oxford University Press.

Lin, K. J., & Schneeweiss, S. (2016). Considerations for the analysis of longitudinal electronic health records linked to claims data to study the effectiveness and safety of drugs. *Clinical Pharmacology and Therapeutics*, *100*(2), 147–159. https://doi.org/10.1002/cpt.359

Meystre, S. M., Lovis, C., Bürkle, T., Tognola, G., Budrionis, A., & Lehmann, C. U. (2017). Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. *Yearbook of Medical Informatics*, *26*(1), 38–52. https://doi.org/10.15265/IY-2017-007

Nijdam, M., Baas, M., Olff, M., & Gersons, B. (2013). Hotspots in trauma memories and their relationship to successful trauma-focused psychotherapy: A pilot study. *Journal of Traumatic Stress*, *26*, 38–44. https://doi.org/10.1002/jts.21771

Olff, M. (2020). To share or not to share –10 years of European Journal of Psychotraumatology. *European Journal of Psychotraumatology*, *11*(1). https://doi.org/10.1080/20008198.2020.1844955

Overhage, J. M., & Overhage, L. M. (2013). Sensible use of observational clinical data. *Statistical Methods in Medical Research*, *22*(1), 7–13. https://doi.org/10.1177/0962280211403598

Pakhomov, S., Weston, S. A., Jacobsen, S. J., Chute, C. G., Meverden, R., & Roger, V. L. (2007). Electronic Medical Records for Clinical Research: Application to the Identification of Heart Failure. *The American Journal of Managed Care*, *13*(1), 281–288.

Peek, N., & Rodrigues, P. P. (2018). Three controversies in health data science. *International Journal of Data Science and Analytics*, *6*(3), 261–269.

Safran, C. (2017). Update on Data Reuse in Health Care. *Yearbook of medical informatics*, *26*(1), 24–27. https://doi.org/10.15265/IY-2017-013

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. https://doi.org/10.1145/505282.505283

Sherman, R. E., Anderson, S. A., Dal Pan, G. J., Gray, G. W., Gross, T., Hunter, N. L., LaVange, L., Marinac-Dabic, D., Marks, P. W., Robb, M. A., Shuren, J., Temple, R., Woodcock, J., Yue, L. Q., & Califf, R. M. (2016). Real-World Evidence — What is it and what can it tell us? *New England Journal of Medicine*, *375*(23), 2293–2297. https://doi.org/10.1056/nejmsb1609216

Sools, A. M., Tromp, T., & Mooren, J. H. (2015). Mapping letters from the future: Exploring narrative processes of imagining the future. *Journal of Health Psychology*, *20*(3), 350–364. https://doi.org/10.1177/1359105314566607

Spencer, R., Bell, B., Avery, A. J., Gookey, G., & Campbell, S. M. (2014). Identification of an updated set of prescribing-safety indicators for GPs. *British Journal of General Practice*, *64*(621), 181–190. https://doi.org/10.3399/bjgp14X677806

van der Lei, J. (1991). Use and abuse of computer-stored medical records. *Methods of Information in Medicine*, *30*(02), 79–80.

van Dalen, M. T., Suijker, J. J., MacNeil-Vroomen, J., van Rijn, M., Moll van Charante, E. P., de Rooij, S. E., & Buurman, B. M. (2014). Self-report of healthcare utilization among community-dwelling older persons: A prospective cohort study. *PloS one*, *9*(4), e93372.

Verheij, R. A., Curcin, V., Delaney, B. C., & McGilchrist, M. M. (2018). Possible sources of bias in primary care electronic health record data use and reuse. *Journal of Medical Internet Research*, *20*(5), e9134. https://doi.org/10.2196/jmir.9134

Vuokko, R., Mäkelä-Bengs, P., Hyppönen, H., & Doupi, P. (2015). Secondary Use of Structured Patient Data: Interim Results of A Systematic Review. *Studies in Health Technology and Informatics*, *210*, 291–295. https://doi.org/10.3233/978-1-61499-512-8-291

Weber, G. M., Mandl, K. D., & Kohane, I. S. (2014). Finding the missing link for big biomedical data. *JAMA - Journal of the American Medical Association*, *311*(24), 2479–2480. https://doi.org/10.1001/jama.2014.4228

Weinfurt, K. P., Hernandez, A. F., Coronado, G. D., DeBar, L. L., Dember, L. M., Green, B. B., Heagerty, P. J., Huang, S. S., James, K. T., Jarvik, J. G., Larson, E. B., Mor, V., Platt, R., Rosenthal, G. E., Septimus, E. J., Simon, G. E., Staman, K. L., Sugarman, J., Vazquez, M., … Curtis, L. H. (2017). Pragmatic clinical trials embedded in healthcare systems: Generalizable lessons from the NIH Collaboratory. *BMC Medical Research Methodology*, *17*(1), 1–10. https://doi.org/10.1186/s12874-017-0420-7

World Health Organization. (2004). *The international statistical classification of diseases and health related problems ICD-10: Tenth revision. volume 1: Tabular list* (Vol. 1).

2

Fitness for purpose of routinely recorded health data to identify patients with complex diseases: The case of Sjögren's syndrome

This chapter was published as: Wiegersma, S., Flinterman, L.E., Seghieri, C., Baldini, C., Paget, J., Barrio Cortés, J., & Verheij, R.A. (2020). Fitness for purpose of routinely recorded health data to identify patients with complex diseases: The case of Sjögren's syndrome. *Learning Health Systems, 4*(4), e10242. https://doi.org/10.1002/lrh2.10242

Abstract

This study assesses the usability (fitness for purpose) of routinely recorded primary care and hospital claims data for the identification and validation of patients with complex diseases such as primary Sjögren's syndrome (pSS). pSS is an underdiagnosed, long-term autoimmune disease that affects particularly salivary and lachrymal glands. We identified pSS patients in primary care by translating the formal inclusion and exclusion criteria for pSS from secondary care into a patient selection algorithm using data from Nivel Primary Care Database (PCD), which covers 10% of the Dutch population between 2006 and 2017. As part of a validation exercise, the pSS patients found by the algorithm were compared to Diagnosis Related Groups (DRGs) recorded in the national hospital insurance claims database (DIS) between 2013 and 2017. International Classification of Primary Care (ICPC) coded general practitioner (GP) contacts combined with the mention of "Sjögren" in the disease episode titles, were found to best convert the formal classification criteria to a selection algorithm for pSS. A total of 1,462 possible pSS patients were identified in primary care (mean prevalence 0.7‰, against 0.61‰ reported globally). The DIS contained 208,545 patients with a Sjögren related DRG or ICD-10 code (prevalence 2017: 2.73‰). A total of 2,577,577 patients from Nivel PCD could be linked to the DIS database. In total, 716 of the linked pSS patients (55.3%) were confirmed based on the DIS. Our study found that GP electronic health records (EHRs) lack the granular information needed to apply the formal diagnostic criteria for pSS. The developed algorithm resulted in a patient selection that approximated the expected prevalence and patient characteristics, although only slightly over half of the possible pSS patients were confirmed using the DIS. Without more detailed diagnostic information, the fitness for purpose of routine EHR data for patient identification and validation could not be determined.

2.1. Introduction

2.1.1. The diagnosis of Sjögren's syndrome

Primary Sjögren's syndrome (pSS) is an underdiagnosed, long-term autoimmune disease that affects particularly salivary and lachrymal glands but that may involve any organ and system (Daniels & Fox, 1992). Despite generally benign, pSS may be characterized by severe rare complications including non-Hodgkin's lymphoma (NHL) with an unneglectable impact on patients' quality of life (Brito-Zeron et al., 2016; Cafaro et al., 2019) To date, health policy and management research for pSS are quite rare, especially on pSS diagnosis and management in primary health care (Seghieri et al., 2019).

A study on the epidemiology of Sjögren's syndrome by Patel and Shahane (2014) concluded that: *"there is no accepted universal classification criterion for the diagnosis of Sjögren's syndrome. There are a limited number of studies that have been published on the epidemiology of Sjögren's syndrome, and the incidence and prevalence of the disease varies according to the classification criteria used. The data is further confounded by selection bias and misclassification bias, making it difficult for interpretation."* [p.247]. In fact, international consensus on the classification criteria for pSS was only reached in 2016, resulting in the American College of Rheumatology/European League Against Rheumatism (ACR/EULAR) classification criteria for pSS (Shiboski et al., 2017), making it difficult to estimate the exact prevalence of the disease. Consequently, estimates of the prevalence of pSS vary greatly across studies (ranging from 0.11‰ to 37.9‰), depending on the setting and the definition used and the population investigated (Qin et al., 2015).

Besides population, geographical, and diagnostic differences, diagnosis may be delayed or patients may be misclassified as another rheumatic disease due to the insidious onset and the broad spectrum of clinical manifestations of the disease. In addition, Sjögren's Syndrome (SS) can occur on its own (primary SS) or in association with other systemic autoimmune diseases (secondary SS). Given the vast availability of electronic health records (EHRs) for the general population, computational phenotyping may help to improve the diagnosis and timely referral of patients with complex diseases such as pSS to the medical specialist. Computational phenotyping algorithms are automated patient selection algorithms to identify a patient population of interest (Wiley, 2020). Such algorithms are increasingly used to identify and characterize patients with complex medical conditions from heterogeneous EHR data in order to improve efficiency of health care delivery and clinical outcomes (Chen, 2018).

2.1.2. Primary care data

Primary care EHRs are a rich source of information about people´s health and health service utilization. In countries with a gatekeeping system, general practitioners (GPs) have a fixed practice population and they are normally the first point of contact with the health care system. Routinely recorded electronic health care data in primary care may be used to develop early detection models or estimate population prevalences for diseases such as pSS defined as "complex with rare com-

plications" (Romão et al., 2018) and in general, to study the disease in a "real life" situation, outside the setting of a specialized clinical center (Maciel et al., 2018).

In the Netherlands, and in many other countries in Europe (e.g., the United Kingdom, Italy, and Spain), primary care practices use an EHR system to record the care delivered to their patients and the health problems presented (Verheij et al., 2018). The diagnoses that are recorded can be assessed by the GP, but also in other sectors of the health care system, such as medical specialists. For many diseases, GPs are unable to diagnose the patients themselves so patients are referred to a medical care specialist for diagnosis and treatment. Diagnoses recorded in the GP EHR data are therefore not necessarily diagnoses made by the GPs but also include those of other health care specialists.

Two characteristics make it worthwhile to investigate primary care EHR data in relation to Sjögren's syndrome:

1. The GP is the first point of contact with the health care system. This allows us to identify the patient's first symptoms and to analyze the care trajectories that eventually lead to the diagnosis of Sjögren and its treatment in primary care and eventually in secondary care.

2. There is a fixed patient list. This means that the data recorded in primary care are population based and that there is an epidemiological denominator available.

One of the difficulties in identifying patients with pSS, or any other relatively rare disease, from EHRs is the coding system used in primary care. GPs in the Netherlands use the International Classification of Primary Care (ICPC) coding system to record diagnoses and symptoms. The ICPC coding system was especially devised for primary care settings (Lamberts & Wood, 1987). In contrast with for example the International Classification of Diseases coding system used in secondary care (ICD; World Health Organization, 2004), ICPC has separate entries for symptoms (such as belly ache) and for diagnoses (such as urinary tract infection). However, as there are only about 700 separate entries, the level of granularity of ICPC coded primary care records is lower than that of the ICD coded records in secondary care (Cardillo et al., 2015).

Due to the low granularity of the ICPC coding system, there is no separate ICPC code for pSS. pSS is recorded under "Musculoskeletal disease other (L99)", as are for example Systemic Lupus Erythematosus and Systemic Sclerosis, which are autoimmune disorders that can occur in association with Sjögren's syndrome (Pasoto et al., 2019). An important consequence for our purposes is the fact that there is no simple way to identify pSS patients from primary care EHRs and no gold standard available to validate any patient selection made based on alternative rules or criteria. However, this information may be available from other sources, such as insurance claims data from secondary care.

2.1.3. Secondary care data

As the GP is the first point of contact in the Netherlands, undiagnosed patients will first visit their GP with any complaints typical for Sjögren's syndrome. When

the GP suspects Sjögren's syndrome, the GP will refer the patient to the Rheumatologist, Internist, or Ophthalmologist for specialized care and diagnosis. After formal diagnosis, general care for Sjögren's patients consists of follow-up appointments (medical checkups) with the medical specialist and symptomatic treatment (e.g., artificial tears or artificial saliva to reduce the symptoms of drought). After first description of these drugs by the specialist, repeat prescriptions are generally prescribed by the GP. The medical specialist informs the GP of the diagnosis made, which is then included by the GP in the patient's primary care EHR. The fact that all suspected Sjögren's patients are eventually referred to secondary care for diagnosis and treatment means that all Sjögren's patients should ultimately show up in secondary care records. Diagnostic information can be retrieved from hospital claims data using two classification systems; the diagnosis related groups (DRG) for hospital reimbursements and aforementioned ICD coding system for diseases. Both systems contain explicit codes for Sjögren's disease.

This study investigates to what extent routinely recorded EHR data can be used to identify patients with complex diseases. To this aim we first examined how formal inclusion and exclusion criteria for pSS could be translated into a computational phenotyping algorithm to identify pSS patients in primary care. As the primary care data do not contain a gold standard to validate the algorithm, we secondly assessed whether secondary care data could be used as an alternative validation method, by comparing the resulting patient selection with DRG and ICD codes retrieved from hospital claims data. In order to assess the overall fitness for purpose of routinely recorded health care data for the identification of patients with complex diseases such as pSS, we finally compared prevalence rates and patients' demographic characteristics to those reported in literature.

2.2. Methods

2.2.1. Data sets

General practitioner electronic health records

Nivel is a research institute that is part of the Dutch national health knowledge infrastructure. Nivel is commissioned by the Dutch Ministry of Health to collect data from EHRs in primary care, in Nivel Primary Care Database (Nivel PCD). Nivel PCD collects routinely recorded data from health care providers to monitor the health of patients and the utilization of health services in a representative sample of the Dutch population. Data are extracted periodically, and patients can be followed through the health care system longitudinally when the Nivel data is linked to other national databases.

For this study, data were extracted for the years 2006-2017, containing consultations, diagnoses, prescriptions, referrals, and patient characteristics (Nivel, n.d.). Diagnoses are recorded routinely in general practices and GPs use the ICPC classification system. Due to privacy regulations, the database contains no information stored in free text fields, apart from the titles of the disease episodes. This project has been approved by the governance bodies of Nivel PCD under no. NZR-00317.057.

Hospital claims database

The national claims data set is provided by Diagnosis Related Groups Information System (DIS) and is accessible and linkable through Statistics Netherlands, a government institution that makes data available for policy development and scientific research. The data set includes claims data, using the DRG classification system for hospital reimbursements (Hasaart, 2011), for all hospitals in the Netherlands.

DRG codes were available for the years 2013-2017 at the time of research (November 2019). DRG codes for Sjögren's syndrome are recorded under three medical specialisms; Rheumatology (DRG code 0324-03-00-0308), Internal Medicine (DRG code 0313-05-00-0524), and Ophthalmology (DRG code 0301-40-00-0404). For the most recent years (2016-2017), ICD-10 codes are increasingly available, although not complete. The ICD-10 code for Sjögren's syndrome is M35.0 (sometimes recorded as M350).

Population

In the Netherlands, all non-institutionalized inhabitants are compulsorily listed with a general practice, even if they do not visit their GP regularly. Nivel PCD contains primary care data of 1.7 million individuals (10% of the Dutch population), enlisted in approximately 500 GP practices. The practices included in Nivel PCD and patients enlisted in each practice may vary over the years. Patients can be tracked over time and linked to other sources based on pseudonymized citizen numbers. In total we analyzed the EHRs of 3,056,928 unique patients enlisted in any practice included in Nivel PCD over the years 2006-2017. The DIS database contains DRG coded insurance claims data for 12,991,265 unique patients who consulted a medical specialist in the Netherlands between 2013-2017.

2.2.2. Developing the algorithm

The first aim of this study was to assess whether primary care electronic health care data could be used to identify pSS patients from primary care electronic health care records. The formal ACR/EULAR classification criteria for pSS were used as a starting point to define inclusion and exclusion criteria for patient selection, but additional information available from the primary care database, such as drug prescriptions and disease episode titles, was also explored.

Formal classification criteria for Sjögren's syndrome

Inclusion criteria The ACR/EULAR criteria (Shiboski et al., 2017) include patients who report at least one symptom of ocular or oral dryness and score above a certain threshold on certain weighted criteria items. Ocular or oral dryness is assessed by diagnostic questions regarding recent eye complaints, use of artificial tears, reporting of dry mouth, and difficulty swallowing food. The weighted criteria concern labial salivary gland histopathology, anti-SSA/Ro antibodies, ocular staining score, Schirmer's test, and unstimulated whole saliva flow rate.

Exclusion criteria The ACR/EULAR criteria (Shiboski et al., 2017) exclude patients with a prior diagnosis of the conditions: history of head and neck radiation

treatment, Active hepatitis C infection (with confirmation by polymerase chain reaction), AIDS, Sarcoidosis, Amyloidosis, Graft-versus-host disease, or IgG4-related disease.

Secondary Sjögren's syndrome In order to distinguish specifically primary Sjögren's syndrome, Systemic Lupus Erythematosus, Systemic Sclerosis, and Rheumatoid Arthritis should additionally be excluded (Pasoto et al., 2019).

Data recorded in primary care

To identify the possible pSS patients, the formal criteria were translated into a set of rules relating to coded diagnoses, comorbidities, and diagnostic test results. We additionally explored drug prescriptions and disease episode titles. These rules were applied in the form of automated queries on the database. Except for the disease episodes title, no free text fields could be used.

ICPC codes In secondary care the patients with pre-specified diseases can be included and excluded using ICD-10 codes. To apply the ACR/EULAR criteria to primary care data, the ICD-10 codes were converted to the corresponding ICPC codes using the WHOFIC Thesaurus ICPC2-ICD10 (WHO Collaborating Centre for the Family of International Classifications, 2012). The resulting ICPC codes were applied to ICPC coded GP contacts (e.g., consults, prescriptions) and disease episodes.

Diagnostic test results The ACR/EULAR criteria (Shiboski et al., 2017) mention several diagnostic tests that can aid in the diagnosis of pSS. Although Nivel PCD contains a range of diagnostic test results, these cover only the results of tests issued or conducted by GPs. Diagnostic test results are recorded in Nivel PCD using NHG lab codes, defined by the Nederlands Huisartsen Genootschap (Dutch College of General Practitioners) for the classification of laboratory and other diagnostic tests and results (Westerhof & Bastiaanssen, 2011). It was checked how many of the diagnostic tests defined by Shiboski et al. (2017) were recorded in Nivel PCD.

Prescriptions The prescriptions in Dutch primary care are coded using the international Anatomical Therapeutic Chemical (ATC) Classification system for medicines (WHO Collaborating Centre for Drug Statistics Methodology, n.d.). In order to strengthen the patient selection, we examined the use of certain medication known to be much used by pSS patients (Ramos-Casals et al., 2020). These are Artificial tears (ATC S01XA20), Hydroxyclorochine (ATC P01BA02), Cortisone (ATC H02AB10 / S01BA03), Pilocarpine (ATC N07AX01), and Ciclosporin (ATC S01XA18). Especially the combined use of Artificial tears, Hydroxychlorochine, and Pilocarpine was expected to be a strong indicator of pSS.

Disease episode titles Finally, a text query was applied on all disease episode titles recorded between 2006 and 2017. The text query was based on a number of variations in spelling of the word "Sjögren" (namely "sjogren", "sjorgen", "sjogern", "sjögren", "sjorgren", "sjoegren", "sogren") in the disease episode titles in Nivel

PCD. The text strings of found cases were then manually checked and scored as to whether they described primary Sjögren syndrome by two of the authors as: 1) 'primary Sjögren'; 2) 'perhaps primary Sjögren'; or 3) 'not Sjögren' or explicitly 'secondary Sjögren'. All cases in which the term Sjögren was followed by a question mark were assigned to category 2. Cases explicitly described as secondary were scored as category 3. This, however, does not necessarily mean that all cases with score 1 are indeed primary Sjögren cases.

2.2.3. Validating the algorithm

As there is no formal diagnosis available to use as a gold standard to validate the developed algorithm, the second aim of this study was to assess to what extent hospital claims data, which contain more fine-grained DRG treatment and ICD-10 diagnosis codes for Sjögren, might be suitable as an alternative validation method. We additionally compared prevalence rates and demographic characteristics of pSS patients identified in primary and secondary care with those reported in literature.

Data linkage

EHR data of patients from Nivel PCD were linked to insurance claims data available from the DIS database on the basis of pseudonymized national citizen numbers. A Sjögren related DRG or ICD-10 code was regarded as a formal diagnosis used to confirm whether pSS patients in found in primary care were also recorded as pSS patients in secondary care.

Because Nivel PCD covers 10% of the Dutch population and the DIS database covers 100% of the Dutch population, it was expected that 10% of the patients found in the DIS database would be retrieved from Nivel PCD. Linkage is done using the patients' citizen service number (BSN), a unique personal number allocated to every registered Dutch citizen. The BSN is used by all recognized care providers, such as GPs, hospitals, and health insurance companies, to identify patients that need care. The BSN is included in Nivel PCD since the year 2014 and as such is not known for patients who did not consult the GP after 2013. For these patients, linkage on BSN level is not possible, leading to a linkage loss of around 10%.

Validation scores

Based on the linked data set, it is possible to compare the pSS patients found with the algorithm from Nivel PCD with formal diagnoses based on recorded DRGs and ICD-10 codes related to Sjögren within the health insurance claims data set. Based on the combined data sets each patient is flagged as a true positive, true negative, false positive, or false negative, as shown in Table 2.1:

- True positives (Tp): labelled as pSS by the algorithm, confirmed based on DRG codes secondary care claims database.
- True negatives (Tn): not labelled as pSS by the algorithm (hence not included in our data set), confirmed based on absence of DRG code related to Sjögren recorded in the secondary care claims data.

Table 2.1: Comparison of identified patients in Nivel and DIS databases

Patient	pSS Nivel PCD	pSS DIS database	Check
1	Yes	No	Fp
2	Yes	Yes	Tp
...	No	Yes	Fn
X	No	No	Tn

Note. pSS = primary Sjögren's syndrome. Nivel PCD = Nivel Primary Care Database, DIS = Dutch National Insurance Claims Database, Fp = False positives, Tp = True positives, Fn = False negatives, Tn = True negatives.

- False positives (Fp): labelled as pSS by the algorithm but not confirmed based on DRG codes secondary claims database.
- False negatives (Fn): not labelled as pSS by the algorithm, but DRG codes related to Sjögren recorded in the secondary claims database.

The total number of Tps, Tns, Fps, and Fns can be used to calculate the accuracy and other performance scores of the algorithm:

$$Accuracy = \frac{Tp + Tn}{Total} \tag{2.1}$$

$$Sensitivity\ (recall) = \frac{Tp}{Tp + Fn} \tag{2.2}$$

$$Specificity = \frac{Tn}{Tn + Fp} \tag{2.3}$$

$$Postive\ Pedictive\ Value\ (PPV, precision) = \frac{TP}{Tp + Fp} \tag{2.4}$$

$$Negative\ Predictive\ Value\ (NPV) = \frac{Tn}{Fn + Tn} \tag{2.5}$$

$$F_1\text{-}score = 2 \times \frac{precision \times recall}{precision + recall} \tag{2.6}$$

Prevalence rates

The prevalence rates are reported from the year 2011. The current Nivel PCD started in 2010, but as this was still a transition year, data for 2010 should be used with caution. The former database, known as the Netherlands Information Network database (LIN; Schweikardt et al., 2016), constituted of a different set of patients, practices, and reference population. This makes prevalence rates calculated from both databases incomparable.

The prevalence rate is calculated for each year by dividing the number of newly identified or existing pSS patients by the number of patients of the population in that year.

$$Prevalence\ rate = \frac{N(patients\ with\ new\ or\ existing\ pSS\ diagnosis)}{N(patient\ years\ of\ the\ population)} \times 1000 \tag{2.7}$$

2.3. Results

2.3.1. Patient selection algorithm

Several patient selection approaches (e.g., based on diagnoses, comorbidities, diagnostic test results, prescriptions, and disease episode titles) were compared to find the most applicable rules for the phenotyping algorithm. Details of the data set and the final phenotyping algorithm are provided in Appendix 2-A.

ICPC codes

Table 2.2 lists the ICPC codes (including counts) used to include and exclude patients with diseases related to Sjögren's syndrome based on the ACR/EULAR criteria.

Diagnostic test results

Of the diagnostic tests used for diagnosing pSS defined by Shiboski et al. (2017), the NHG lab codes include only the autoantibodies anti-Ro/SSA. Schirmer's test, salivary flow and ocular staining tests are generally conducted by the Rheumatologist or Ophthalmologist and as such are not recorded in primary care. Therefore test results were not used as input for the patient selection algorithm. After finalizing the patient selection, we did check for how many patients autoantibodies anti-Ro/SSA values were recorded in Nivel PCD; this was only for four of the 1,462 selected pSS patients.

Prescriptions

In addition to the ICPC codes, we assessed the use of Artificial tears, Hydroxychlorochine, Cortisone, Pilocarpine, and Ciclosporin, and the combined use of Artificial tears, Hydroxychlorochine, and Pilocarpine in specific. However, these medications are barely prescribed by the GP in the Netherlands. For example, for all Dutch patients in Nivel PCD in the period 2006-2017 (N = 3,056,928), prescription rates for Cortisone (N = 308), Pilocarpine (N = 200), and Ciclosporin (N = 143) are low. When applying the combination of the three prescribed medications to the final patient selection, only 24 of the patients that met the defined pSS selection criteria from Table 2.2 remained. The (combined) prescription use thus does not seem to be a feasible selection criterion for pSS.

To gain insight in the prescriptions that were used a lot by possible pSS patients, Table 2.3 shows the prescriptions with the highest recording rates over the complete period. In total 928 different medications were prescribed to the possible pSS patients found in Nivel PCD. Of these, especially artificial tears, proton pump inhibitors (Omeprazole and Pantoprazole), beta blocking agents (Metoprolol), and thyroid hormones (Levothyroxine) were highly used. Apart from artificial tears, these are among the highest used drugs in the general population and are probably related to other morbidities than pSS.

Table 2.2: Inclusion and exclusion criteria applied to Nivel PCD

Disseases (ICPC code)		N(patients)
Inclusions		
Patient has one or more of:	Other musculoskeletal diseases (L99)	347,082
	Other disease eye (F99)	267,361
	Non-Hodgkin's disease (B72.02)	3,674
Exclusions		
Patient has one or more of:	Hepatitis (incl. hepatitis C infection) (D72)	8,580
	Other infections of the lungs (R83)	62,091
	HIV (B90)	3,045
	Sarcoidosis (B99)	4,913
	Graft-versus-Host disease (A87)	22,345
	Amyloidosis (T99)	23,526
	IgG4-related disease (B99)	4,913
Exclusions secondary Sjögren*		
Patient has:	Rheumatoid arthritis (L88)	31,472

Note. ICPC = International Classification of Primary Care.

* Two of the three exclusion criteria for secondary Sjögren defined by Pasoto et al. (2019), Systemic Lupus and Systemic Sclerosis, could not be excluded because these are recorded under the ICPC code L99 ("Musculoskeletal disease other"), which is also the ICPC code for Sicca/Sjögren.

Disease episode titles

In total, one of the defined variations of the word 'Sjögren' occurred in the disease episode titles of 3,259 unique patients. The majority of GPs used the term 'Sjogren' (N = 2,944), followed by 'Sjögren' (N = 256), and various misspellings 'Sjorgen' (N = 31), 'Sjoegren' (N = 16), 'Sogren' (N = 7), 'Sjogern' (N = 3), and 'Sjorgren' (N = 2).

The distinction between primary and secondary Sjögren was not often explicitly made in the episode texts. For only 71 patients Sjögren was specifically defined as primary (indicated by 'prim', 'prim.', or 'primary') and for 65 patients as secondary (indicated by 'sec', 'sec.', or 'secondary'). When the GP was unsure of a patient having Sjögren this was often indicated by a question mark: e.g., 'Sjögren?' (N = 348). However, as the episode titles are free text fields, the variation in used text strings was high and each patient was assigned to one of the categories manually by taking into account the complete textual context.

Text strings interpreted as primary Sjögren were mainly clear and short state-

Table 2.3: Top 10 of in total 928 unique prescriptions used by pSS patients

ATC code	Description	N(records)
S01XA20	Artificial tears and other indifferent preparations	11,838
A02BC01	Omeprazole	7,584
A02BC02	Pantoprazole	5,483
C07AB02	Metoprolol	4,415
H03AA01	Levothyroxine	4,327
B01AC06	Acetylsalicylic acid	4,262
C10AA01	Simvastatin	4,090
P01BA02	Hydroxychloroquine	3,752
C03AA03	Hydrochlorothiazide	2,946
N05CD07	Temazepam	2,905

Note. ATC = Anatomical Therapeutic Chemical Classification system for medicines.

ments such as 'm. Sjögren', 'morbus Sjögren', and 'Sjögren's syndrome', without the mention of any secondary diseases (Table 2.2). Text strings interpreted as perhaps Sjögren contained words such as 'suspicion of Sjögren' or 'possibly Sjögren', or the use of a question mark. Text strings interpreted as not or secondary Sjögren clearly stated 'no(t) Sjögren', 'secondary Sjögren', 'Sjögren' combined with one of the secondary diseases, or regarded a family member having Sjögren or the patient only being afraid of having Sjögren. This resulted in the following counts per category: 1) 'primary Sjögren' (N = 2,319); 2) 'perhaps primary Sjögren' (N = 672); or 3) 'not Sjögren' or explicitly 'secondary Sjögren' (N = 268).

Final algorithm

The selection criteria based on the formal ACR/EULAR classification criteria (listed in Table 2.2), combined with the mention of "Sjögren" (or variations) in the disease episode titles were found to be the most suitable identifiers for pSS in primary care EHRs. To be defined as pSS patient, one or more of the ICPC inclusion criteria should be recorded in the patient journal in the defined period and "Sjögren" (or variations) should be mentioned in the disease episode titles. Only a record of one or more of the inclusion criteria and no mention of "Sjögren" (N = 623,700), or vice versa (N = 729), was not sufficient to be included as a pSS patient. Any patients for which any of the exclusion criteria were recorded were subsequently excluded from the selection. This resulted in a total sample of 1,462 plausible pSS patients that were retrieved from Nivel PCD, leading to a prevalence of 0.81 per 1,000 patients in 2017.

The flowchart in Figure 2.1 shows the inclusion and exclusion rules applied to the total number of patients extracted from the primary care database for the years 2006-2017 (N = 3,056,928). Of these, 625,809 patients visited the GP for one or more of the diseases related to Sjögren (L99, F99, or B72.02). Since it is possible for patients to visit the GP for either one or multiple defined diseases in the given

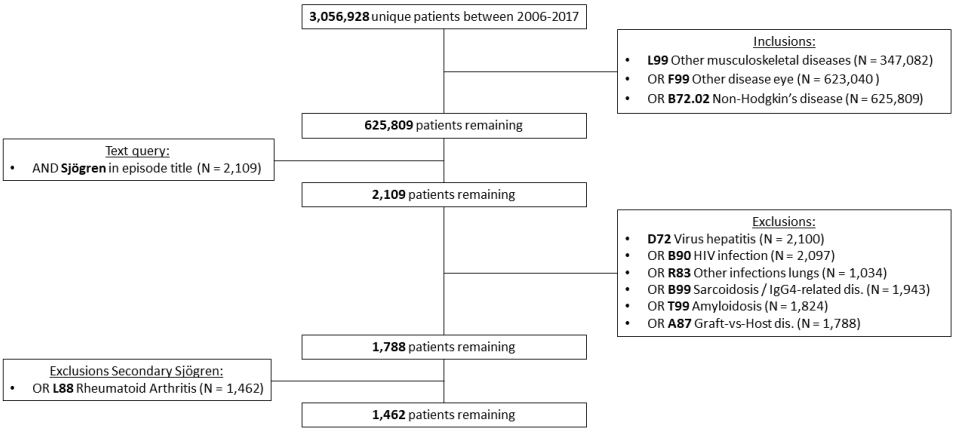


Figure 2.1: Flowchart of inclusion and exclusion criteria applied to Nivel PCD data (cumulative numbers)

period, the flowchart displays cumulative numbers per inclusion and exclusion step instead of absolute numbers per disease (which can be found in Table 2.2).

First, 347,082 patients were included because they visited the GP for complaints recorded under "Other musculoskeletal diseases" (ICPC code L99). Second, 275,958 additional patients recorded under "Other disease eye" (ICPC code F99) were included, leading to 623,040 patients with codes L99 or F99. Third, an additional 2,769 non-Hodgkin's disease (ICPC code B72.02) patients were included, leading to a total of 625,809 included patients who met at least one of the inclusion criteria.

Of these, only 2,109 also had "Sjögren" or any of the defined textual variations mentioned in the disease episode titles, leading to 2,109 remaining patients. Of these in total 321 patients were excluded because they visited the GP for one or more of the defined exclusion diseases (D72, B90, R83, B99, T99, or A87), leaving 1,788 patients. Finally, 326 of these patients were excluded as these were recorded as having Rheumatoid Arthritis (ICPC code L88), which was defined as a criterion for secondary Sjögren's disease, leaving 1,462 pSS patients.

2.3.2. Algorithm validation

The claims data indicate that on average around 54,000 unique patients per year visit the hospital for a treatment recorded under one of the Sjögren related DRGs or the ICD-10 code for Sjögren. Based on the estimated global prevalence of 61 per 100,000 (e.g., Qin et al., 2015) and a total Dutch population of 17 million, we would expect slightly over 10,000 patients. Table 2.4 shows the number of patients for whom Sjögren related DRGs were recorded in the years 2013-2017, or who had an ICD-10 recorded Sjögren diagnosis in the years 2016-2017. The majority of

Table 2.4: Number of pSS patients in secondary care

	N(unique patients)						
		DRG		ICD-10			
Year	Rheuma-tology (N=10,045)	Internal Medicine (N=1,447)	Ophthal-mology (N=201,648)	M35.0, M350 (N=34,933)	Total unique patients (N=208,545)	Reference population[a]	Population prevalence (per 1,000)
2013	4,740	703	56,862	n.a.	60,995	16,779,575	3.64
2014	5,089	629	53,250	n.a.	57,656	16,829,289	3.43
2015	5,073	584	50,439	n.a.	54,854	16,900,726	3.25
2016	4,870	550	46,044	22,104	50,427	16,979,120	2.97 (1.30)[b]
2017	4,896	554	42,277	18,958	46,621	17,081,507	2.73 (1.11)[b]

Note. DRG = Diagnosis Related Groups, ICD-10 = International Classification of Diseases. DRG data is available from 2013, one year after the implementation of the updated DRG system in 2012.

[a] Retrieved from StatLine Open Data provided by Statistics Netherlands (https://opendata.cbs.nl/statline), retrieved November 2019.

[b] Prevalence based only on patients with recorded ICD-10 code for Sjögren.

the identified patients were treated at the Ophthalmology department, followed by Rheumatology and Internal Medicine.

As the number of Sjögren patients in secondary care defined based on DRG and ICD-10 codes is higher than expected, we compared the recorded DRGs with the available ICD-10 codes as the ICD-10 codes are more explicit diagnoses and DRG codes might be too broad. In the years for which ICD-10 codes were available (2016 and 2017), much overlap was found between the Sjögren DRGs recorded in the Rheumatology and Internal Medicine departments. For Rheumatology, 4,397 of the 4,870 (90.3%) patients for which a Sjögren related DRG was recorded also had the ICD-10 Sjögren diagnosis recorded in 2016. For 2017 this was the case for 4,501 of the 4,896 patients with a Rheumatology DRG (91.9%). For Internal Medicine 444 of the 550 (80.7%) patients had both the Sjögren related DRG and ICD-10 diagnosis code in 2016, and 458 of the 554 (82.7%) in 2017. For the Ophthalmology department this overlap was a lot smaller; only 17,806 of the 46,044 (38.7%) patients with a Sjögren DRG in 2016 and 14,596 of the 42,277 (34.5%) patients with a Sjögren DRG in 2017 also had the ICD-10 code for Sjögren recorded. For the patients for which no Sjögren ICD code was recorded, the ICD code was mainly missing, or referred to an "Unspecified Illness" (R69), Myositis Ossificans Progressiva (M61.19), Congenital malformation syndromes predominantly associated with short stature (Q87.1), or Other disorders of lacrimal gland (H04.1). Especially the latter was highly recorded for patients with a Sjögren DRG at the Ophthalmology department.

To check whether the Sjögren DRGs for each department included any diseases related to Secondary Sjögren, we checked for the presence of ICD-10 codes related to the three secondary diseases listed in Table 2.2 among the patients with Sjögren related DRGs in the period 2016-2017. For the Internal Medicine department, none

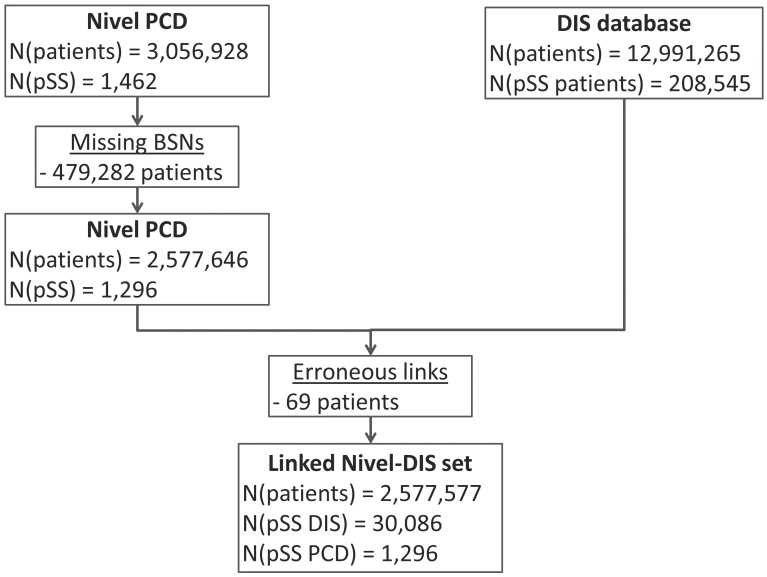


Figure 2.2: Linkage process primary and secondary care data

of the patients with a Sjögren DRG was diagnosed with any of the secondary diseases. For the Rheumatology department, 75 unique patients were diagnosed with Rheumatoid Arthritis and ≤ 10 patients with Systemic Lupus Erythematosus or Systemic Sclerosis. For the Ophthalmology department, ≤ 10 patients were diagnosed with Systemic Lupus Erythematosus or Systemic Sclerosis and none with Rheumatoid Arthritis.

Linked patients

For 208,545 of the 12,991,265 unique patients who visited a medical specialist in any hospital in the Netherlands between 2013-2017, a Sjögren related DRG or ICD-10 code was recorded. In total, 2,577,577 of the 3,056,928 patients included in Nivel PCD could be linked to the secondary care data in the DIS database. Among the linked patients, 30,086 of the initial 208,545 patients with a Sjögren related DRG or ICD-10 code from the DIS database remained, against 1,296 of the 1,462 pSS patients found in Nivel PCD, as shown in Figure 2.2.

Validation scores

The matrix in Table 2.5 visualizes the performance of the algorithm applied to Nivel PCD by comparing the patients found in Nivel PCD to the formally diagnosed patients in the DIS database. The cells contain the true and false positives and negatives. The number of true positives (Tp) shows that 716 out of the 1,296 (55.3%) linked patients that were likely to have pSS in Nivel PCD, indeed visited the hospital medical specialist for a Sjögren related treatment (DRG) between 2013-2017 or were recorded as a Sjögren patient (ICD-10) during a visit to the hospital

Table 2.5: Confusion matrix

		Formal pSS diagnosis (DIS)		
		pSS	Non-pSS	Total
Possibly pSS (Nivel PCD)	pSS	**716** (Tp)	580 (Fp)	1,296
	Non-pSS	29,370 (Fn)	**2,546,911** (Tn)	2,576,281
	Total	30,086	2,547,491	2,577,577

Note. pSS = primary Sjögren syndrome, Nivel PCD = Nivel Primary Care Database, DIS = Dutch National Insurance Claims Database, Tp = True positives, Fn = False negatives, Fp = False positives, Tn = True negatives

in 2016-2017. 580 of the 1,296 linked patients remained unconfirmed based on the DRG data set, meaning that these pSS patients did not visit the hospital for a Sjögren related treatment in the years 2013-2017 or did not receive a formal diagnosis recorded by a specialist (ICD-10 codes) in the years 2016-2017.

Table 2.5 shows 580 of the 1,296 linked patients who were identified as possible pSS patients based on Nivel PCD data were not confirmed based on information from the DIS database. These may not be pSS patients, or pSS patients that did not visit a hospital for a Sjögren related treatment in the years 2013-2017. Of the 580 unconfirmed pSS patients in primary care, 213 visited the hospital in the defined period for other DRGs (e.g., Cataract (N = 71), Chest pain (N = 65), Perceptive hearing loss (N = 46), or Osteoarthritis of the knee (N = 41)), whereas 367 did not visit the hospital at all. 29,370 of the 30,086 patients who visited the hospital for a Sjögren related treatment were not identified as a possible pSS patient in Nivel PCD. The values from the matrix lead to the following performance scores; Accuracy (92.4%), Sensitivity/recall (2.38%), Specificity (99.98%), PPV/precision (55.25%), NPV (98.84%), F_1-score (4.56%).

Patient characteristics

Table 2.6 shows the mean age and gender of the total population included in Nivel PCD and for the selected pSS patients over the years. It also shows the number of new and known pSS patients for each year. The number of new patients in a given year is the number of patients for which "Sjögren" was mentioned for the first time in the ICPC episode title in that year. The total number of patients in a given year is the number of new patients in that year added to the number of patients known from previous years.

As the first year of diagnosis we used the first date in which a record was found in the journal in which "Sjögren" was mentioned in the ICPC text episode title. This date was unknown for 810 pSS patients, probably because the diagnosis was made before the patient had been listed as patient in the practice for which data is included in the database. For the patients for which this date could be retrieved, the majority was between 50-70 years of age at the first year of diagnosis (see

Table 2.6: Nivel PCD sample and population characteristics

Year	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Total
						Total subjects in database							
Population	164,678	221,887	217,025	288,795	885,226	1,313,607	1,558,192	1,803,696	1,870,279	1,926,817	1,890,690	1,810,851	3,056,928
Gender*													
Male	81,918	109,253	106,617	142,134	435,822	647,149	767,072	888,692	922,460	952,151	934,120	894,863	
Female	82,755	112,628	110,407	146,660	449,404	666,458	791,120	915,003	947,407	974,600	956,203	915,745	
Unknown	5	6	1	1	0	0	0	1	412	66	367	243	
Age													
Mean	37.18	37.80	38.76	39.00	39.07	39.72	39.53	40.05	43.79	44.77	42.86	43.47	
SD	21.94	22.31	22.38	22.41	22.72	22.91	22.87	23.31	82.03	89.80	62.15	68.24	
						Selected pSS patients							
New	17	29	39	37	40	61	48	58	56	61	77	68	
Total	888	917	856	993	1,033	1,094	1,142	1,200	1,256	1,317	1,394	1,462	1,462
Non-patients	163,790	220,970	216,169	287,802	884,193	1,312,513	1,557,050	1,802,496	1,869,023	1,925,500	1,889,296	1,809,389	
Prevalence (per 1,000)	-	-	-	-	-	0.83	0.73	0.67	0.67	0.68	0.74	0.81	
Gender													
Males	100	107	113	118	122	124	133	142	148	157	167	177	177
Females	788	810	843	875	911	970	1,009	1,058	1,108	1,160	1,227	1,285	1,285
Age*													
Mean	54.74	55.81	56.73	57.62	58.82	59.55	60.35	61.15	62.05	62.81	63.65	64.43	
SD	33.56	33.10	32.49	32.05	31.54	30.89	30.41	29.85	29.37	28.85	28.21	27.78	

Note. pSS = Primary Sjögren syndrome.

* This is the mean age of the patients included in the database in the concerning year. We do not report mean age in year of first diagnosis here, because the first diagnosis year is unknown for a large group of patients who were diagnosed before inclusion in Nivel PCD (N = 810).

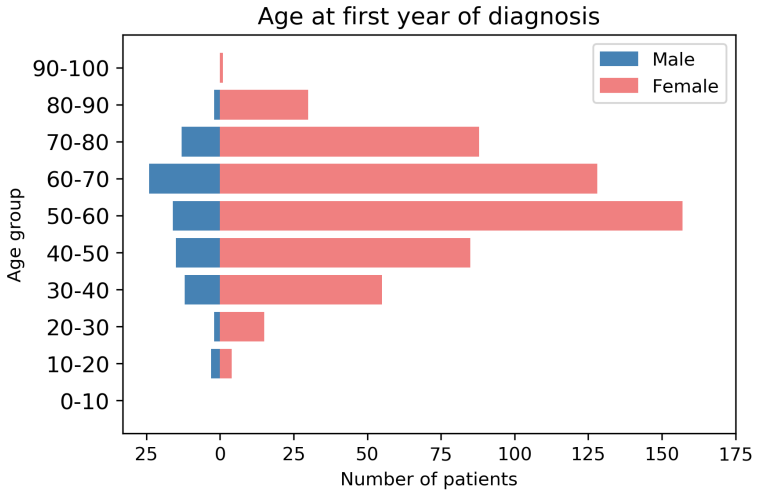


Figure 2.3: Age distribution at first diagnosis year

Figure 2.3, with a mean of 65.8 years (SD 15.1).

Prevalence rates

Table 2.6 displays the prevalence rates, in which the total and new number of pSS patients are compared to the total patient population in Nivel PCD. These rates show the prevalence has slightly increased in the most recent years, after a slight decrease in the first years of the new database. On average the prevalence of pSS patients in Nivel PCD was 0.7‰.

2.4. Discussion

This study illustrates the potential use of routinely recorded primary and secondary care EHR data to identify and validate patients with complex diseases such as pSS. A patient selection algorithm was developed based on known inclusion and exclusion criteria used in the diagnosis of patients with Sjögren's syndrome. ICPC coded diseases combined with keywords extracted from episode text titles were found to be the most suitable for identifying possible pSS patients in primary care, resulting in 1,462 possible pSS patients identified in primary care. The patients selected by the algorithm were compared to patients treated for Sjögren's syndrome in secondary care, resulting in a confirmation of 716 of the 1,296 linked pSS patients (55.3%).

The first part of our study focused on the question how formal inclusion and exclusion criteria for pSS used by medical specialists in secondary care could be applied to EHR data recorded in primary care. The exact ACR/EULAR classification criteria for pSS could not be easily applied to the available primary care data. The ICPC codes are less granular than the specified ICD-10 codes used in secondary care, and cover more diseases than the ones specified as a single inclusion or

exclusion criterion. In addition, GPs often record only the main ICPC disease codes and not always the more specific sub codes. This complicated the inclusion and exclusion of explicit sub diseases such as Hepatitis C infection, which was now excluded using the overarching main category "Hepatitis". Another consequence of the broader ICPC codes was that secondary Sjögren's diseases Systemic Lupus Erythematosus and Systemic Sclerosis could not be excluded, as these are both recorded under ICPC code L99 (Musculoskeletal disease other), which is also the code generally used for Sjögren's Syndrome. The translation of the formal disease based classification criteria to criteria applicable to primary care data thus may have resulted in a less precise selection of pSS patients.

Consequently, the selection of pSS patients based on ICPC codes only was not specific enough. Combining the ICPC codes with a mention of Sjögren in the disease episode title narrowed the selection down to more a plausible number of patients. Besides the ICPC disease codes and the episode titles, we examined ATC coded medication prescribed by the GP that was expected to be frequently used by pSS patients, and NHG coded diagnostic test results conducted by the GP. However, prescription rates for the defined medications were quite low. This could be because this type of medication is not used a lot in the Netherlands, possibly because some are not covered by the general health insurance, or because these are prescribed by specialists in the hospital and not by the GP (and therefore cannot be retraced in our database). With regard to the diagnostic tests, it was found that only one of five tests that can be used to diagnose pSS (Shiboski et al., 2017) is used by GPs in the Netherlands, and recordings of their use are very limited. Although primary care EHRs are quite extensive, only a limited amount of the information needed to apply the formal diagnostic criteria for pSS was available in primary care. Based on the information that was available in the GP records, an alternative phenotyping algorithm could be developed to define a plausible set of pSS patients.

The second part of our study focused on the question of whether DRG and ICD codes retrieved from hospital claims data could be used to validate the primary care algorithm and resulting patient selection. The number of Sjögren related DRGs in the DIS database seems highly inflated when compared to known global prevalence estimates. When using both the DRGs and ICD-10 codes recorded at the Rheumatology, Internal Medicine, and Ophthalmology departments, the relative number of pSS patients found and the corresponding prevalence rates are much higher than those found in Nivel PCD and reported by Qin et al. (2015).

There may be several reasons for this overestimation in secondary care. First, it may be the consequence of strategic recording behavior. DRGs that are used as a basis for reimbursement (as is the case for the DIS database) have been found to be at risk for upcoding (Steinbusch et al., 2007; Verheij et al., 2018). A second reason could be that the recorded DRGs may only be indicative of a suspected Sjögren diagnosis, for which the treatment results turn out to be negative. However, the comparison of DRGs with the ICD-10 diagnosis codes recorded at each department for the years 2016 and 2017 showed a high overlap between the DRGs and ICD codes recorded at the Rheumatology and Internal Medicine departments. This may indicate that the DRGs of these departments do not suffer from upcoding and reflect

the formal diagnoses recorded by means of the ICD code. This does not seem to apply to the Ophthalmology department, for which less than half of the recorded DRGs overlapped with the ICD codes.

Another reason for the high number of pSS patients recorded in the claims data set might be that the DRGs include cases of primary as well as secondary Sjögren. A check for the presence of ICD-10 diagnosis codes related to secondary Sjögren diseases showed that the DRGs recorded at the Rheumatology department included a small number of patients (N = 75) diagnosed with the secondary Sjögren disease Rheumatoid Arthritis. The other secondary diseases were only recorded for very little (N ≤ 10) patients at the Rheumatology and Ophthalmology departments. No patients with secondary Sjögren diseases were included in the DRGs recorded at the Internal Medicine department. This shows that the DRGs include mainly primary and only very few secondary Sjögren's patients. Although DRGs from hospital claims data may not provide sufficiently accurate diagnostic information to be reliably used for the validation of patient selection algorithms, our analyses did show that, especially for Rheumatology and Internal Medicine, DRGs are a suitable alternative for ICD codes when ICD codes are not available.

Despite the high number of recorded Sjögren DRGs, the comparison of pSS patients found in primary care with those treated in secondary care resulted in a relatively low number of confirmed patients. There may be several explanations for this:

1. Some patients found via the algorithm in general practice may not have been referred to specialized care (yet). This is a plausible explanation, as the average time to diagnosis of Sjögren's syndrome, the time it takes for a patient to be referred to a specialist to get a formal diagnosis, is known to be long.

2. Some patients' last visit to the hospital for a Sjögren related treatment had taken place before 2013.

3. Some patients have received secondary care treatment (DRGs) or diagnosis (ICD-10 codes) other than the ones defined by us.

4. Despite meeting the criteria from the algorithm, some of these patients may not have been pSS patients, meaning the algorithm incorrectly identified some patients as possible Sjögren patients. In order to examine this further, the characteristics of the confirmed and unconfirmed patient groups were compared (check for significant differences).

5. In spite of claims regulations, DRG groups in claims data may not represent true pSS patients.

In future research we will first focus on exploring and confirming these possible explanations by comparing the primary and secondary care characteristics of the confirmed and unconfirmed pSS patients. Second, we aim to fine-tune the patient selection algorithm for primary care and the resulting patient selection by studying the characteristics of the pSS patients that were included in the DIS database but

that were not found in Nivel PCD based on the initial selection criteria. This may result in additional pSS identifiers in primary care, to be implemented in an improved, more precise algorithm for the selection of pSS patients in general practice. Third, we will develop a timeline displaying the average combined primary and secondary care trajectory of pSS patients in the Netherlands, using the linked Nivel PCD and DIS data of the confirmed pSS patients. This timeline will provide more insight into the used health care and the diagnostic process.

When looking at the prevalence rates based on the Dutch primary care database, we see the average prevalence based on our final algorithm (0.7‰) is comparable to the global population prevalence of 0.61‰ reported by Qin et al. (2015). Our mean age at diagnosis (Figure 2.3) is comparable to the average age of 56.16 years reported by Qin et al. (2015). The female:male ratio in our sample is 7:1, which is to be expected as pSS primarily affects peri- and postmenopausal women. Our female:male ratio is lower than the ratio in the prevalence data reported by Qin et al. (2015), which was 11:1. The proportion and characteristics of the pSS patients in primary care identified by the phenotyping algorithm are thus mostly in line with those reported in literature. The number of pSS patients in secondary care, however, highly exceeded the number expected based on the general population prevalence. Even when using only ICD-10 codes, which might be a more accurate source of diagnostic information, the prevalence found for the Netherlands still exceeds global estimates. There is not enough information to assess whether this discrepancy can be attributed to the sources and methods used to identify pSS patients in secondary care or the possibility that literature reported global prevalence rates might not be accurate for the Netherlands. This has a major impact on our study results in that it is unclear whether insurance claims records are a suitable source to compare and confirm the results obtained from primary care data with and, consequently, we cannot draw unambiguous conclusions regarding the quality of our patient selection and the developed phenotyping algorithm.

This study shows the possibilities of using EHR data for studying complex medical conditions. It is clear that population-based health records provide a lot of longitudinal medical information and insight in the use of care for a large range of diseases. However, the study of patients with low prevalence, uncoded diseases is more challenging, as those cannot be as easily identified from primary care data as patients with more general diseases. The lack of a granular coding system for symptoms and diseases also makes it difficult to apply diagnostic criteria used in secondary care to data recorded in primary care. The possibility to link primary to secondary care databases on patient level allows one to (iteratively) try different patient selection algorithms and compare those to patients referred to specialized care, and to study patient and care characteristics in primary care of patients thus far only known in secondary care. As such, these combined medical data should be considered a rich source of information for the epidemiological study of low prevalence, complex diseases, patients' early symptoms, diagnosis paths, and overall treatment trajectories in primary and eventually secondary care. However, without the formal diagnostic information required to validate the developed phenotyping algorithm and patient selection, we have insufficient information to affirm that routine

EHR data are fit for the identification and study of patients with complex diseases such as pSS.

2.5. References

Brito-Zeron, P., Baldini, C., Bootsma, H., Bowman, S. J., Jonsson, R., Mariette, X., Sivils, K., Theander, E., Tzioufas, A., & Ramos-Casals, M. (2016). Sjögren syndrome. *Nature Reviews Disease Primers*, *2*(1), 1–20. https://doi.org/10.1038/nrdp.2016.47

Cafaro, G., Croia, C., Argyropoulou, O. D., Leone, M. C., Orlandi, M., Finamore, F., Cecchettini, A., Ferro, F., Baldini, C., & Bartoloni, E. (2019). One year in review 2019: Sjögren's syndrome. *Clinical and Experimental Rheumatology*, *37*(Suppl 118), S3–15.

Cardillo, E., Chiaravalloti, M. T., & Pasceri, E. (2015). Assessing ICD-9-CM and ICPC-2 use in primary care. An Italian case study. *Proceedings of the 5th International Conference on Digital Health 2015*, 95–102. https://doi.org/10.1145/2750511.2750525

Chen, R. (2018). *Tackling chronic diseases via computational phenotyping: Algorithms, tools and applications* (Doctoral dissertation). Georgia Institute of Technology, Atlanta, GA.

Daniels, T. E., & Fox, P. C. (1992). Salivary and oral components of Sjögren's syndrome. *Rheumatic Disease Clinics of North America*, *18*(3), 571–589.

Hasaart, F. (2011). *Incentives in the diagnosis treatment combination payment system for specialist medical care: A study about the behavioral responses of medical specialists and hospitals in the netherlands*. Maastricht University. https://doi.org/10.26481/dis.20111104fh

Lamberts, H., & Wood, M. (1987). *ICPC, international classification of primary care*. Oxford University Press.

Maciel, G., Servioli, L., Nannini, C., Berti, A., Crowson, C. S., Achenbach, S. J., Matteson, E. L., & Cornec, D. (2018). Hospitalisation rates among patients with primary Sjögren's syndrome: A population-based study, 1995–2016. *RMD open*, *4*(1), e000575. https://doi.org/10.1136/rmdopen-2017-000575

Nivel. (n.d.). *Nivel Primary Care Database* [Accessed: 2019-11-01]. https://www.nivel.nl/en/nivel-primary-care-database/

Pasoto, S., de Oliveira, M., & Bonfa, E. (2019). Sjögren's syndrome and systemic lupus erythematosus: Links and risks. *Open Access Rheumatology*, *11*, 33–45. https://doi.org/10.2147/OARRR.S167783

Patel, R., & Shahane, A. (2014). The epidemiology of sjögren's syndrome. *Clinical Epidemiology*, *6*, 247–255. https://doi.org/10.2147/CLEP.S47399

Qin, B., Wang, J., Yang, Z., Yang, M., Ma, N., Huang, F., & Zhong, R. (2015). Epidemiology of primary sjögren's syndrome: A systematic review and meta-analysis. *Annals of the Rheumatic Diseases*, *74*(11), 1983–1989. https://doi.org/10.1136/annrheumdis-2014-205375

Ramos-Casals, M., Brito-Zerón, P., Bombardieri, S., Bootsma, H., De Vita, S., Dörner, T., Fisher, B. A., Gottenberg, J.-E., Hernandez-Molina, G., Kocher, A., et al. (2020). Eular recommendations for the management of sjögren's syndrome with topical and systemic therapies. *Annals of the Rheumatic Diseases*, *79*(1), 3–18. https://doi.org/10.1136/annrheumdis-2019-216114

Romão, V. C., Talarico, R., Scirè, C. A., Vieira, A., Alexander, T., Baldini, C., Gottenberg, J.-E., Gruner, H., Hachulla, E., Mouthon, L., et al. (2018). Sjögren's syndrome: State of the art on clinical practice guidelines. *RMD open*, *4*(Suppl 1), e000789. https://doi.org/10.1136/rmdopen-2018-000789

Schweikardt, C., Verheij, R. A., Donker, G. A., & Coppieters, Y. (2016). The historical development of the dutch sentinel general practice network from a paper-based into a digital primary care monitoring system. *Journal of Public Health*, *24*(6), 545–562. https://doi.org/10.1007/s10389-016-0753-4

Seghieri, C., Lupi, E., Exarchos, T. P., Ferro, F., Tzioufas, A. G., & Baldini, C. (2019). Variation in primary sjögren's syndrome care among european countries. *Clinical and Experimental Rheumatology*, *37*(3), 27–28.

Shiboski, C. H., Shiboski, S. C., Seror, R., Criswell, L. A., Labetoulle, M., Lietman, T. M., Rasmussen, A., Scofield, H., Vitali, C., Bowman, S. J., et al. (2017). 2016 american college of rheumatology/european league against rheumatism classification criteria for primary sjögren's syndrome: A consensus and data-driven methodology involving three international patient cohorts. *Arthritis & Rheumatology*, *69*(1), 35–45. https://doi.org/10.1002/art.39859

Steinbusch, P. J., Oostenbrink, J. B., Zuurbier, J. J., & Schaepkens, F. J. (2007). The risk of upcoding in casemix systems: A comparative study. *Health policy*, *81*(2-3), 289–299. https://doi.org/10.1016/j.healthpol.2006.06.002

Verheij, R. A., Curcin, V., Delaney, B. C., & McGilchrist, M. M. (2018). Possible sources of bias in primary care electronic health record data use and reuse. *Journal of Medical Internet Research*, *20*(5), e9134. https://doi.org/10.2196/jmir.9134

Westerhof, H., & Bastiaanssen, E. (2011). Nieuwe versie nhg-tabel diagnostische bepalingen. *SynthesHis*, *10*(3), 8–9. https://doi.org/10.1007/s12494-011-0044-3

WHO Collaborating Centre for Drug Statistics Methodology. (n.d.). *ATC/DDD Index 2019* [Accessed: 2019-11-01]. https://www.whocc.no/atc_ddd_index/

WHO Collaborating Centre for the Family of International Classifications. (2012). *Thesaurus ICPC-2 ICD-10* [Accessed: 2019-11-01]. https://www.whofic.nl/media/901

Wiley, L. (2020). Introduction to Computational Phenotyping. [Online course].

World Health Organization. (2004). *The international statistical classification of diseases and health related problems ICD-10: Tenth revision. volume 1: Tabular list* (Vol. 1).

Appendix 2-A
Phenotyping algorithm and data set details

Phenotyping algorithm

Patients = all patients enlisted in any Nivel PCD practice at any moment during the years 2006-2017

```
Result = possible pSS patients based on the input data

Journal_To_Include = (L99 or F99 or B72.02)
Episode_To_Include = ('sjogren' or 'sjorgen' or 'sjogern' or
                      'sjögren' or 'sjorgren' or 'sjoegren' or
                      'sogren')
Journal_To_Exclude = (D72 or B90 or R83 or B99 or T99 or A87 or L88)

for patient in Patients:
    if (patient.Journal in Journal_To_Include and
            patient.Episode contains Episode_To_Include
        ) and not (patient.Journal in Journal_To_Exclude):
        add to Result
```

Data set description

The data set covers GP recorded medical information on patient level for 3,056,928 patients enlisted in any practice included in Nivel PCD at any moment during the period 2006-2017.

The data set includes data on the topics: patient, practice, journals, prescriptions, episodes, and test results. Data for each topic can be analyzed on patient level using pseudonymized patient IDs. Practice information can be analyzed based on the pseudonymized practice ID.

Table 2.A.1: Data set details

Topic	Description	Variables
Patient	All patients enlisted in any Nivel PCD practice at any moment in defined period.	• Patient ID • Patient ID • Practice ID • Year of birth • Gender • Date in practice from • Date in practice until
Practice	All practices included in Nivel PCD at any moment in defined period.	• Practice ID • Practice type: unknown, solo, duo, health center • Practice size
Journals	All ICPC coded GP contacts per patient.	• Patient ID • Practice ID • ICPC: disease diagnosis code related to contact • ICPCepi: disease diagnosis code for episode under which contact was recorded • Date: recording date journal entry
Prescriptions	All ATC coded prescriptions of enlisted patients.	• Patient ID • Practice ID • ATC: prescription code • Prescriber: employee type • First prescriber: care provider that made the first prescription • Frequency of use • Amount prescribed • Repeat prescription: indicator for repeat prescription • Date: prescription date • End date: final date of the prescription
Episodes	All ICPC coded disease episodes of enlisted patients.	• Patient ID • Practice ID • Title: disease episode title • ICPC: diagnosis code related to episode • Epistart: start date of episode • Epistop: stop date of episode
Results	All NHG coded diagnostic test results conducted at the GP's office.	• Patient ID • Practice ID • NHG code: diagnostic test result code • Result value: outcome of diagnostic test • Result unit: measurement unit of diagnostic test • Date: testing date

Note. pSS = primary Sjögren syndrome, Nivel PCD = Nivel Primary Care Database, DIS = Dutch National Insurance Claims Database, Tp = True positives, Fn = False negatives, Fp = False positives, Tn = True negatives

3

Automated supervised text classification tool (ASTeCT): Text mining applied to psychological assessments

Wiegersma, S., Van Noije, A.J., Sools, A.M., & Veldkamp, B.P.

ASTeCT can be freely downloaded from: https://github.com/Sytskee/TextTool

Abstract

Whereas in the previous chapter we made use of a simple keyword search, supervised text classification may result in a more accurate identification of cases from text. Supervised text classification is a popular text mining application in which textual objects are assigned to a set of predefined class labels using a classification model. Supervised text classification is increasingly used by researchers to process, organize, or analyze unstructured text data more efficiently, while also improving research consistency and reproducibility. To make this method available for researchers with little to no experience in computer science, statistical modeling, or programming, this study provides a step-by-step instruction to develop new binary and multiclass classification models. The study addresses the complete text classification pipeline including model selection and evaluation using nested K-fold cross-validated parameter grid search. The elements of the pipeline (preprocessing, feature extraction, feature selection, and machine learning using support vector machines) are described and the main parameters are reviewed. In addition, an Automated Supervised Text Classification Tool (ASTeCT) is provided, which enables researchers to apply the complete procedure directly to their own text data set to generate their own classification models. The study ends with an example in which the tool was applied to a Dutch data set from psychological research practice. ASTeCT was also tested on a public English test data set for classification research, which showed that the procedure and tool can be applied to text data from different contexts and in different languages.

3.1. Introduction

The growing amount of digital psychological text data currently collected online through for example e-mental health applications and online self-help forums, makes it more and more interesting for psychological researchers to use text mining (TM) methods. By using TM methods, information can automatically be extracted from unstructured text documents (Feldman & Sanger, 2007), which may lead to new insights or help answering research questions. TM is a relatively young and interdisciplinary research area in which techniques from fields such as machine learning, information retrieval, natural language processing, and statistics are combined (Berry & Kogan, 2010; Gupta & Lehal, 2009). TM has been successfully used in a broad range of studies in the field of psychology, for example to screen for mental disorders (He et al., 2017; He et al., 2012), to study clinical dialogue contents (Angus et al., 2012), to analyze patient-caretaker communication (Cretchley et al., 2010; Wallace et al., 2013), and to predict treatment adherence (Howes et al., 2012). TM could also be used for treatment evaluation by for example analyzing patient opinions or patient records, or even in other research phases like the literature study (Abbe et al., 2016).

TM could be seen simply as an aid in processing text data (Krippendorff, 2004), enabling researchers to scale up their studies by including more cases, variables, or repeated measurements, but it has more advantages. For example, coding texts using automated algorithms keeps researchers from making premature decisions or assumptions, which could influence the research process and introduce bias (Yu et al., 2011). Moreover, since in TM information is retrieved and coded according to a previously defined set of methods and rules, the text analysis process and its outcomes are more consistent and reproducible. This can greatly improve inter-rater as well as test-retest reliability, two aspects of (qualitative) research that have led to some concern in the past (Armstrong et al., 1997; Carey et al., 1996).

Text mining adopts various methodologies to process text data and identify or explore patterns across large document collections. In general, natural language processing (NLP) is used to transform the unstructured text documents into normalized, structured input for machine learning (ML) algorithms. In NLP, computers are used to learn, "understand", or produce natural language; any spoken or written language used by humans in everyday life (Bird et al., 2009; Hirschberg & Manning, 2015). NLP techniques can be used for example to find differences in writing and speaking styles, detect emotion and sentiments, or give automated responses to human utterances by counting word frequencies and exploring patterns in texts (Bird et al., 2009). The potential benefit of NLP for the fields of psychology and psychiatry was already described in the early nineties by Garfield et al. (1992). From their review they concluded that NLP could be seen as a broad and useful research tool that enabled both clinicians and researchers to model and test new language comprehension or psychopathological theories. They found that differences between patient populations or shifts in communication over time could be easily studied on both patient and population level by coding for example syntactic, semantic, and pragmatic language characteristics.

In ML, computers are programmed to learn to optimize the parameters of a

mathematical model based on training data or previous experience. The resulting model and its learned parameters can be aimed at making predictions on future data (predictive models) or gaining knowledge from data (descriptive models) (Alpaydin, 2004). ML methods are frequently used in data-intensive sciences to learn models, make new discoveries, or characterize complex or unusual patterns from large data sets (Mitchell, 2006). The main advantages of ML are that it is highly effective, requires far less expert labor power than manual coding, and is easily applicable in different domains (Sebastiani, 2002).

Machine learning is generally divided into three categories: supervised learning, unsupervised learning, and semi-supervised learning (see James et al., 2013, for more on the different learning tasks). In supervised learning, for each input observation (input document x) the associated output variable (class label y) is known, and an algorithm is used to learn the mapping function from each input to output. This mapping function can then be applied to new input documents (x) to predict the corresponding class labels (y). Classification and regression are examples of supervised learning tasks. In unsupervised learning, the input data (x) is known whilst the output variables (y) are not. As there are no prespecified class labels to predict, unsupervised learning aims to discover and present the underlying structure of the data, for example by clustering or association rules. For semi-supervised tasks, class labels (y) are known only for a part of the input documents (x). For these data, a combination of supervised (for the labeled input) and unsupervised (for the unlabeled input) learning can be used.

When working with text data, supervised text classification is a popular approach as this can be used to simply organize documents (Sebastiani, 2002) or extract valuable knowledge (Gupta & Lehal, 2009). Supervised text classification involves assigning textual objects to a set of predefined class labels using a text classification model (Bird et al., 2009). However, developing and evaluating a new classification model is quite challenging, especially for researchers with little to no background in computer science, statistical modeling, or programming. The large amount of documentation, publications, and example scripts available can be overwhelming and often focus on specific elements, like feature extraction (Guyon & Elisseeff, 2006; Shen et al., 2006), feature selection (Forman, 2003; Yang & Pedersen, 1997), or model evaluation (Eichelberger & Sheng, 2013; Sokolova & Lapalme, 2009).

Although basic, ready to use text classification tools are available online, many of those are web-based, requiring users to send their text data over the internet, and in some cases logging the input text which may be problematic when working with privacy sensitive data. Moreover, it is not always possible to specify parameter settings and preferences, or to download the trained text classification algorithm for later use. This study therefore provides a freely available, local Automated Supervised Text Classification Tool (ASTeCT) for researchers to develop their own text classification model. This tool enables users to classify and analyze privacy sensitive data and apply the resulting model to new, future text documents. We first describe the text classification and model development process, after which the tool will be applied to existing Dutch and English data sets to demonstrate its use.

Africa, December 2013

Dear me,

I have just arrived at the office of a company in Africa and look out of the window while the sun shines at my desk. I am surrounded by books. It is calm and peaceful. I finished my study and now have this lovely job and plenty of time to enjoy the people in Africa. I became aware of how important making decisions is for living a good life. Try to suck the marrow out of life.

Myself

Figure 3.1: Example of a present-oriented imaginative letter from the "Letters from the Future" data set. From "Mapping letters from the future: Exploring narrative processes of imagining the future," by A.M. Sools, T. Tromp, and J.H. Mooren, 2015, *Journal of Health Psychology, 20*, p.360. Copyright [2015] by Sage Publishing. Reprinted with permission.

3.2. Methods

3.2.1. Corpus: Letters from the Future

ASTeCT was applied to an existing data set from psychological research practice, to illustrate the development and interpretation of binary and multiclass models for psychological assessments. The data set was previously collected online using a narrative based mental health promotion instrument called "Letters from the Future", in which participants are asked to write a letter from a particular moment and situation in the future to someone in the present. Informed consent to reuse these letters for ongoing research was obtained. Because the purpose of this example was to show how to develop and evaluate a new text classifier, only Dutch letters that were clearly assigned to one of the classes were used, resulting in a data set of 351 letters. More information on the data collection process, the composition of the participant group, and the types of letters can be found in Sools and Mooren (2012) and Sools et al. (2015). An example letter, reprinted from Sools et al. (2015) is shown in the boxed text in Figure 3.1.

As shown in Table 3.1, the data set can be split into either two (imaginative and generic), three (retrospective, prospective, and present-oriented), or six (letter types one to six) classes. For the binary classifier the imaginative versus generic classes were used; imaginative letters contain a more exhaustive imaginative account of a future situation, which is lacking or limited in the generic letters. For the multiclass classifier the retrospective, prospective, and present-oriented classes were used, because not all six letter types contained enough data to properly train and test a letter-specific multiclass classifier. Retrospective letters contain more retrospective evaluation, prospective letters contain more prospective orientation, and present-oriented letters contain neither evaluative nor orientative components.

Table 3.1: Characteristics of the example data set "Letters from the Future" used to test the described model development procedure

Characteristic	Imaginative	Generic	Total
Retrospective evaluation			
Letter type	1	4	
N(letters)	137	19	156
Mean N(words)	324	292	320
Prospective evaluation			
Letter type	2	5	
N(letters)	47	9	56
Mean N(words)	303	196	286
Present-oriented			
Letter type	3	6	
N(letters)	94	45	139
Mean N(words)	289	270	283
Total			
N(letters)	278	73	351
Mean N(words)	309	267	300

Note. The data set is split into two classes (imaginative - generic) to develop a binary classification model and into three classes (retrospective evaluation - prospective evaluation - present-oriented) to develop a multiclass classification model.
Mean N(words) = the mean number of words per text document.

3.2.2. ASTeCT

ASTeCT is a freely available supervised text classification tool that was developed as part of this dissertation. ASTeCT does not require any programming skills or prior experience with supervised text mining. The graphical user interface enables the user to specify several analytic preferences and parameter settings, for which background information is provided in this chapter. The only thing needed is a labeled data set, organized in one input folder that contains a separate subfolder with plain text files (.txt) for each class. The tool returns log files of the complete development process and a .pkl file containing the developed model, which can later be applied to new input data using the same tool (see Figure 3.2). ASTeCT is saved locally and runs from the user's hard drive, allowing researchers to securely process sensitive text data.

ASTeCT was written in Python 3.5.2 using Scikit-learn; a Python module harnessing a wide range of machine learning algorithms for (un)supervised problems (Pedregosa et al., 2011), the Natural Language Toolkit (NLTK; Loper & Bird, 2002); a Python library for natural language processing and text analysis, and NumPy; a package for scientific computing (Oliphant, 2006). ASTeCT was tested on Dutch and English data sets, but can be easily applied to data sets in any of the languages

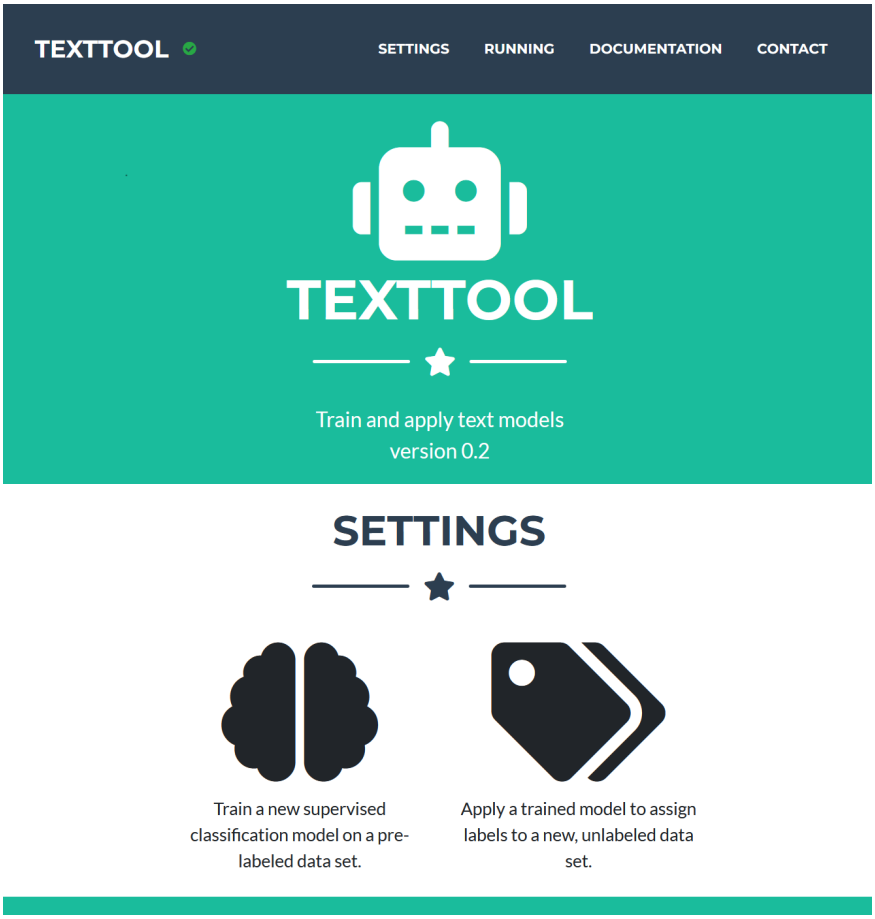


Figure 3.2: Screenshot of ASTeCT homepage: develop a new model or apply a previously trained model.

supported by the Snowball stemmer (see Normalization). This is because almost all the text processing steps from the pipeline are generic elements, not influenced by the language of the data. Only the stemming algorithm and the stop word list are language-specific, and can be defined by setting the "language" variable in the tool.

3.2.3. Model development

Figure 3.3 gives an overview of the overall procedure for developing a supervised text classification model. The two main stages in model development are model selection and model evaluation. A central element in both stages is model validation. In the model selection stage, validation is applied to compare different models and select the best combination of parameter settings as the final model. In the model evaluation stage, validation is used to assess the performance of the final model on

a new, previously unseen data set (Stone, 1974). To avoid generalization problems like overfitting and overly optimistic performance estimations, separate data sets must be used in both stages (Bird et al., 2009; Hastie et al., 2009; Rao et al., 2008).

Validation strategy

Ideally, holdout validation (Kurtz, 1948) is used, in which the data set is split into three independent samples: a training set to train the classifier, a validation set to select the best classifier, and a test set used only to evaluate the prediction capability and generalization performance of the final classifier. However, for small data sets (in this context meaning human-annotated data sets of a few hundred documents in total, with about 50-100 training documents per class (Raudys & Jain, 1991)), holdout validation has some drawbacks. First, when two parts of the data set are kept apart for model selection and evaluation, less data remain to train the model. Second, when the performance estimates are based on only one single set, the performance metrics can be misleading and biased by the way the data set was split.

A resampling method like K-fold cross-validation (K-fold CV; Breiman et al., 1984) is a suitable alternative for small data sets. In K-fold CV, the data set is split into K different, complementary training and validation sets (called folds). The model is trained K times on the altering training sets and tested on the K corresponding validation sets. Model performance is then assessed by averaging the performance scores over the K iterations, resulting in a mean CV performance score. By averaging the performance estimates, the variance problem that may occur in holdout validation is reduced, since all input documents are used for both training and validation (Gutierrez-Osuna, 2002). The main drawback of K-fold CV is that it is quite computationally expensive, as the statistical learning method is fitted K times.

ASTeCT applies a nested K-fold cross-validation strategy consisting of an inner loop for model selection and an outer loop for model evaluation. In the model selection stage, an exhaustive K-fold cross-validated grid search is conducted over a parameter grid to find the best performing model (for more on Scikit-learn's "GridSearchCV", see Buitinck et al., 2013). The parameter grid is a predefined subset of parameter values, which is specified manually based on existing literature or experience. Since the runtime of the pipeline increases exponentially with each additional parameter that is included in the grid search, it is advised to base parameter settings on existing literature as much as possible. However, when there is no clear consensus on which parameter value performs best, the parameter should be included in the grid search. During the grid search, all possible parameter combinations are fitted on the data set. The combination of parameter values that results in the highest mean cross-validated performance score is selected as the final model. In the outer loop, the selected model was trained on the complete development set and applied to the held-out test set to evaluate model generalizability.

To deal with class imbalance, stratified sampling is used in both validation stages. In stratified samples the proportions of the different classes are equal in each training, validation, and test sample, as such representing the distribution of the classes

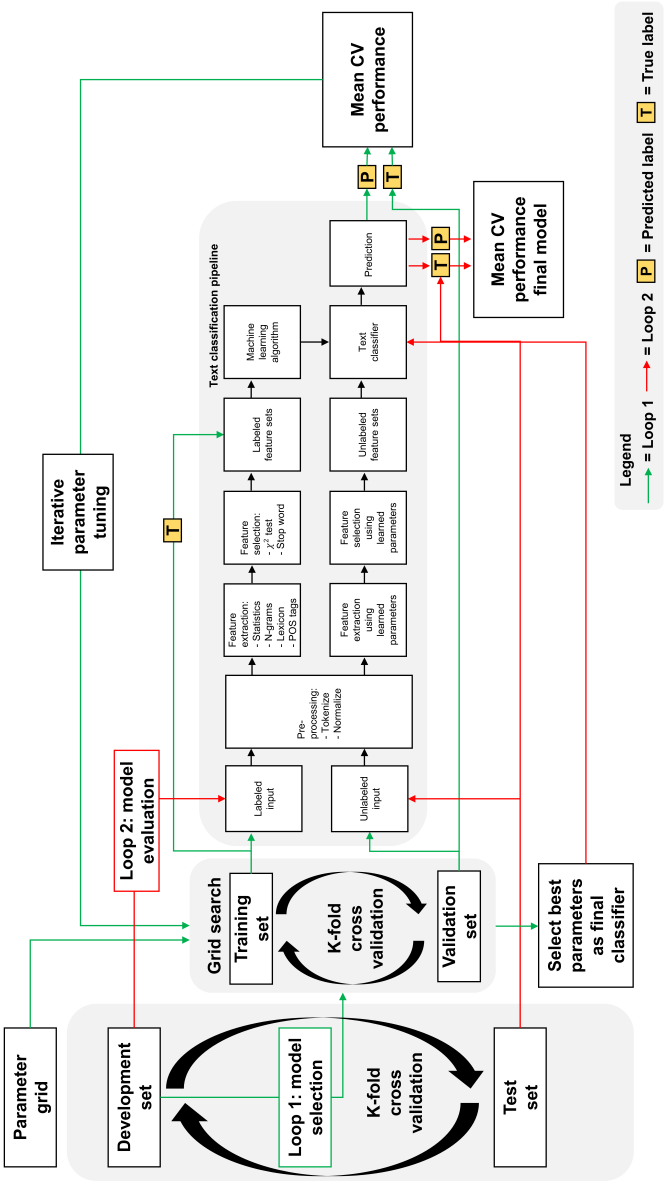


Figure 3.3: Model development procedure. In the first run the model is trained on the training set and tested on the validation set. An exhaustive K-fold cross-validated grid search is conducted to select the best model. In the second run K-fold cross-validation is used to evaluate the generalizability of the selected model.

in the complete data set. Kohavi (1995) showed that stratified cross-validation works well for selecting and evaluating supervised classification models, although stratification should be used cautiously, as it influences the cross-validation heuristics (Arlot & Celisse, 2010; Krstajic et al., 2014).

3.2.4. Text classification pipeline

The different samples are used as input for the text classification pipeline; a sequence of text processing elements in which the output of each element is the input of its succeeding element (for more on Scikit-learn's Pipeline module, see Buitinck et al., 2013). The two main steps in the classification pipeline are training and prediction (Bird et al., 2009). To train the model, preprocessing, feature extraction, and feature selection are used to transform each labeled input document to a labeled feature set. These labeled feature sets are used as input for the machine learning algorithm to train the model (also called classifier).

Preprocessing

Preprocessing is a standard and straight-forward step that is the same for all documents, regardless of which sample (training, validation, or test set) the document belongs to. The preprocessing element consists of tokenization and normalization, and results in a vocabulary.

Tokenization To be able to process and analyze the texts on word level, each input text is split into smaller parts, which is called tokenization. Texts can be split into paragraphs, which can be split further into sentences, which on their turn can be split into tokens like words, numerical expressions, punctuation marks or symbols. The splits are based for example on the use of punctuation and capital letters at the end and the beginning of sentences, or the occurrence of white spaces marking the beginning and end of each word (Perkins, 2014).

Normalization All words are then normalized by removing punctuation (like the dots in "U.S.A.", or the hyphen in "well-known"), converting all capital letters to lower case letters (called case folding), and stripping off accents. This way "U.S.A." and "USA" are converted to "usa" and the word "Dear" at the beginning of a sentence is converted to "dear". Next, a language-specific stemming algorithm is used to remove the affixes from the remaining words (Perkins, 2014). By stemming, all morphological variants of a word are brought back to one core, meaning bearing stem (Jurafsky & Martin, 2009). For example, the words "translation", "translating", "translated", and "translator" all result in the stem "translat". Normalization, and stemming in specific, is done so that all variants of an extracted word are of the same form, which makes it easier to match and compare words extracted from different documents or samples (Jurafsky & Martin, 2009). For stemming, the Snowball stemmer (Porter, 2001) is used because this stemmer provides stemming algorithms for many different languages. This enables ASTeCT to process texts in different languages. At the time of writing 14 languages were supported:

Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Norwegian, Portuguese, Romanian, Russian, Spanish, and Swedish.

Vocabulary The resulting set of normalized words is called the vocabulary (Bird et al., 2009). The vocabulary consists of word tokens and word types, where the set of word tokens refers to the total number of words that are used in a document, and the set of word types refers to the total number of different words used in a document. So, the phrase "to be or not to be" consists of four word types ("to", "be", "or", "not") and six word tokens. The number of word types in the vocabulary equals the dimensionality of the feature (or vector) space. In text classification, dimensionality can be excessively high due to the many possible unique words or word combinations (phrases) that can occur in text documents. Only a few machine learning algorithms can deal with high-dimensional feature spaces (Yang & Pedersen, 1997). In addition, using less features improves computation efficiency and leads to simpler, more robust and accurate models (Alpaydin, 2004; Forman, 2003; Guyon & Elisseeff, 2006). Therefore dimensionality is generally reduced using feature extraction, feature selection, or both.

Feature extraction

In feature extraction, the data are transformed from a high-dimensional to a lower dimensional vector space. The normalized words extracted from the text documents are converted to a structured set of features that can be used as input for the classifier. Feature extraction consists of the steps document representation and vectorization.

Document representation Documents are represented using certain document representation schemes. Well-known schemes are the bag-of-words model for unigrams (single words), and language model based representations like N-grams or N-multigrams (phrases). In the bag-of-words model, the simplest and most efficient representation, every word is treated as an independent, separate feature (Manning & Schütze, 1999; Perkins, 2014), not taking into account word ordering or constituency (Jurafsky & Martin, 2009). One of the drawbacks of the bag-of-words model is that it does not take into account the relationship between consecutive words (e.g., in the case of denial) or the context in which words that can have multiple meanings occur (Shen et al., 2006). N-gram models (sequences of N words, like bigrams (sequences of two words), trigrams (sequences of three words), and so on), or N-multigram models (variable-length sequences with a maximum of N words (see more in Shen et al., 2006) could be used to deal with this (Bekkerman & Allan, 2003; Tan et al., 2002). In addition to N-(multi)grams, text documents can also be represented by basic linguistic variables like the total number of words (tokens and types) and sentences, word and sentence length, word diversity, or word repetition (e.g., see Paap et al., 2015). However, that is beyond the scope of this chapter.

Vectorization The unigrams or N-grams that occur in the data set are called terms. Using the algebraic Vector Space Model (or Term Vector Model) of Salton (1971), a vector of features representing the terms that occur within the data set is then used to represent each document. This process is called vectorization. A text document (j) for example is represented as the vector $\overrightarrow{d_j} = (w_{1,j}, w_{2,j}, \ldots, w_{m,j})$, where each element resembles the weight for the corresponding term (term weight). The complete set of documents from the data set can be represented as a (sparse) term-by-document matrix $(A_{m,n})$. In this matrix (shown in Equation 3.1), the columns and rows represent the documents and terms respectively, and $w_{i,j}$ is the term weight; the weight of term i (of m terms) in document j (of n documents):

$$A_{m,n} = \begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,n} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,1} & w_{m,2} & \cdots & w_{m,n} \end{pmatrix} \tag{3.1}$$

Selecting suitable terms and term weights to represent the documents has a substantial influence on the effectiveness and performance of a classification model (Jurafsky & Martin, 2009; Shen et al., 2006). Terms can be weighted by:

- Term frequency ($tf_{i,j}$): the number of times a term occurs within a document (Luhn, 1957). The underlying idea is that frequently occurring terms give a better reflection of the content of a document, and thus get higher values than less frequently occurring terms.
- Inverse document frequency (idf_i): the total number of documents in the data set (N) divided by the number of documents (n) in which the term occurs ($log(N/n_i)$; Spärck Jones, 1972). The underlying idea is that terms occurring in a small number of documents are more discriminative, and thus get higher values than terms occurring in a large number of documents.
- Term frequency-inverse document frequency ($tf\text{-}idf_{i,j}$): a combination of the term frequency and inverse document frequency ($tf_{i,j} \times idf_i$; Jurafsky & Martin, 2009). The underlying idea is that terms that frequently occur in a specific document but not in the overall data set are more informative and thus get higher values.

Of these, $tf_{i,j}$ and $tf\text{-}idf_{i,j}$ are the most commonly used weights. To prevent bias towards longer documents the term frequency is usually normalized by document length, as suggested by Forman (2003). From this point on, term frequency (or $tf_{i,j}$) denotes the normalized term frequency.

Feature selection

The process of selecting the most relevant and informative features and discarding the remaining, noninformative features is called feature selection (or sometimes subset selection). The objectives of feature selection can vary from finding the minimal (Kira & Rendell, 1992) to the optimal (Narendra & Fukunaga, 1977) subset

of features that maximally contribute to the performance of the model (Alpaydin, 2004). To find the most discriminative features, stop word removal, minimal document frequency, and a Pearson's χ^2 test are used.

Stop word removal In information retrieval and NLP contexts, commonly used stop words like "I", "the", or "it" are often removed because these words do not contribute specifically to the meaning of the texts (Perkins, 2014), and carry little semantic weight (Jurafsky & Martin, 2009). However, Campbell and Pennebaker (2003) show that particles (the most commonly used words in English) and especially pronouns like "I", "you", "it", "who", and "what" indicate (changes in) writing styles and related health improvements. Moreover, when stop words are removed, it is harder to select informative phrases, as these can be expected to include one or more stop words (Jurafsky & Martin, 2009). This shows there is no clear consensus on stop word removal. In addition to stop words, sometimes words that occur less than a certain number of times or in less than a certain number of different training documents are removed as well, to avoid needlessly large feature vectors (Joachims, 1998).

Pearson's chi-squared test An efficient and statistically robust method for feature selection is the filter method (Guyon & Elisseeff, 2006). In the filter method, each feature is scored independently based on its occurrence in positive and negative training documents using a feature selection metric (Forman, 2003). Based on this metric, the features are ranked and a subset of features is selected using a certain cut-off point in the ranking (Suman & Thirumagal, 2013). A common feature selection metric for text classification is Pearson's chi-squared test (χ^2), a highly efficient univariate statistical hypothesis test that measures the independence between corpora by comparing the observed and expected feature occurrences in each class (Forman, 2003). It has been successfully used in models for example for document classification (Oakes et al., 2001), identification of differences between male and female vocabulary characteristics (Rayson et al., 1997), and assessment of patients' self-narratives (He et al., 2017). Following Oakes et al. (2001), 2 x 2 contingency tables (see Table 3.2) are compiled for each feature type, where C_{pos} = the positive class, C_{neg} = the negative class, and F = a unique feature type in the data set. The values in each cell (a, b, c, d) are called the observed frequencies (O).

Next, the expected frequencies (E) are calculated for each cell in the contingency table based on the marginal probabilities, using the formula:

$$E_{i,j} = \frac{column_i\ total \times row_j\ total}{grand\ total} \tag{3.2}$$

Finally, the χ^2 score is calculated separately for each feature by summing the differences between the observed and expected frequencies for each cell in the table using the formula:

$$\chi^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{3.3}$$

Table 3.2: Contingency table with observed frequencies (O) for feature F

	Class	
Feature	Positive (C_{pos})	Negative (C_{neg})
F	a	b
$\neg F$	c	d

Note. Observed frequencies for feature type F versus the other feature types in the training set for the positive (C_{pos}) and negative (C_{neg}) class. a = number of times feature F occurs in class C_{pos}; b = number of times feature F occurs in class C_{neg}; c = total number of features (tokens) in class C_{pos} that are not F; d = total number of features (tokens) in class C_{neg} that are not F.

The features are then ranked in descending order based on their χ^2 scores, after which the most informative (discriminative) features, with the highest scores, are selected for the classification model.

Some studies, e.g., Manning and Schütze (1999), suggest removing features that occur less than five times in the training set to ensure reliability of the χ^2 calculation. However, since this option is not offered by Scikit-learn's standard "GridSearchCV" implementation for nested K-fold cross-validation, scarcely used features were not removed before calculating the χ^2 scores.

Supervised learning

In order to predict the class labels for new input documents, the classifier first needs to learn the boundaries that separate the input documents of each class from those of the other class(es). These boundaries are used to develop a classifier (h) that approximates the unknown target function that designates documents from the training set to their corresponding classes as much as possible (Alpaydin, 2004; Sebastiani, 2002). Using the vector representation described before, each input document (j) from the training set is represented as the vector $\overrightarrow{d_j}$ with input features $(w_{1,j}, w_{2,j}, \ldots, w_{m,j})$ and document label (r) denoting the class, as in Equation 3.4:

$$r = \begin{cases} 1 \text{ if } \overrightarrow{d_j} \text{ is a positive example} \\ 0 \text{ if } \overrightarrow{d_j} \text{ is a negative example} \end{cases} \tag{3.4}$$

Combining the input feature vector and document label, each of the n documents in the training set is converted to a labeled feature set represented by the ordered pair $(\overrightarrow{d_j}, r)$. Equation 3.5 shows the training set (S), with t indexing the different documents and n the total number of documents in the training set

$$S = \left\{\overrightarrow{d_j}^t, r^t\right\}_{t=1}^{n} \tag{3.5}$$

The classifier (h) can make the following predictions for an input document (j):

$$h(\overrightarrow{d_j}) = \begin{cases} 1 \text{ if } h \text{ classifies } \overrightarrow{d_j} \text{ as a positive example} \\ 0 \text{ if } h \text{ classifies } \overrightarrow{d_j} \text{ as a negative example} \end{cases} \tag{3.6}$$

For multiclass classification with K classes (denoted C_i, where $i = 1, \ldots, K$), the training set is denoted by

$$S = \left\{\overrightarrow{d_j}^t, r^t\right\}_{t=1}^{n} \tag{3.7}$$

where r has K classes and

$$r_i^t = \begin{cases} 1 \text{ if } \overrightarrow{d_j}^t \in C_i \\ 0 \text{ if } \overrightarrow{d_j}^t \in C_j, j \neq i \end{cases} \tag{3.8}$$

The input documents from the training set that belong to class C_i are positive examples, whereas the rest of the input documents from the other classes are negative examples for the classifier h_i (Alpaydin, 2004; Sebastiani, 2002). As such, for a K-class classification task, K classifiers need to be learned:

$$h_i(\overrightarrow{d_j}^t) = \begin{cases} 1 \text{ if } \overrightarrow{d_j}^t \in C_i \\ 0 \text{ if } \overrightarrow{d_j}^t \in C_j, j \neq i \end{cases} \tag{3.9}$$

Algorithm There are many machine learning algorithms for text classification, like naive bayes (NB), support vector machines (SVM), or decision trees (DT). For a thorough description of these methods is referred to Hastie et al. (2009). The support vector machine (SVM; Vapnik, 1995) is used in the pipeline, as this is perceived as one of the best performing and most robust classification algorithms (Joachims, 1998). Moreover, the SVM algorithm is found to deal well with high-dimensional data (Joachims, 1998) and to have less problems handling imbalanced training data than other learning algorithms such as the NB (Rennie, 2001).

The idea behind the SVM algorithm is to find a hyperplane that perfectly separates the documents from the training sample according to their class labels; the optimal separating hyperplane. Following Alpaydin (2004), using the labels -1 / +1 and the training sample $S = \left\{\overrightarrow{d_j}^t, r^t\right\}$, with $r^t = +1$ if $\overrightarrow{d_j}^t \in C_1$ and $r^t = -1$ if $\overrightarrow{d_j}^t \in C_2$, the aim is to find the weight factor (w) and the threshold (w_0) for which $w^T\overrightarrow{d_j}^t + w_0 \geq +1$ for $r^t = +1$, and $w^T\overrightarrow{d_j}^t + w_0 \leq -1$ for $r^t = -1$. Or more compact: $r^t(w^T\overrightarrow{d_j}^t + w_0) \geq +1$. The optimal separating hyperplane is the hyperplane that maximizes the margin; the distance between the hyperplane and the closest data points on either side. As described by Alpaydin (2004), this margin is maximized for best generalization by solving the quadratic optimization problem:

$$\min \frac{1}{2}||w||^2 \text{ subject to } r^t(w^T\overrightarrow{d_j}^t + w_0) \geq +1, \forall t \tag{3.10}$$

For the model to generalize well to new data, at least two SVM hyperparameters ("higher level" model parameters) need to be set: a kernel parameter (γ) and a regularization parameter (C) (Duan et al., 2003). The kernel parameter (the degree of the polynomial kernel and the width for the Gaussian kernel) controls the flexibility of the classifier (Ben-Hur & Weston, 2010). The linear kernel, which is the lowest degree polynomial, is generally used in text classification, since most text classification problems are linearly separable (Joachims, 1998) and the linear kernel was found to perform better than (Yang & Liu, 1999) or equal to (Rennie, 2001) nonlinear kernels. For the regularization parameter, which controls the trade-off between a minimal training error and a minimal testing error (Duan et al., 2003), a value equal or close to the number of classes to be predicted is found to perform well (Mattera & Haykin, 1999).

Decomposition strategy SVMs were originally intended for binary classification tasks. Multiclass (K-class) classification tasks are therefore typically executed as K binary classification tasks using a decomposition strategy (Lorena et al., 2008), also known as binarization (Galar et al., 2011). Two widely applied decomposition strategies are the One-against-One (O-a-O, also called One-versus-One) and the One-against-All (O-a-A, also called One-versus-Rest) strategy. O-a-O uses $\binom{K}{2}$ pairwise classifiers to distinguish between each pair of classes (Galar et al., 2011; Hastie et al., 2009) whereas O-a-A uses K classifiers to distinguish between each single class and the remaining classes (Galar et al., 2011; Hastie et al., 2009). ASTeCT applies the O-a-A approach, which is the most commonly used due to its computational efficiency and interpretability, using the "Linear Support Vector Classifier" based on the LIBLINEAR library (Fan et al., 2008).

Prediction

During prediction, the same preprocessing, feature extraction, and feature selection steps are used to transform each unlabeled input document to an unlabeled feature set. The trained model is then used to predict class labels for each unlabeled input document. The differences between the true and predicted class labels are used to assess model performance (see Figure 3.3).

Following Alpaydin (2004), for binary classification problems the labels are predicted by calculating the function $g(\overrightarrow{d_j}) = w^T\overrightarrow{d_j^t} + w_0$. Depending on the sign of $g(\overrightarrow{d_j})$, a document is assigned to C_1 if $g(\overrightarrow{d_j}) > c$, where c can be any constant threshold value, and to label C_2 otherwise. For multiclass (K-class) classification problems, K binary SVMs $(g_i(\overrightarrow{d_j}), i = 1, \ldots, K)$ are learned. All $g_i(\overrightarrow{d_j})$ are calculated and the maximum is selected as the predicted label. The predicted labels are then compared to the true class labels to assess the model performance.

Performance metrics

To assess model performance, a confusion matrix (Table 3.3) is generated where the columns represent the instances for the predicted classes and the rows represent the instances for the true classes. The cells on the diagonal show the number

Table 3.3: Confusion matrix to assess model performance

	Predicted class	
True class	Positive (C_{pos})	Negative (C_{neg})
Positive (C_{pos})	**True positive** (tp)	False negative (fn)
Negative (C_{neg})	False positive (fp)	**True negative** (tn)

Note. Comparison of true (rows) and predicted (columns) class labels for the classes C_{pos} and C_{neg}. The values on the diagonal (in boldface) show the correctly predicted class labels.

of correctly predicted class labels, whereas the off-diagonal cells show the number of errors (Bird et al., 2009). The number of correctly predicted labels (true positives and true negatives) and errors (false positives and false negatives) from the confusion matrix are used to calculate the performance metrics accuracy, precision, recall, and F_1-score (see Table 3.4 for a description). Of these, accuracy and F_1-score are the most commonly used metrics to evaluate the performance of supervised text classification models, although F_1-score (and weighted F_1-score for the multiclass classifier) is preferred when working with imbalanced data sets. The metrics are calculated separately for each individual class C_i based on that class' counts (true positives (tp_i), true negatives (tn_i), false positives (fp_i), and false negatives (fn_i). The output scores for all metrics are between 0 (worst) and 1 (best). For binary classification generally only the performance of the positive class is reported.

As described under Decomposition strategy, K-class classification tasks are executed as K binary classification tasks. The overall performance score for each metric is then computed by averaging the performance scores of all K binary classifiers. There are several different averaging methods, of which micro- and macro-averaging are the most commonly used. In micro-averaging, the average of the K binary performance scores is computed giving an equal weight to each document (class instance). This way, micro-averaged scores are dominated by the frequently occurring classes as these contain more instances than classes with lower occurrence frequencies (Yang, 1999). In macro-averaging, the average of the K binary performance scores is computed giving an equal weight to each class. As a result, the (typically low) scores for infrequent classes count just as much as the scores for the (typically higher scoring) frequent class(es) (Yang, 1999). The micro-averaged scores for each performance metric (M) are computed by:

$$M_{micro} = M(\sum_{i=1}^{K} tp_i, \sum_{i=1}^{K} fp_i, \sum_{i=1}^{K} fn_i, \sum_{i=1}^{K} tn_i) \tag{3.11}$$

Table 3.4: Model performance metrics and functions

Metric (M)	Description	Function
Accuracy	Average per-class effectiveness of the classifier	$\frac{tp_i+tn_i}{tp_i+fn_i+fp_i+tn_i}$
Precision	Average per-class agreement of the true and predicted class labels	$\frac{tp_i}{tp_i+fp_i}$
Recall	Average per-class effectiveness of the classifier to identify class labels	$\frac{tp_i}{tp_i+fn_i}$
F_1-score	Harmonic mean of precision and recall	$2 \times \frac{precision \times recall}{precision+recall}$

Note. The true positives (tp), false positives (fp), false negatives (fn), and true negatives (tn) from the confusion matrix in Table 3.3 are used to calculate scores for the model performance metrics (Accuracy, Precision, Recall, and F_1-score) for each class C_i.

and the macro-averaged scores for each performance metric (M) by:

$$M_{macro} = \frac{1}{K}\sum_{i=1}^{K} M(tp_i, fp_i, fn_i, tn_i) \tag{3.12}$$

where M = the concerning performance metric, K = the total number of classes, and $i = 1, \ldots, K$. The macro-averaged scores can be altered to deal with class imbalance by weighting the scores for each class by the occurrence frequency of the concerning class in the data set, this is called the weighted average.

3.2.5. Model selection

As stated before, the complete pipeline is run twice; once for model selection and once for model evaluation. In the model selection stage (the inner loop) only the development set is used, which is split into training and validation sets. Since K-fold cross-validation is used, this is an iterative process in which the model is trained K times on altering training sets, and class labels are predicted K times for the corresponding validation sets. The mean cross-validated performance is then assessed by calculating the performance scores for all K iterations and then taking the mean. The value for K can be specified in the tool, but is set to 5 by default in both the inner and the outer loop (see Figure 4.2 for a schematic representation).

Parameter grid search

To select the best performing model, different models and model parameters are compared. Section 3.2.4 described various text processing settings and parameters that can strongly influence the performance of the classifier. Although some parameter settings could be based on existing literature, for most parameters there is no clear consensus in the literature on which values generate the best performance.

This could be because the effects of specific parameter settings can be data dependent and can strongly interact with other parameter settings. Therefore, these parameters are included in an exhaustive parameter grid search, in which all possible combinations of parameters are fitted on the data set. The grid search can be guided by any of the four described performance metrics. The parameter combination that generates the highest mean cross-validated performance score is selected as the final model.

Except for preprocessing, which is a standard step, for all elements of the pipeline parameters are included in the grid search. For feature extraction these are stop word removal, representation schemes and term weights. Stop words are either retained or removed using a language-specific stop word list included in the NLTK library (available in multiple languages, see Normalization). In addition, only terms occurring in at least x different training documents are selected, with the values for x ranging from 1-3. To find which representation scheme generates the best results, different N-gram ranges (unigrams, bigrams, trigrams and 3-multigrams) are compared. Finally, the two most commonly used term weights ($tf_{i,j}$ and $tf\text{-}idf_{i,j}$) are compared. For feature selection, only the k features with the highest χ^2 scores are selected. The best value for k (the cut-off point) is determined by comparing values ranging from 10-500, increasing with steps of 20 features, or 'all'. For machine learning, several C values of around and further off the number of classes (1, 2, 3, 100, and 1,000) are compared. In addition, a class weight parameter is included in the grid search to test whether adjusting class weights to be inversely proportional to the class frequencies in the training data performs better than using no class weights. All parameters settings can be configured in ASTeCT, as displayed in Figure 3.4. Table 3.5 gives an overview of all the parameters and the corresponding subset of parameter values that are included in the grid search.

3.2.6. Model evaluation

In the model evaluation stage (the outer loop) the development set and the test set are used. As for model selection, this is an iterative process using K-fold cross-validation, splitting the complete data set into K (5 by default) folds and alternately defining K-1 folds as the development set for model selection and setting aside one fold as a test set for assessing the final model performance and generalization. The final model, with the best performing combination of parameter settings, is fit one final time on the complete development set to take full advantage of all the available training data and is then used to predict the class labels for the test set to evaluate model generalizability. This shows how well a model trained and validated on the labeled input data predicts the correct labels for new, future data (Alpaydin, 2004).

3.2.7. Save and apply final model

The output files and the final model are saved locally to the user's hard drive. A second feature of ASTeCT is that researchers can upload a previously developed model together with a new, unlabeled data set in order to predict class labels for each new input document.

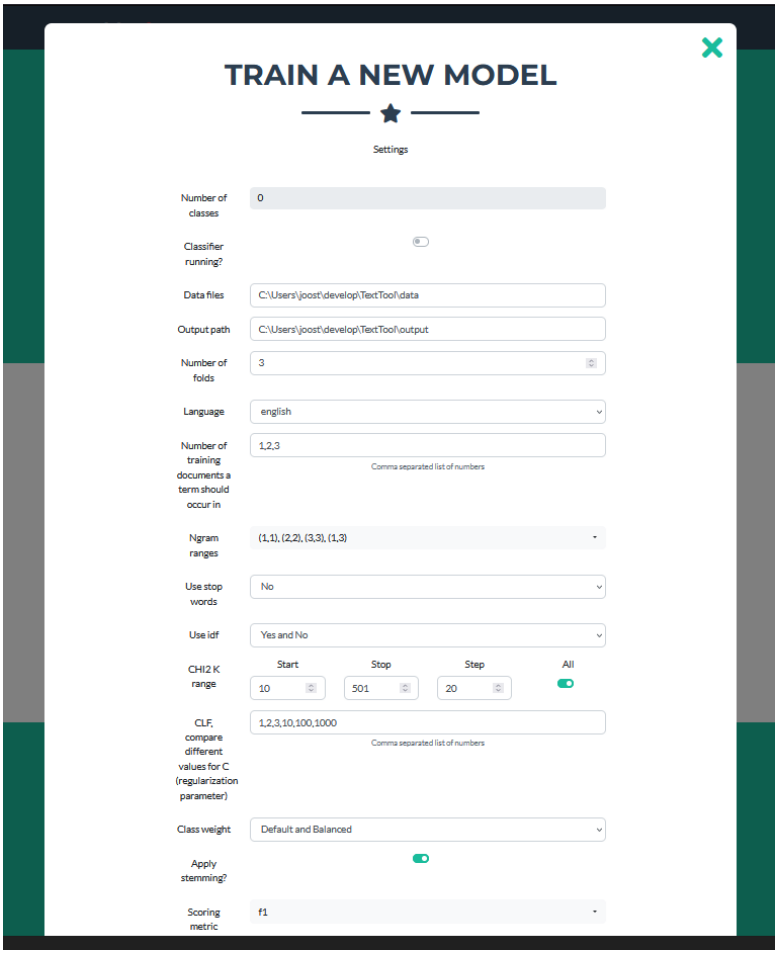


Figure 3.4: Screenshot of exhaustive parameter grid search settings to be configured in ASTeCT.

3.3. Results

To test its performance, ASTeCT was first applied to the Letters from the Future data set to develop a binary classifier (to distinguish between imaginative and generic letters) and a multiclass classifier (to distinguish between retrospective, prospective, and present-oriented letters). Second, the tool was applied to three classes of a standard English test data set in order to test the tool's performance on a large, balanced data set in a different language.

3.3.1. Binary classifier

In the exhaustive grid search in the inner 5-fold cross-validation loop, all possible combinations of parameter values listed in Parameter grid search were compared to find the model with the highest performance score. The parameter combination

Table 3.5: Parameters and parameter values used in the grid search

Parameter	Description
Remove stop words	Do/do not remove (Dutch) stop words
Minimal x documents	Minimal number of documents a term should occur in, with x ranging from 1-3
Representation schemes	Test uni-, bi-, tri-, and N-multigrams ranging from 1-3
Term weights	Use $tf_{i,j}$ or $tf\text{-}idf_{i,j}$ weights
Select k best features	Select k features with highest χ^2 score, with k ranging from 10-500 with steps of 20 features, or all
Regularization parameter C	Compare values 1, 2, 3, 10, 100, 1000 for C
Class weights	Weighted ('balanced') versus non-weighted classes to account for class imbalance

Note. Parameters and defined subset of parameter values that are included in the exhaustive cross-validated grid search for model selection.

that generated the highest cross-validated F_1-score on the validation set is shown in Table 3.6. The best results on the validation set were generated by the Linear Support Vector Classifier with regularization parameter C = 2 (F_1-score = 0.507). Adjusting class weights to be inversely proportional to the class frequencies in the training set was found to perform better than using no class weights. Moreover, removing stop words did not result in a higher mean cross-validated F_1-score. The grid search further showed that documents could best be represented by unigrams, using tf term weights.

Based on the outcomes of the grid search, 230 unigrams that occurred in at least one document in the training set were included in the binary model. The ten most informative features for each class are shown in Table 3.7. The table shows that the χ^2 values were low and close to each other. In fact, for none of the features the χ^2 value was high enough to indicate significant differences in occurrence between the imaginative and generic letters (a χ^2 value > 3.84 is required to indicate significant differences between two classes ($p \leq .05$, df = 1)). However, as stated by Suman and Thirumagal (2013) and He and Veldkamp (2012), the χ^2 values are used only to rank the features from most to least informative to choose the best cut-off point. As such, the significance of the χ^2 test is not taken into account.

The true and predicted class labels are shown in the confusion matrix in Table 3.8. The cells on the diagonal show that the classifier predicted the correct class label for 58 of the 70 test documents, leading to an accuracy score of 0.83. The classifier did a good job on defining the imaginative letters, correctly labeling 49 of the 56 imaginative letters (87.5%) and mislabeling only seven as generic. The classifier had a bit more difficulty identifying the generic letters, labeling nine of the 14 generic letters correctly (64.3%), and mislabeling five generic letters as

Table 3.6: Best parameter values for binary and multiclass classification models

	Best parameter value	
Parameter	Binary model	Multiclass model
Remove stop words	no	no
Minimal x documents	1	2
Representation schemes	unigrams (1, 1)	unigrams (1, 1)
Term weights	$tf_{i,j}$	$tf_{i,j}$
Select k best features	230	70
Regularization parameter C	2	1000
Class weights	balanced	balanced

Note. The combination of parameter values that generated the highest mean cross-validated performance score during the grid search. These parameter values are selected for the final binary and multiclass classification models.

imaginative.

Table 3.9 shows the performance scores of the final model. The performance metrics show that, although the overall accuracy score of the binary classifier was very high (accuracy = 0.83), there was a big difference in performance for both separate classes (for the generic letters F_1-score = 0.60, for the imaginative letters F_1-score = 0.89). A reason for this could be that, with only 59 generic letters in the development set and 14 in the test set, there was not enough information available to sufficiently train the classifier on recognizing generic letters. The weighted F_1-score over the test set was 0.600. This is the estimated generalization performance, the performance that can be expected when the final model would be applied to new data sets in the future.

3.3.2. Multiclass classifier

As for the binary classifier, the grid search for the multiclass classifier was guided by the weighted F_1-score. The parameter combination that generated the highest cross-validated weighted F_1-score is shown in Table 3.6. For the multiclass model this was the Linear Support Vector Classifier with regularization parameter C = 1000 (weighted F_1-score = 0.567). As for the binary model, adjusting class weights to be inversely proportional to the class frequencies in the training data was found to perform better than using no class weights, and removing stop words did not result in a higher mean cross-validated weighted F_1-score. The grid search further showed that documents could best be represented by unigrams, using tf term weights.

Only the 70 best features that occurred in at least two different training documents were included in the model. The ten most informative features for each class are shown in Table 3.10. As for the binary classifier, the χ^2 values were low (for none of the features the χ^2 value indicateed significant differences in occurrence between the retrospective, prospective, and present-oriented letters) and close to

Table 3.7: Ten most informative features per class for the binary classification model

Feature (Dutch)	Feature (translation)	χ^2 value	Feature counts	
			Imaginative	Generic
Je	You	1.6928	2506	821
We	We	1.0184	415	39
Verjaardag	Birthday	0.9062	2	7
Rad	To guess	0.8102	4	10
In	In	0.7640	1144	190
Tentamenphas	Exam phase	0.6804	0	6
Jezelf	Yourself	0.6543	112	55
Zal	Shall	0.6152	99	54
Persoonsnaam[a]	Person name[a]	0.5895	0	7
Trainer	Trainer	0.5719	0	7
Hop	To hope	0.5300	48	21
Benieuwd	Curious	0.5044	7	6
Feestdag	Holiday	0.4853	0	2
Hart	Heart	0.4385	20	11
Accepter	To accept	0.4356	5	8
Een	A/an	0.4219	1491	297
Up	Up	0.4089	0	3
Miss	To miss	0.4083	5	4
Masterdiploma	Master's degree	0.4011	0	1
Persoonsnaam[a]	Person name[a]	0.3799	0	3

Note. The ten features with the highest χ^2 scores for the imaginative and generic letters. The first column contains the stemmed features in Dutch, the second column contains the corresponding translations (unstemmed) in English.

[a] Place and person names were anonymized.

Table 3.8: Confusion matrix final binary classification model

	Predicted letter type	
True letter type	Imaginative	Generic
Imaginative	**49**	7
Generic	5	**9**

Note. Comparison of true (rows) and predicted (columns) letter types for the binary model. The values on the diagonal (in boldface) show the number of correctly predicted letter types.

Table 3.9: Performance scores final binary and multiclass models

	Precision	Recall	F_1-score	Accuracy[a]	N(letters) test set
Binary model					
Imaginative	0.91	0.88	0.89		56
Generic	0.56	0.64	0.60		14
Weighted average, total[b]	0.84	0.83	0.83	0.83	70
Multiclass model					
Retrospective	0.65	0.71	0.68		31
Prospective	0.29	0.45	0.36		11
Present-oriented	0.68	0.46	0.55		28
Weighted average, total[b]	0.61	0.57	0.58	0.57	70

Note. Per class and average performance scores of the final binary and multiclass models on the test set.

[a] Accuracy is the overall accuracy of the classifier, and can therefore not be calculated for each separate class.

[b] Total N(letters) for complete test set.

each other.

Table 3.11 shows the confusion matrix with the true and predicted class labels. The cells on the diagonal show that the multiclass classifier predicted the correct class label for 40 of the 70 test documents. This leads to an accuracy score of 0.57. Overall, 22 of the 31 retrospective letters (71.0%), five of the 11 prospective letters (45.5%), and 13 of the 28 present-oriented letters (46.4%) were assigned to the correct class.

Table 3.9 shows the performance metrics of the final model. The table contains the performance metrics for each individual class, as well as the weighted average. The table shows that the multiclass model classified the retrospective and present-oriented letters moderately well (with F_1-score = 0.68 for the retrospective letters and F_1-score = 0.55 for the present-oriented letters), but poorly identified the prospective letters (with F_1-score = 0.36). As for the binary classifier, it seems that in order to develop a classifier that performs satisfactory for all classes, a minimum amount of documents has to be present in each separate class. The weighted F_1-score over the test set was 0.577. This is the estimated generalization performance, the performance that can be expected when the final model would be applied to new data sets in the future.

Table 3.10: Ten most informative features per class for the multiclass classification model

Feature (Dutch)	Feature (translation)	χ^2 value	Feature counts		
			Retro-spective	Pro-spective	Present-oriented
Was	Was	1.1848	283	59	121
Hop	To hope	1.0438	30	32	23
Ik	I	0.9558	1370	307	977
Will	To want	0.9187	15	22	21
Persoonsnaam[a]	Person name[a]	0.8371	0	6	0
Altijd	Always	0.8320	173	19	81
Wild	Wild	0.8166	60	10	12
Is	Is	0.7597	464	175	522
Jullie	You	0.7582	97	67	116
Mijn	Mine	0.7114	454	103	257
Trouw	Marry	0.6784	5	7	2
Je	Your	0.6211	1	6	2
Plaatsnaam[a]	Placename[a]	0.6164	5	13	5
Verjaardag	Birthday	0.6017	0	4	4
Kilometer	Kilometer	0.5981	0	3	0
Wandel	To walk	0.5512	0	7	4
Kon	Could	0.5444	60	8	19
Generatie	Generation	0.5380	4	10	1
Geregeld	Regularly	0.5359	2	5	0
Uitgekom	Came through	0.5343	10	8	3
We	We	0.5215	182	103	188
Lijkt	Seems	0.5071	8	1	21
Geen	None	0.5035	62	13	81
Mezelf	Myself	0.4938	31	3	6
Mens	Human	0.4851	81	17	98
Gelop	Walked	0.4772	6	6	0
Plaatsnaam[a]	Placename[a]	0.4669	18	0	1
Water	Water	0.4503	0	1	12
Auto	Car	0.4192	4	3	23
Miss	To miss	0.4078	2	3	4

Note. The ten features with the highest χ^2 scores for the retrospective, prospective and present-oriented letters. The first column contains the stemmed features in Dutch, the second column contains the corresponding translations (unstemmed) in English.

[a] Place and person names were anonymized.

Table 3.11: Confusion matrix final multiclass classification model

	Predicted letter type		
True letter type	Retrospective	Prospective	Present-oriented
Retrospective	**22**	5	4
Prospective	4	**5**	2
Present-oriented	8	7	**13**

Note. Comparison of true (rows) and predicted (columns) letter types for the multiclass model. The values on the diagonal (in boldface) show the number of correctly predicted letter types.

3.3.3. English classifier

To test whether ASTeCT functioned correctly using English input data, the tool was applied to a standard test collection for text classification research during development. This is a common way to test natural language processing algorithms, scripts, or tools. A widely used public test collection is the English "20 Newsgroups" data set originally collected by Lang (1995), which consists of in total approximately 20,000 Usenet posts on twenty different topics (classes), so around 1,000 documents per class. This data set was fetched directly by Scikit-learn using the "sklearn.datasets.fetch_20newsgroups" function. A multiclass classifier was developed by applying the pipeline to three of the twenty classes (namely "Atheism", "Graphics", and "Religion"). The selected classifier resulted in high performance scores for each class (overall accuracy score = 0.83, per class F_1-scores of 0.82, 0.92, and 0.71 respectively). This indicated that the described pipeline and corresponding tool functioned appropriately.

3.4. Discussion

This study provides a step-by-step description and tool for anyone who wants to start using supervised text classification models. The concept of text classification and the model development process (including model validation, selection, and evaluation strategies) are addressed and the main model parameters are reviewed. The provided tool, ASTeCT, is ready-to-use and enables researchers to develop their own binary and multiclass text classification models without any technical programming skills.

To illustrate how supervised text classification can contribute to psychological research, ASTeCT was applied to the Dutch "Letters from the Future" data set from psychological research practice. In previous studies by Sools and Mooren (2012) and Sools et al. (2015), this data set was classified into different letter types by manually coding each letter on sentence level, clustering narrative processes, and comparing patterns in letter components. Sools et al. (2015) suggest that these narrative processes and patterns could relate to the writer's (mental) health and well-being, although further analysis of the letters is required to investigate this.

There are several ways in which supervised classification models can be used to advance the study of these letters. First of all, automatically classifying the letters into letter types without manual intervention saves a lot of time and expert labor power, which makes it possible to process a larger quantity of letters at once. Secondly, the content of the different letter types can be studied more extensively and efficiently by automatically recognizing the letter components or narrative processes operating within the letters.

The example application showed the development of a binary and a multiclass classifier to distinguish between two (imaginative - generic) and three (retrospective - prospective - present-oriented) overarching letter types. The results showed that although the binary classifier performed quite well, the multiclass classifier had more difficulty assigning the letters to the correct classes. This was particularly the case for the prospective letters. This could be because the prospective letters were underrepresented in the data set or because the data set did not contain sufficient information to properly train the classifier for all classes. For text mining and machine learning applications a data set of 351 documents, or classes of only 45 training documents, are considered very small. The features in the used data set were not discriminative enough to distinguish between the multiple classes. It can be expected that adding more training data would improve the classification performance of both models.

ASTeCT was also applied to the English "20 Newsgroups" data set, a standard test set for text classification research. A multiclass classifier was developed to distinguish between three classes of Usenet messages (atheism - graphics - religion), which resulted in high performance scores. These results not only indicate that ASTeCT does work well on larger and more balanced data sets, it also shows that the tool can be easily applied to data sets in other languages (in this case English). Thanks to the way ASTeCT is organized, using a pipeline with an integrated cross-validated grid search for model selection, it can be easily applied to data sets in any of the supported languages (see Normalization). This is because almost all the text processing steps from the pipeline are generic elements, not influenced by the language of the data. The only language-specific elements are the stemming algorithm and the stop word list, which can be set by defining the "language" variable in the tool.

The two example applications showed that the described method and tool can be easily applied to unstructured text data sets in various languages and from different contexts. For psychological research, this could involve data regarding treatment content (e.g., ego documents, therapy session transcripts, or patient diaries), treatment administration (e.g., medical records, doctor notes, or patient feedback), or literature (e.g., scientific papers or research reports). Supervised text classification is not only a very efficient way to process such data, it can also improve research consistency and reproducibility because information is retrieved or coded according to a predefined, fixed set of rules (Yu et al., 2011). Using such a standardized information extraction process enables researchers to objectively assess differences between patient groups or shifts over longer periods of time. This way, text classification models could be used for example to monitor progression or to provide

information on treatment impact or adherence. As such, classification models could supplement or potentially replace patient reported outcomes like questionnaires, which are more indirect and are sometimes perceived as burdensome or interfering (Valderas et al., 2008).

All in all, supervised text classification can be very beneficial for (psychological) research and practice. However, getting a solid grasp of the complete text classification and model development process can be quite challenging, especially for researchers with limited to no experience in computer programming. This study provides researchers with the basic knowledge of supervised text classification and model development that is required to develop their own binary and multiclass classifiers. Using the provided tool ASTeCT, the described procedure can be easily applied to any given data set, so that interested researchers can directly use this method in their own (research) practice.

3.5. References

Abbe, A., Grouin, C., Zweigenbaum, P., & Falissard, B. (2016). Text mining applications in psychiatry: A systematic literature review. *International Journal of Methods in Psychiatric Research*, *25*, 86–100. https://doi.org/10.1002/mpr.1481

Alpaydin, E. (2004). *Introduction to machine learning*. MIT Press.

Angus, D., Watson, B., Smith, A., Gallois, C., & Wiles, J. (2012). Visualising conversation structure across time: Insights into effective doctor-patient consultations. *PLoS ONE*, *7*(6), 1–12. https://doi.org/10.1371/journal.pone.0038014

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79. https://doi.org/10.1214/09-SS054

Armstrong, D., Gosling, A., Weinman, J., & Marteau, T. (1997). The place of inter-rater reliability in qualitative research: An empirical study. *Sociology*, *31*, 597–606. https://doi.org/10.1177/0038038597031003015

Bekkerman, R., & Allan, J. (2003). *Using bigrams in text categorization* (Report IR-408). Center for Intelligent Information Retrieval, UMass. Amherst, MA, University of Massachusetts. http://ist.psu.edu

Ben-Hur, A., & Weston, J. (2010). A user's guide to support vector machines. In O. Carugo & F. Eisenhaber (Eds.), *Data mining techniques for the life sciences* (pp. 223–239). Humana Press.

Berry, M. W., & Kogan, J. (2010). *Text mining: Applications and theory*. John Wiley & Sons, Ltd.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'reilly Media, Inc.

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth International Group.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Müller, A. C., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). API design for machine learning software: Experiences from the Scikit-learn project. *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 1–15. http://arxiv.org/abs/1309.0238

Campbell, R., & Pennebaker, J. (2003). The secret life of pronouns: Flexibility in writing style and physical health. *Psychological Science*, *14*(1), 60–65. https://doi.org/10.1111/1467-9280.01419

Carey, J. W., Morgan, M., & Oxtoby, M. J. (1996). Intercoder agreement in analysis of responses to open-ended interview questions: Examples from tuberculosis research. *Cultural Anthropology Methods*, *8*(3), 1–5. http://www.cdc.gov

Cretchley, J., Gallois, C., Chenery, H., & Smith, A. (2010). Conversations between carers and people with schizophrenia: A qualitative analysis using Leximancer. *Qualitative Health Research*, *20*, 1611–1628. https://doi.org/10.1177/1049732310378297

Duan, K., Keerthi, S., & Poo, A. (2003). Evaluation of simple performance measures for tuning SVM hyper parameters. *Neurocomputing*, *51*, 41–59. https://doi.org/10.1016/S0925-2312(02)00601-X

Eichelberger, R. K., & Sheng, V. S. (2013). Does one-against-all or one-against-one improve the performance of multiclass classifications? *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, 1609–1610. http://www.aaai.org

Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, *9*, 1871–1874. http://www.jmlr.org/

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, *3*, 1289–1305. http://www.jmlr.org/

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*, *44*, 1761–1776. https://doi.org/10.1016/j.patcog.2011.01.017

Garfield, D. A. S., Rapp, C., & Evens, M. (1992). Natural language processing in psychiatry: Artificial intelligence technology and psychopathology. *The Journal of Nervous and Mental Disease*, *180*, 227–237. http://journals.lww.com/jonmd

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, *1*, 60–76. http://www.jetwi.us

Gutierrez-Osuna, R. (2002). Pattern analysis for machine olfaction: A review. *IEEE Sensors Journal*, *2*, 189–202. https://doi.org/10.1109/JSEN.2002.800688

Guyon, I., & Elisseeff, A. (2006). An introduction to feature extraction. In I. Guyon, M. Nikravesh, S. Gunn, & L. A. Zadeh (Eds.), *Feature extraction: Foundations and applications* (pp. 1–25). Springer Berlin-Verlag Heidelberg. https://doi.org/10.1007/978-3-540-35488-8

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer Science & Business Media.

He, Q., & Veldkamp, B. P. (2012). Classifying unstructured textual data using the Product Score Model: An alternative text mining algorithm. In T. J. H. M. Eggen & B. P. Veldkamp (Eds.), *Psychometrics in practice at RCEC* (pp. 47–62). RCEC.

He, Q., Veldkamp, B. P., Glas, C. A. W., & de Vries, T. (2017). Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining [Advance online publication]. *Assessment*. https://doi.org/10.1177/1073191115602551

He, Q., Veldkamp, B., & De Vries, T. (2012). Screening for posttraumatic stress disorder using verbal features in self narratives: A text mining approach.

Psychiatry Research, *198*(3), 441–447. https://doi.org/10.1016/j.psychres.2012.01.032

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*, 261–266. https://doi.org/10.1126/science.aaa8685

Howes, C., Purver, M., McCabe, R., Healey, P. G. T., & Lavelle, M. (2012). Predicting adherence to treatment for schizophrenia from dialogue transcripts. *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 79–83. http://dl.acm.org/citation.cfm?id=2392814

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Proceedings of the 10th European Conference on Machine Learning*, 137–142. https://doi.org/10.1007/BFb0026683

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall.

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In D. Sleeman & P. Edwards (Eds.), *Machine Learning. Proceedings of the 9th International Workshop* (pp. 249–256). Morgan Kaufmann Publishers.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence*, *14*, 1137–1143. http://dl.acm.org

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology.* SAGE Publications, Inc.

Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, *6*(10), 1–15. https://doi.org/10.1186/1758-2946-6-10

Kurtz, A. K. (1948). A research test for the Rorschach test. *Personnel Psychology*, *1*(1), 41–51. https://doi.org/10.1111/j.1744-6570.1948.tb01292.x

Lang, K. (1995). Newsweeder: Learning to filter netnews. In A. Prieditis & S. Russel (Eds.), *Machine Learning. Proceedings of the 12th International Conference on Machine Learning* (pp. 331–339). Morgan Kaufmann Publishers.

Loper, E., & Bird, S. (2002). NLTK: The Natural Language Toolkit. *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, 62–69. https://doi.org/10.3115/1118108.1118117

Lorena, A. C., De Carvalho, A. C. P. L. F., & Gama, J. M. P. (2008). A review on the combination of binary classifiers in multiclass problems. *Artificial Intelligence Review*, *30*, 19–37. https://doi.org/10.1007/s10462-009-9114-9

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, *1*, 309–317. https://doi.org/10.1147/rd.14.0309

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Mattera, D., & Haykin, S. (1999). Support vector machines for dynamic reconstruction of a chaotic system. In B. Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in kernel methods* (pp. 211–241). MIT Press.

Mitchell, T. M. (2006). *The discipline of machine learning* (Report CMU-ML-06-108). Pittsburgh, PA, Carnegie Mellon University. http://www.cs.cmu.edu

Narendra, P. M., & Fukunaga, K. (1977). A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers*, *26*, 917–922. https://doi.org/10.1109/TC.1977.1674939

Oakes, M., Gaaizauskas, H., Rand Fowkes, Jonsson, A., Wan, V., & Beaulieu, M. (2001). A method based on the chi-square test for document classification. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 440–441. https://doi.org/10.1145/383952.384080

Oliphant, T. E. (2006). *A guide to NumPy*. Trelgol Publishing.

Paap, M. C. S., He, Q., & Veldkamp, B. P. (2015). Selecting testlet features with predictive value for the testlet effect: An empirical study. *SAGE Open*, *5*(2), 1–12. https://doi.org/10.1177/2158244015581860

Pedregosa, G., Fand Varoquaux, Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. T., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. http://www.jmlr.org/

Perkins, J. (2014). *Python 3 text processing with NLTK 3 cookbook*. Packt Publishing Ltd.

Porter, M. (2001). Snowball stemmer [Stemming algorithm]. http://snowballstem.org/

Rao, R. B., Fung, G., & Rosales, R. (2008). On the dangers of cross-validation. An experimental evaluation. *Proceedings of the 2008 SIAM International Conference on Data Mining*, 588–596. https://doi.org/10.1137/1.9781611972788.54

Raudys, S. J., & Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*, 252–264. https://doi.org/10.1109/34.75512

Rayson, P., Leech, G. N., & Hodges, M. (1997). Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, *2*, 133–152. https://doi.org/10.1075/ijcl.2.1.07ray

Rennie, J. D. M. (2001). *Improving multi-class text classification with naive bayes* (Report No. 2001-004). Cambridge, MA, Massachusetts Institute of Technology. http://publications.csail.mit.edu/

Salton, G. (1971). *The SMART retrieval system: Experiments in automatic document processing.* Prentice Hall.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. https://doi.org/10.1145/505282.505283

Shen, D., Sun, J. T., Yang, Q., & Chen, Z. (2006). Text classification improved through multigram models. *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, 672–681. https://doi.org/10.1145/1183614.1183710

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*, 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Sools, A. M., & Mooren, J. H. (2012). Towards narrative futuring in psychology: Becoming resilient by imagining the future. *Graduate Journal of Social Science*, *9*(2), 203–226. http://www.gjss.org

Sools, A. M., Tromp, T., & Mooren, J. H. (2015). Mapping letters from the future: Exploring narrative processes of imagining the future. *Journal of Health Psychology*, *20*(3), 350–364. https://doi.org/10.1177/1359105314566607

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*, 11–21. https://doi.org/10.1108/eb026526

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, *36*(2), 111–147. http://www.jstor.org

Suman, K., & Thirumagal, S. (2013). Feature subset selection with fast algorithm implementation. *International Journal of Computer Trends and Technology*, *6*, 1–5. http://www.ijcttjournal.org

Tan, C. M., Wang, Y. F., & Lee, C. D. (2002). The use of bigrams to enhance text categorization. *Information Processing and Management*, *38*(4), 529–546. https://doi.org/10.1016/S0306-4573(01)00045-0

Valderas, J. M., Kotzeva, A., Espallargues, M., Guyatt, G., Ferrans, C. E., Halyard, M. Y., Revicki, D. A., Symonds, T., Parada, A., & Alonso, J. (2008). The impact of measuring patient-reported outcomes in clinical practice: A systematic review of the literature. *Quality of Life Research*, *17*, 179–193. https://doi.org/10.1007/s11136-007-9295-0

Vapnik, V. (1995). *The nature of statistical learning*. Wiley.

Wallace, B. C., Trikalinos, T. A., Laws, M. B., Wilson, I. B., & Charniak, E. (2013). A generative joint, additive, sequential model of topics and speech acts in patient-doctor communication. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1765–1775. https://aclweb.org/

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval*, *1*, 69–90. https://doi.org/10.1023/A:1009982220290

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 42–49. https://doi.org/10.1145/312624.312647

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In D. H. Fisher (Ed.), *Machine Learning. Proceedings of the*

14th International Conference on Machine Learning (pp. 412–420). Morgan Kaufmann Publishers Inc.

Yu, C. H., Jannasch-Pennell, A., & DiGangi, S. (2011). Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. *The Qualitative Report*, *16*, 730–744. http://nsuworks.nova.edu/tqr

4

Improving web-based treatment intake for multiple mental and substance use disorders by text mining and machine learning

This chapter was published as: Wiegersma, S., Hidajat, M., Schrieken, B., Veldkamp, B.P., & Olff, M. (2022). Improving Web-based Treatment Intake for Multiple Mental and Substance Use Disorders by Text Mining and Machine Learning: Algorithm Development and Validation. *JMIR Mental Health, 9*(4):e21111. https://doi.org/10.2196/21111

Abstract

Text mining and machine learning are increasingly used in mental health care practice and research, potentially saving time and effort in the diagnosis and monitoring of patients. Previous studies showed that mental disorders can be detected based on text, but they focused on screening for a single predefined disorder instead of multiple disorders simultaneously. The aim of this study is to develop a Dutch multiclass text classification model to screen for a range of mental disorders, in order to refer new patients to the most suitable treatment. Based on patients' (N = 5,863) textual responses to a questionnaire currently used for intake and referral, a seven-class classifier was developed to distinguish among anxiety, panic, posttraumatic stress, mood, eating, substance use, and somatic symptom disorders. A linear support vector machine (SVM) was fitted using nested cross-validation grid search. The highest classification rate was found for eating disorders (82%). The scores for panic (55%), posttraumatic stress (52%), mood (50%), somatic symptom (50%), anxiety (35%), and substance use disorders (33%) were lower, likely because of overlapping symptoms. The overall classification accuracy (49%) was reasonable for a seven-class classifier. In conclusion, the developed classification model could screen text for multiple mental health disorders. The screener resulted in an additional outcome score that may serve as input for a formal diagnostic interview and referral. This may lead to a more efficient and standardized intake process.

4.1. Introduction

Mental and substance use disorders such as anxiety, mood, alcohol and drug use, eating, and depressive disorders have been listed among the leading causes of global disability over the past years (Vos et al., 2016). Annual studies show that between 2010 and 2016, these disorders accounted for approximately 18%-19% of the global burden of disease, measured in years lived with disability (Global Burden of Disease Collaborative Network, 2017). The proportion of people living with a mental disorder has remained practically unchanged in recent years (approximately 15.6%, 17.6%, and 19.0% for the global, European, and Dutch populations, respectively). However, because of population growth, absolute numbers of people diagnosed with a mental disorder have increased by 72 million globally and by 2 million in Europe between 2010 and 2016. For the Netherlands, despite an initial decrease of 15,000 from 2010 to 2014, numbers increased by 4,000 between 2014 and 2016.

This growing number of people requiring mental health care each year makes preventing and detecting mental disorders, implementing early interventions, and improving treatments and mental health care access top public health and research priorities (P. Y. Collins et al., 2011; World Health Organization, 2015). Mental health disorders are usually treated through medication or psychotherapy such as cognitive behaviour therapy (CBT), of which psychotherapy is generally seen as the first-line treatment (Linden & Schermuly-Haupt, 2014). However, mental health treatments are often underused (K. A. Collins et al., 2004) or delayed for many years (Kohn et al., 2004). Especially in low- and middle-income countries, there is a huge treatment gap in mental health care; 75% of the people experiencing anxiety, mood, impulse control, or substance use disorders remain untreated (Magruder et al., 2017). Reasons for this could be individual patient factors (e.g., embarrassment, lack of time, and geographic influences); provider factors (e.g., underdetection and lack of skill in treating mental health problems); or systemic factors such as limited access to, or limited availability of, mental health providers, resulting in waiting lists (K. A. Collins et al., 2004).

This calls for more efficient, accurate, and accessible screening and treatment methods (Bourla et al., 2018; Olff, 2015). Modern technologies are increasingly recognized as a means of improving the accessibility of care and advancing the assessment, treatment, and prevention of mental health disorders. Creative, low-cost approaches should be used to increase access to (trauma-focused) CBT and other treatments (Frewen et al., 2017). An example of such an approach is web-based self-help, which is an increasingly available alternative for a range of disorders. Web-based self-help can be therapist-guided or not, and although some studies reported equal effects for guided and unguided web-based treatment, e.g., for social anxiety disorders (Berger, Caspar, et al., 2011) and depression (Berger, Hämmerli, et al., 2011), most research endorses the importance of at least minimal, regular therapist guidance in psychological interventions (Ruwaard et al., 2012; Spek et al., 2007). Web-based therapist-guided treatment, such as computerized CBT (CCBT), is found to be approximately as effective as face-to-face treatment for several mental health disorders (e.g., depression, anxiety, and burn-out) (Cuijpers et al., 2010;

Emmelkamp, 2005; Kaltenthaler et al., 2006).

One party offering web-based, therapist-assisted CBT in the Netherlands is Interapy, a web-based mental health clinic approved by the Dutch health regulatory body. Interapy conducts screening, treatment, and outcome measurement online. Patient intake and diagnosis is performed using validated self-report instruments, followed by a diagnostic interview by telephone, after which patients are referred to a protocolled disorder-specific treatment. The treatment consists of a fixed set of evidence-based homework assignments provided through the Interapy application and uses standardized instructions that are tailored to the patient by a therapist. After submitting the homework assignments, the patient receives asynchronous personal feedback and new instructions (Ruwaard et al., 2012).

This form of web-based therapy generates large quantities of digital text data to be processed manually by the treating therapist. Textual data contains a lot of information that could be used more efficiently in the screening and treatment process through the application of text mining techniques. Text mining is generally used to automatically explore patterns and extract information from unstructured text data (Feldman & Sanger, 2007). There is a large body of literature on text mining applications in the field of psychiatry and mental health. Two recent systematic literature reviews provide a useful overview of the scope and limits, general analytic approaches, and performance of text mining in this context (Abbe et al., 2016; Wongkoblap et al., 2017). Abbe et al. (2016) concluded that text mining should be seen as a key methodological tool in psychiatric research and practice, because of its ability to deal with the ever growing amount of (textual) mental health data derived from, for example, medical files, online communities, and social media pages. However, despite the amount of data that is generated, assembling large, high-quality mental health text data sets has been found to be difficult (Wongkoblap et al., 2017). With regard to the analytic approach, in most studies predictive models are developed using supervised learning algorithms such as support vector machines (SVMs), and verified using k-fold cross-validation (Wongkoblap et al., 2017).

A way in which text mining can be put to use in mental health care practice concerns the detection of mental disorders. Previous studies showed that text mining can be used successfully in screening for posttraumatic stress disorder (PTSD) and depression (He et al., 2017; Neuman et al., 2012). He et al. (2017) developed an automatic screening model for PTSD using textual features from online self-narratives posted on a forum for trauma survivors. On the basis of a set of highly discriminative keywords and word combinations extracted from the narratives using text mining techniques, they developed a text classifier that could accurately distinguish between trauma survivors with and without PTSD. They concluded that automatic classification based on textual features is a promising addition to the current screening and diagnostic process for PTSD that can be easily implemented in web-based diagnosis and treatment platforms for PTSD and other psychiatric disorders. Neuman et al. (2012) developed an automatic screening system for depression using a "depression lexicon" based on metaphorical relations and relevant conceptual domains related to depression harvested from the internet. This lexicon was used to screen texts from open questions on a mental health website and a

set of general blog texts for signs of depression and was found to classify texts that included signs of depression very accurately.

Although both studies showed the technical potential of automatic text classification in screening for mental disorders, they applied a proxy or a self-reported diagnosis instead of a direct, formal diagnosis by a psychiatrist as the classification criterion. In addition, both studies developed a binary classifier that focused on recognizing only a single specific disorder (PTSD or depression) at a time, which is the case in most studies that apply text mining to detect mental disorders (Abbe et al., 2016; Wongkoblap et al., 2017). However, in practice, for many patients who register with mental health complaints or sign up for web-based treatment, it is not clear beforehand which disorder they should be screened for. In this case, a multiclass classifier, screening for multiple different mental disorders at once, would be more useful than a binary classifier screening for only a single prespecified disorder. Finally, it is pointed out that most natural language processing tools are currently designed for exploring English texts (Abbe et al., 2016). Although indeed, text mining and language processing tools are mainly developed for the English language, the methods and techniques underlying the text analysis process are not necessarily language dependent. The development of models for different languages depends mainly on the availability of training and testing corpora and not so much on the methods and techniques used, as will be demonstrated in this study.

This study investigates if and to what extent automatic text classification can improve the current web-based intake procedure of a Dutch web-based mental health clinic. The current intake questionnaire (see Online questionnaire) consists of open and multiple-choice questions. The multiple-choice answers are converted to scores on four scales (somatization, depression, distress, and anxiety) as well as estimates of symptom severity, required level of care, suicide and psychosis risk, and drug dependence. These scores lead to an automatically generated indicative referral advice. This advice and the answers to the open questions are used by the therapist as input for the subsequent diagnostic telephone interview to come to a formal diagnosis and referral advice. However, the current questionnaire does not cover all disorders for which treatment is offered by Interapy, and the textual answers to the open questions remain to be processed and interpreted by the therapist. An automatic text screener may provide the therapist with more specific additional information, making the intake process more efficient and standardized.

Therefore, a multiclass text classification model has been developed to screen for a range of different mental disorders with the aim of referring newly registered patients to the most fitting treatment. The focus is on a selection of treatments currently offered by Interapy for anxiety and panic disorders, PTSD, mood disorder (including depressive disorders), eating disorder, substance use disorders, and somatic symptom disorders. These will be referred to respectively as "Anxiety", "Panic", "PTSD", Mood", "Eating", "Addiction", and "Somatic" throughout the rest of this chapter. The choice for these treatments was made based on the amount of text data that was readily available from the Interapy database at the time of this research. This study adds to existing research in that 1) the patients in our

sample have an official clinical diagnosis made by a therapist; 2) our data set consists of patients with a variety of mental health disorders, enabling us to develop a multiclass text classifier; and 3) the derived texts and the resulting classifier are in Dutch and as such provide an example of non-English text mining efforts applied in mental health care research and practice.

4.2. Methods

The multi-disorder screening model was developed based on text and questionnaire data collected through the web-based intake environment of Interapy. This section describes the methods and techniques used to develop the supervised text classification model and evaluate its performance.

4.2.1. Data set

We used pretreatment scores on a self-reported questionnaire and text data derived from three open questions collected within the online intake environment. The patients are Dutch adults and adolescents who were referred to one of Interapy's web-based treatments by their general practitioner and diagnosed by a therapist. All participants have given permission for their treatment data to be used for anonymized research by Interapy to improve and evaluate their treatments through informed consent. The electronic patient database was queried in July 2017. For each treatment, all available data were retrieved, excluding incomplete or double entries. For treatments for which large quantities of data were available, a random sample of 1,100 patients was drawn to distribute the available data across the classes more evenly.

Online questionnaire

After signing up, new patients were asked to fill in the Digitale Indicatiehulp Psychische Problemen (DIPP; Digital Indication Aid for Mental Health Problems) questionnaire, an approved and validated decision support tool developed by Interapy and the HSK group, a national organization for psychological care in the Netherlands (Interapy, 2015; Van Bebber et al., 2017). The DIPP questionnaire consists of the Dutch version (Terluin, 1996) of the Four-Dimensional Symptom Questionnaire (4DSQ) (Terluin et al., 2004; Terluin et al., 2006), complemented with several multiple-choice and open questions. The 4DSQ contains 50 multiple-choice questions measuring distress, depression, anxiety, and somatization, which are dimensions of common psychopathology (Terluin et al., 2004). The complementary questions relate to current symptoms, treatment goals, anamnesis, psychosis risk, substance use, and medication. The DIPP questionnaire was originally developed, validated, and published in Dutch. A translated version of the questionnaire is provided in Appendix 4-A. The answers to the following three open questions were used to develop the text classification model:

1. Can you briefly describe your main symptom(s)?

2. What would you like to achieve with a treatment?

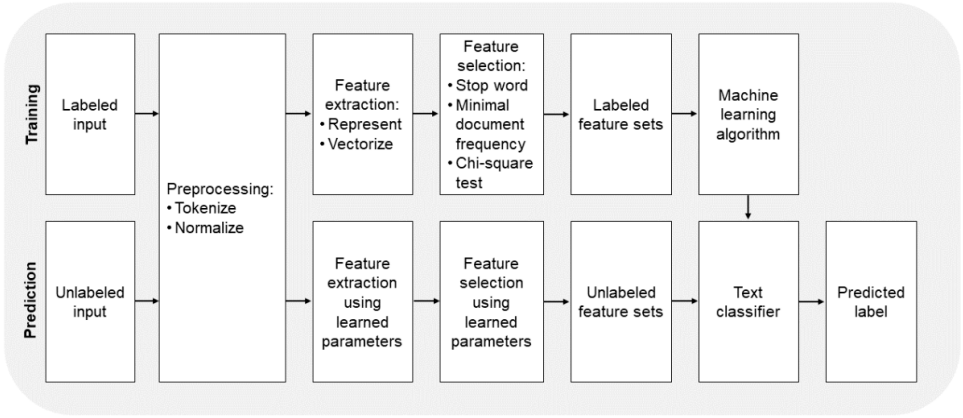


Figure 4.1: Supervised text classification model procedure. In the training phase the model is trained on labeled feature sets extracted from the input texts. In the prediction phase the trained model is used to predict labels for new, unlabeled feature sets extracted from the input texts.

3. Have there been any events (such as a divorce, loss of job, or accident) that, in your opinion, affect your current symptoms, and if so, what are they?

The information collected through the DIPP questionnaire results in scores on four scales: somatization, depression, distress, and anxiety. Each patient is then assigned a weight to indicate symptom severity and level of care (no care, general practice mental health care, basic mental health care - short, basic mental health care - moderate, basic mental health care - intensive, and specialist mental health care). The outcome is verified by a semi-structured diagnostic interview over the telephone, which results in a formal referral advice and diagnosis. Intake, diagnosis, referral, and treatment are all conducted by a CBT-certified health psychologist.

4.2.2. Automated text screening model

To screen future textual answers on the three open questions of the DIPP questionnaire for the presence of anxiety and panic disorders, PTSD, mood disorders, eating disorders, substance addiction, or somatic symptom disorders, a supervised multiclass text classifier was developed. It is called a supervised classifier because it was developed based on an existing set of text fragments provided with the correct diagnostic labels. The answers to all three questions were combined into one text document per patient. The formal referral advice based on the DIPP questionnaire scores and the diagnostic interview was used as the diagnostic label to be predicted by the model. The classifier is multiclass because the model refers each input text to one of multiple classes; the seven disorders present in the input corpus. The development of a supervised classification model follows a two-phase strategy: a model training phase and a label prediction phase. This section explains the steps taken in each phase. The complete classification procedure is shown graphically in Figure 4.1.

Training

During training, text features (words or word combinations) are extracted from each input text, converting the texts to labeled feature sets. These labeled feature sets are used as input for the machine learning algorithm, which generates a multiclass model by selecting the most informative features for each class.

Preprocessing

Standard preprocessing steps such as tokenization (splitting texts into separate tokens such as words, numerical expressions, and punctuation) and normalization (removing punctuation, converting capital letters to lower case letters, and stripping off accents) were applied to process all texts at the word level (Perkins, 2014). All words were brought back to their core, meaning-baring stem using the Snowball Stemmer, a standard stemming algorithm available for many languages including Dutch (Porter, 2001). The resulting set of words for each input text is termed the vocabulary and consists of tokens, all used words or word combinations, and types, all unique words or word combinations used (Bird et al., 2009).

Feature extraction

To convert the resulting vocabularies to feature sets suitable as input for the machine learning algorithm, the dimensionality of the feature space was reduced by feature extraction and feature selection techniques. For feature extraction, different document representation and vectorization schemes were compared. The document representations considered were unigrams, N-grams, and N-multigrams, which are single words, sequences of N words, and variable-length sequences of maximum N words, respectively (Shen et al., 2006). The vectorization schemes refer to the specified term weights, for which we used normalized term frequency (tf; Forman, 2003) or term frequency-inverse document frequency ($tf\text{-}idf$; Jurafsky & Martin, 2009).

Feature selection

Stop word removal, minimal document frequency, and the Pearson's χ^2 test were used to select the most informative features. Stop word removal was considered because stop words are generally not expected to contribute to the meaning of the text (Perkins, 2014), although other studies contradict this (Campbell & Pennebaker, 2003). In addition, words that only occur sparsely throughout the complete corpus (document frequency) may also be removed (Joachims, 1998). The most informative features (features with the highest χ^2 values) are found by ranking features based on their Pearson's χ^2 value, a common and highly efficient method that measures the independence among corpora by comparing the observed and expected feature occurrences in each class (Forman, 2003). The optimal number of features to select is determined by an exhaustive parameter grid search, which will be further explained in section 4.2.3.

Machine learning algorithm

The selected features and their corresponding labels from the training set form the labeled feature sets that were used as input for the machine learning algorithm. The SVM (Vapnik, 1995) was used because this is a high-performing and

robust classification algorithm that deals well with high-dimensional data such as text (Joachims, 1998). As SVMs were originally intended for binary classification tasks, multiclass (K-class) classification tasks were split into K binary classification tasks following the One-against-All (O-a-A, also known as One-versus-Rest) or the One-against-One (O-a-O, also known as One-versus-One) decomposition strategy.

The O-a-O strategy, which compares each pair of classes separately (Galar et al., 2011; Hastie et al., 2009), is generally considered a better approach when dealing with class imbalance, as was present in our data set. However, this strategy requires substantially more computational resources because many pairwise SVMs need to be trained. We therefore applied the widely used O-a-A strategy, which compares each single class with the remaining classes (Galar et al., 2011; Hastie et al., 2009). This strategy is the most commonly used, thanks to its computational efficiency and interpretability. To compensate for the class imbalance, a class-weighting scheme was used where classes were weighted to be inversely proportional to the class frequencies in the complete data set (as proposed by King & Zeng, 2001). This puts more emphasis on the information extracted from the smaller classes and prevents the highly present classes from overshadowing the classification model.

The SVM with O-a-A strategy was implemented in the linear support vector classifier within the LIBLINEAR library developed by Fan et al. (2008). Finally, two hyperparameters could be optimized for the SVM model: the kernel parameter γ (Duan et al., 2003), which controls model flexibility (Ben-Hur & Weston, 2010), and the regularization parameter C, which controls training and testing error (Duan et al., 2003). We used a linear kernel as is common in text classification (Joachims, 1998) and optimized the regularization parameter in the grid search (see section 4.2.3).

Prediction

During prediction, text features of new, unlabeled input texts were extracted and converted to feature sets following the same strategy used during training. Following the O-a-A approach, we fitted seven SVMs, one for each disorder, alternately comparing one of the seven classes (the positive class) to the remaining six (together forming the negative class). As described by James et al. James et al. (2013), this results in seven separate binary classification models, each with their own parameters $\beta_{0k}, \beta_{1k}, ..., \beta_{pk}$, with k denoting the k^{th} class and p the number of learned parameters. Each new, unlabeled input text x was provided with the class label for which the confidence score $\beta_{0k} + \beta_{1k}x_1 + \beta_{2k}x_2 + \cdots + \beta_{pk}x_p$ was the largest. This showed that there was a high level of confidence that the input text belonged to this class and not to one of the other six classes.

Confusion matrix

The performance of the classifier was measured by comparing the predicted labels with the known labels for each class using a confusion matrix. A confusion matrix displays the instances in the predicted classes per column and the true classes per row, directly visualizing the number of correctly labeled documents on the diagonal and the errors (mislabeled documents) in the surrounding cells (Bird et al., 2009). Table 4.1 shows the confusion matrix for a seven-class classifier with classes A-G.

Table 4.1: Confusion matrix for seven-class classifier

True label	Predicted label						
	Class_A	Class_B	Class_C	Class_D	Class_E	Class_F	Class_G
Class_A	$\mathbf{TP_A}$	$E_{A,B}$	$E_{A,C}$	$E_{A,D}$	$E_{A,E}$	$E_{A,F}$	$E_{A,G}$
Class_B	$E_{B,A}$	$\mathbf{TP_B}$	$E_{B,C}$	$E_{B,D}$	$E_{B,E}$	$E_{B,F}$	$E_{B,G}$
Class_C	$E_{C,A}$	$E_{C,B}$	$\mathbf{TP_C}$	$E_{C,D}$	$E_{C,E}$	$E_{C,F}$	$E_{C,G}$
Class_D	$E_{D,A}$	$E_{D,B}$	$E_{D,C}$	$\mathbf{TP_D}$	$E_{D,E}$	$E_{D,F}$	$E_{D,G}$
Class_E	$E_{E,A}$	$E_{E,B}$	$E_{E,C}$	$E_{E,D}$	$\mathbf{TP_E}$	$E_{E,F}$	$E_{E,G}$
Class_F	$E_{F,A}$	$E_{F,B}$	$E_{F,C}$	$E_{F,D}$	$E_{F,E}$	$\mathbf{TP_F}$	$E_{F,G}$
Class_G	$E_{G,A}$	$E_{G,B}$	$E_{G,C}$	$E_{G,D}$	$E_{G,E}$	$E_{G,F}$	$\mathbf{TP_G}$

Note. Comparison of true (rows) and predicted (columns) class labels for $\text{classes}_{A\text{-}G}$. The values on the diagonal (in boldface) show the correctly predicted class labels. The off-diagonal values show the prediction errors.

The number of true positives for class A (TP_A) were the number of times a document was labeled with A and the true label was indeed A. The false positives for class A (FP_A) were the instances that were incorrectly labeled by the classifier as A, whereas the true label was not A. This was calculated for class A using the formula: $E_{B,A}+E_{C,A}+E_{D,A}+E_{E,A}+E_{F,A}+E_{G,A}$. The false negatives for class A (FN_A) were the instances with true label A for which the classifier predicted a different label. This was calculated for class A using the formula: $E_{A,B}+E_{A,C}+E_{A,D}+E_{A,E}+E_{A,F}+E_{A,G}$. The number of true negatives for class A (TN_A) was the number of times a document for which the true label was not A was indeed not labeled as A by the classifier. This was calculated for class A by summing all counts within the confusion matrix (both errors and TPs) except for the counts in the column and row for class A. The confusion matrix was normalized by the number of documents in each class to get a more honest view of the proportions (%) of correctly predicted labels per class instead of looking at the absolute number. This is especially useful when working with an unbalanced data set where there are differences in the number of documents in each class, as was the case with our data set.

Performance metrics

The correct predictions (TPs and TNs) and errors (FPs and FNs) were then used to calculate performance metrics for each class. Bird et al. (2009) define several metrics, the simplest of which is accuracy, a measure for the proportion of correctly labeled input texts in the test set. The recall, also called sensitivity or true positive rate, indicates how many of the text documents with a true (known) positive label were identified as such by the classifier and is calculated for each class using the formula: $TP/(TP+FN)$. The precision (also known as positive predictive value) is calculated for each class by using the formula $TP/(TP+FP)$ and concerns the proportion of positively predicted text documents where the true (known) label was

indeed positive. The harmonic mean of the precision and recall, $2 \times (Precision \times Recall)/(Precision + Recall)$, is the F_1-score. The overall performance scores for the classifier were calculated by averaging the performance scores of all classes (i.e., all seven binary SVMs that were fitted following the O-a-A approach). We used weighted macro-averaged scores because this accounts for class imbalance; as this method gives equal weight to each class, it prevents the most occurring classes from dominating the model (Yang, 1999).

4.2.3. Analytical strategy

To prevent model evaluation bias, different subsets of the data were used to train, validate, and test the model. A nested k-fold cross-validation strategy was adopted, using a 5-fold cross-validated grid search in the inner loop for model selection and 5-fold cross-validation in the outer loop for model evaluation (see Figure 4.2 for a schematic representation). To make sure all classes were represented in each fold in approximately the same proportions as in the complete data set, stratified sampling (Kohavi, 1995) was used in both cross-validation loops.

For the outer loop, the data set was first split into five folds, alternately defining four folds as the development set for model selection and setting aside one fold as a test set for assessing final model performance and generalization. To optimize the different model parameters, an exhaustive parameter grid search was conducted in the inner loop. In this grid search all possible combinations of parameter values were fitted on the data set in search of the combination resulting in the highest performance score. The following model parameters and parameter values were compared:

- Choice of representation scheme: unigrams, bigrams, trigrams, or 3-multigrams
- Term weights: tf or tf-idf
- Stop words: included or excluded
- Minimal document frequency: 1, 2, 3
- Optimal number of features: ranging from 1 to 500, increasing with steps of 20
- Regularization parameter C: 1, 2, 3, 10, 100, 1,000

The search can be guided by any performance metric. We used the F_1-score because this is the preferred metric when working with imbalanced data sets. The grid search also uses a 5-fold cross-validation approach, splitting the development set into five folds, alternately using four folds for training and the remaining fold for validation. This is repeated until every fold has been used as the validation set once. The parameter combination that resulted in the highest mean weighted F_1-score over all validation sets was selected as the final model. The generalization performance of the selected model was estimated by again calculating the mean weighted F_1-score, but this time over all test sets from the outer cross-validation loop.

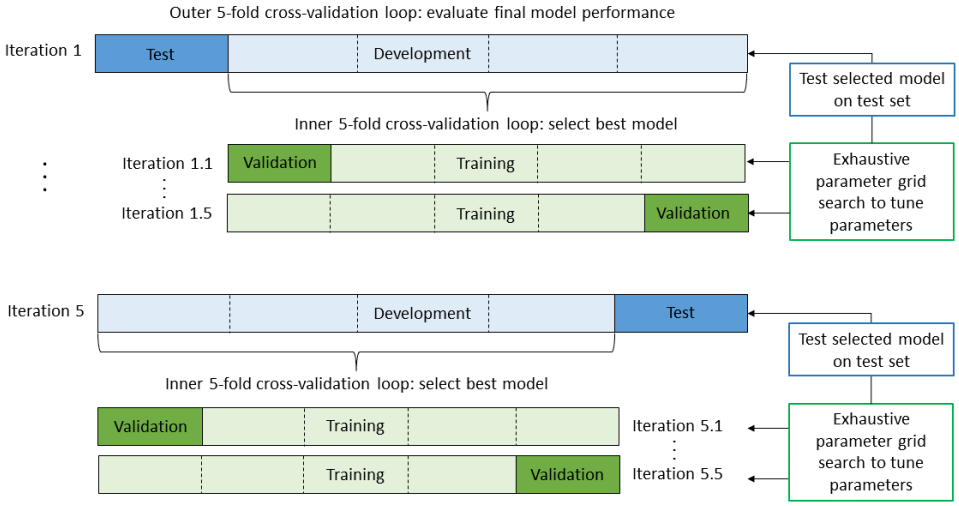


Figure 4.2: Nested 5-fold cross-validation scheme. The validation strategy consists of an inner and an outer 5-fold cross-validation loop. In the inner loop an exhaustive parameter grid search is conducted using data from the development set to select the best combination of parameter settings. The selected model is then tested on the held out test set from the outer loop to evaluate final model performance. Both loops are being iterated 5 times, alternately using each fold as test set (outer loop) or validation set (inner loop) once.

4.2.4. Text classification tool

The process of model development by means of nested stratified k-fold cross-validated grid search is fully automated in a blind text classification tool developed by the authors. This tool can be used to develop and test a text classification model on any available text data set without human insight into the data set (hence "blind"). It can be installed and used locally. After installation, no external packages are required; therefore, there is no need to send sensitive information over the internet for external text processing or analysis. An extensive description of the tool, the model development process, and the results on different test data sets can be found in Chapter 3. The tool was applied and described previously in a master's thesis (Smalbergher, 2017).

4.3. Results

4.3.1. Data set

Table 4.2 shows the demographic characteristics and DIPP questionnaire results of the patients and the lexical characteristics of their documents for each class. The class labels are "Addiction" (substance use disorders), "Panic" (anxiety disorders with panic attacks), "Anxiety" (anxiety disorders without panic attacks), "PTSD" (posttraumatic stress disorder), "Mood" (mood disorders including depressive disorders), "Eating" (eating disorders), and "Somatic" (undifferentiated somatoform and other somatic symptom disorders).

Table 4.2: Patient and lexical characteristics

Variable	Addiction (N=197)	Anxiety (N=1,100)	Panic (N=1,100)	PTSD (N=1,016)	Somatic (N=1,100)	Mood (N=1,100)	Eating (N=250)	Total (N=5,863)
Demographic characteristics								
Gender, N								
Female	18	362	394	498	500	265	180	2,217
Male	34	176	174	119	197	166	8	874
Unknown[a]	145	562	532	399	403	669	62	2,772
Age, M(SD)	37.9(15.0)	36.5(14.2)	36.3(13.8)	36.5(13.1)	41.2(11.7)	39.2(14.4)	30.8(10.0)	37.7(13.6)
DIPP results								
4DKL scales, M(SD)								
Anxiety	5.8(5.3)	8.0(5.0)	11.9(5.5)	9.3(6.3)	5.8(4.9)	6.6(5.3)	5.8(5.6)	8.1(5.8)
Depression	3.9(3.8)	3.3(3.1)	4.1(3.7)	4.8(3.8)	3.5(3.1)	6.3(3.7)	4.4(3.8)	4.4(3.7)
Distress	19.0(8.4)	19.2(7.5)	20.5(7.6)	23.6(6.9)	21.5(6.9)	23.7(6.8)	19.1(8.2)	21.5(7.5)
Somatization	10.5(6.8)	11.1(6.6)	15.3(6.9)	14.7(7.4)	13.6(6.7)	12.6(6.9)	12.4(7.1)	13.3(7.1)
Level of care, N								
No care	15	62	28	31	61	55	13	265
General practice	46	198	165	90	171	183	19	872
Basic short	11	127	92	93	110	102	8	543
Basic moderate	4	90	69	41	84	34	7	329
Basic intensive	23	340	340	244	457	283	29	1,716
Specialist	98	283	406	517	217	443	174	2,138
Lexical characteristics								
N(words), M(SD)	55.1(55.0)	71.7(69.5)	68.0(103.5)	75.1(157.0)	70.9(74.9)	65.5(75.2)	76.4(72.4)	69.9(98.2)

Note. DIPP = Dutch Digital Indication Aid for Mental Health Problems; 4DSQ = Dutch Four-Dimensional Symptom Questionnaire. For the 4DSQ trichotomized 5-point scale responses on each subscale are reported (see (Terluin et al., 2004) for the exact scoring method). Scores are considered moderately elevated (>10, >2, >8, >10) or strongly elevated (>20, >5, >12, >20) for distress, depression, anxiety, and somatization, respectively.

[a] For patients who entered the study through their general practitioner the gender is not registered; as such, gender is unknown for a large group of patients.

Table 4.3: Best parameters selected by exhaustive grid search

Parameter	Best value
Remove stop words	Yes
Minimal x documents	1
Representation scheme	Unigrams
Term weight	Term frequency (tf)
Select k best features	470
Regularization parameter C	1

Note. x = number of documents a feature should be present in; k = number of most informative features to select.

The demographic information (Table 4.2) shows that for those patients whose gender is known, more women than men had registered for all treatments except for Addiction. The mean age of the sample was 37.7 (SD 13.6) years, where patients treated for eating disorders were considerably younger (mean 30.8, SD 10.0) and patients treated for somatic disorders slightly older (mean 41.2, SD 11.7). The DIPP questionnaire results show that patients in treatment for panic attacks had the highest anxiety and somatization scores compared with those in other treatments. Patients treated for mood disorders scored higher on the depression and distress scale than those treated for other disorders. From the lexical characteristics, it can be concluded that the texts written by patients treated for addiction were considerably shorter: the mean number of words was 55.1 (SD 55.0), compared with an overall mean number of words of 69.9 (SD 98.2) for the complete sample. Patients with PTSD and eating disorders wrote relatively longer answers (mean 75.1, SD 157.0, and mean 76.4, SD 72.4, respectively).

4.3.2. Screening model

In the exhaustive grid search in the inner 5-fold cross-validation loop, all possible combinations of parameter values listed in section 4.2.3 were compared to find the model with the highest performance score. This resulted in a linear support vector classifier with a weighted F_1-score of 0.471. The selected model consisted of 470 unigrams (single words) weighted by term frequency. For this model, stop words were excluded and the selected keywords had to occur in at least one of the documents in the training set. The optimal value found for the regularization parameter C was 1. An overview of the selected model parameters is presented in Table 4.3.

Most informative features

The 50 most informative unigrams (from hereon referred to as keywords) are listed in Table 4.4. The keywords are in Dutch, followed by their English translation. The large χ^2 values and highly significant P values (when applying the O-a-A strategy, a χ^2 value > 3.84 is required to indicate significant differences ($P \leq 0.05$, df =

1)) show that there are significant differences between the observed and expected frequencies with which the keywords occur in texts written by patients with different disorders. These keywords are considered informative and were therefore included in the model. The remaining columns show the frequency with which each keyword occurs in each class (classes being the disorders for which the patients are being treated). For each keyword, the class in which it occurs most is presented in boldface. This shows that especially for the eating disorder, many highly distinctive keywords are found: 22 of the 50 keywords have the highest frequency of occurrence in Eating. Some keywords have a high occurrence in several of the classes; for example, the word "fear" occurs often in the classes Panic (N = 574), Anxiety (N = 411), and PTSD (N = 205). Of the top 50, none of the keywords occurs the most in Anxiety and only a few have the highest occurrence in Mood and Addiction.

Table 4.4: The 50 most informative features (keywords) of the multiclass classifier

Stemmed keyword (English)	χ^2	P	Addiction	Anxiety	Eating	Mood	PTSD	Panic	Somatic
eten (food)	436.99	0.000	1	18	**218**	19	20	32	22
eetbui (binge)	407.32	0.000	0	3	**121**	3	3	0	2
angst (fear)	126.63	0.000	17	411	25	98	205	**574**	82
eetstoornis (eating disorder)	100.93	0.000	0	1	**33**	1	3	1	1
paniekaanvall (panic attacks)	96.55	0.000	0	13	2	12	21	**196**	11
brak (to vomit)	93.12	0.000	0	6	**28**	0	2	0	4
boulimia (bulimia)	78.43	0.000	0	1	**26**	0	0	0	0
eetpatron (eating pattern)	75.75	0.000	0	0	**24**	2	1	0	1
gewicht (weight)	69.88	0.000	0	0	**26**	4	1	0	3
overgev (to throw up)	62.16	0.000	2	16	**39**	0	1	19	4
paniek (panic)	57.66	0.000	8	42	4	22	49	**185**	23
eet (eat)	53.40	0.000	2	6	**33**	2	4	7	2
drink (drink)	47.97	0.000	**20**	5	2	8	2	9	1
eetgedrag (eating behavior)	44.44	0.000	0	0	**14**	0	0	0	0
nachtmerries	42.26	0.000	0	7	0	6	**78**	8	1

Table 4.4: The 50 most informative features (keywords) of the multiclass classifier (Continued)

Stemmed keyword (English)	χ^2	P	Addiction	Anxiety	Eating	Mood	PTSD	Panic	Somatic
(nightmares)									
vreetbui (binge)	40.91	0.000	0	0	**12**	0	0	0	0
werk (work)	39.49	0.000	30	214	26	238	172	232	**531**
verled (past)	37.39	0.000	5	74	11	65	**188**	73	47
gezond (healthy)	36.77	0.000	4	21	**50**	30	17	37	20
overet (overeating)	35.59	0.000	0	0	**9**	0	0	0	0
zin (sense)	34.45	0.000	20	41	17	**198**	78	56	103
afvall (to lose weight)	30.60	0.000	2	1	**21**	6	3	3	3
eetproblem (eating problems)	30.25	0.000	0	0	**11**	0	2	1	2
bang (scared)	30.14	0.000	13	205	22	65	131	**206**	54
aanvall (to attack)	29.52	0.000	2	6	1	8	22	**74**	7
compenser (to compensate)	28.26	0.000	0	3	**11**	0	0	0	0
dik (fat)	28.18	0.000	0	3	**12**	2	3	3	1
angstig (anxious)	27.60	0.000	6	152	8	62	102	**168**	43
moe (tired)	27.23	0.000	12	66	10	145	88	66	**214**
paniekaanval (panic attack)	27.05	0.000	1	2	0	1	3	**55**	3
drug (drug)	26.27	0.000	**14**	5	3	5	4	6	3
verkracht (raped)	23.60	0.001	1	2	3	0	**44**	6	4
ongeluk (accident)	23.02	0.001	7	26	1	20	**87**	30	24
overgewicht (overweight)	22.93	0.001	0	1	**8**	2	1	1	1
blow (to blow)	22.55	0.001	**10**	1	1	0	6	0	0
hyperventilatie (hyperventilation)	22.52	0.001	2	3	0	2	4	**51**	7
vermoeid	22.50	0.001	7	33	4	60	35	38	**134**

Table 4.4: The 50 most informative features (keywords) of the multiclass classifier (Continued)

Stemmed keyword (English)	χ^2	P	Addiction	Anxiety	Eating	Mood	PTSD	Panic	Somatic
(tired)									
alcohol (alcohol)	22.47	0.001	**15**	9	5	6	4	6	5
misbruik (abuse)	21.14	0.002	5	9	0	6	**53**	6	4
obsessive (obsession)	21.05	0.002	0	2	**6**	0	0	0	0
flashback (flashback)	20.74	0.002	2	1	0	4	**27**	1	0
eating (eating)	20.18	0.003	0	0	**5**	0	0	0	0
lustelos (heavy-headed)	19.98	0.003	9	40	6	**105**	27	23	74
control (control)	19.62	0.003	9	53	49	45	36	**102**	38
geget (ate)	19.33	0.004	0	0	**7**	1	0	0	0
ondergewicht (underweight)	18.94	0.004	0	0	**6**	1	0	0	0
voeding (nutrition)	18.91	0.004	0	2	**9**	1	1	0	0
somber (gloomy)	18.58	0.005	3	32	6	**112**	32	40	38
normal (normal)	18.43	0.005	8	58	55	44	83	**105**	63
verslav (addictive)	17.87	0.007	**10**	4	4	3	2	1	4

Note. The 50 most informative features with the highest χ^2 values and significant ($P \leq 0.05$) P values. The keyword column contains the stemmed keyword in Dutch, followed by the English translation in parentheses. The remaining columns show occurrence frequencies for each feature in each class (disorder). For each feature, the frequency for the class in which it occurs the most is printed in boldface.

Performance metrics

Table 4.5 reports the performance scores of the final model for each class. The model performs especially well in screening for eating disorders. The high precision (0.75) for this class means that 75% of the patients whom the model classified as having an eating disorder, were indeed referred to a treatment for eating disorders by the therapist. The high recall (0.82) shows that 82% of the patients who were referred to a treatment for eating disorders by the therapist were also identified as such by the model. The model screens the least effective for addiction and anxiety. Only 25% of the patients who were classified by the model as having an addiction and 44% of the patients with anxiety were also identified as such by the therapist.

Table 4.5: Performance metrics final model

Disorder	Precision	Recall	F_1-score	Overall accuracy[a]	N(patients) in test set
Addiction	0.25	0.33	0.28		40
Anxiety	0.44	0.35	0.39		220
Eating	0.75	0.82	0.78		50
Mood	0.44	0.50	0.47		220
PTSD	0.57	0.52	0.54		203
Panic	0.57	0.55	0.56		220
Somatic	0.46	0.50	0.48		220
Weighted average/Total(N)	0.50	0.49	0.49	0.49	1173

Note. Per class and average performance scores for the final model.
[a] Accuracy is the overall accuracy of the classifier averaged over all classes.

Of the patients referred to treatments for addiction and anxiety by the therapist, respectively, only 33% and 35% were also found by the model. The overall accuracy of the classifier is 0.49, meaning that 49% of the predictions made by the model were correct. For a 7-class classifier this exceeds random guessing, which would be 1/7 = 0.14 (14%).

Confusion matrix

The confusion matrix in Table 4.6 contains the absolute counts and normalized values (counts corrected by the number of documents present in each class, in %) for the true and predicted labels. The normalized values are the most useful because these indicate the proportion of correctly predicted labels for each class, independent of the class sizes. The normalized values on the diagonal show that the classifier screens the best for Eating (82% correct), followed by Panic (55%), PTSD (52%), Somatic (50%), Mood (50%), Anxiety (35%), and Addiction (33%). In total, this screener referred 578 of the 1,173 patients (49%) from the test set to the correct treatment.

The normalized confusion matrix is plotted in Figure 4.3 to give a more direct visual presentation of which classes are being misclassified. The darker the blue tones, the higher the proportions in that cell. The perfect classifier would have a dark blue diagonal line, surrounded by white cells. The plot confirms that Eating is rarely misclassified. Most confusion occurs for Addiction, which is often mislabeled as a mood or somatic disorder. In addition, mood and somatic disorders are often confused with each other, as are panic and anxiety disorders.

Table 4.6: Confusion matrix for seven-class classifier

True disorder	Predicted disorder						
	Addiction	Anxiety	Eating	Mood	PTSD	Panic	Somatic
Addiction	**13 (33%)**	3 (7%)	1 (3%)	8 (20%)	3 (7%)	3 (7%)	9 (23%)
Anxiety	11 (5%)	**77 (35%)**	6 (3%)	33 (15%)	27 (12%)	41 (19%)	25 (11%)
Eating	1 (2%)	1 (2%)	**41 (82%)**	4 (8%)	1 (2%)	0 (0%)	2 (4%)
Mood	11 (5%)	26 (12%)	0 (0%)	**110 (50%)**	14 (6%)	10 (5%)	49 (22%)
PTSD	2 (1%)	18 (9%)	0 (0%)	36 (18%)	**105 (52%)**	19 (9%)	23 (11%)
Panic	4 (2%)	27 (12%)	3 (1%)	24 (11%)	18 (8%)	**121 (55%)**	23 (10%)
Somatic	10 (5%)	23 (10%)	4 (2%)	37 (17%)	16 (7%)	19 (9%)	**111 (50%)**

Note. Absolute and normalized values (%) for the true versus predicted class labels. The diagonal cells show the correctly predicted labels (in boldface). The off diagonal cells show the prediction errors for each class.

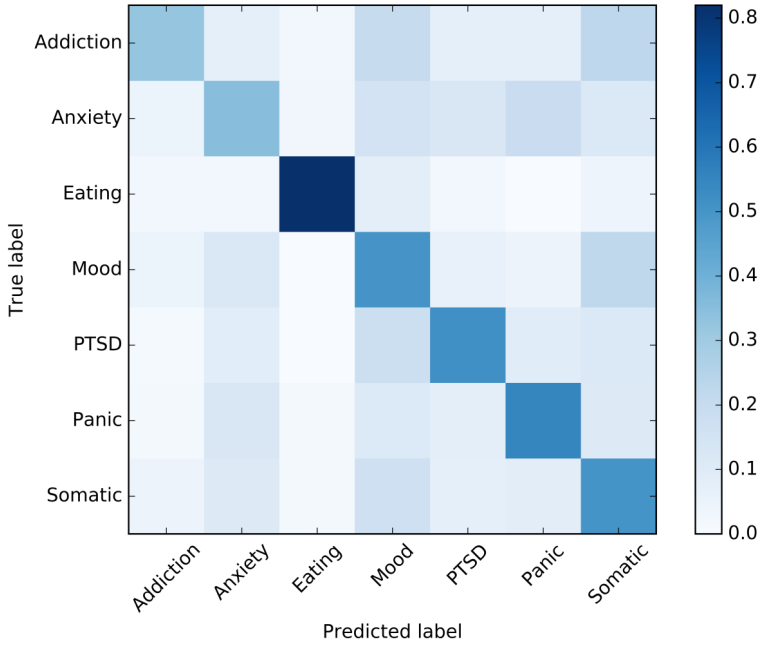


Figure 4.3: Normalized confusion plot. Visual presentation of the true versus predicted class labels. The darker the tone, the higher the proportion in the corresponding cell.

Final model evaluation

The 5-fold cross-validation grid search was conducted five times in the inner loop, iteratively using four of the five folds from the outer loop as the development set once. This resulted in five weighted F_1-scores: one for each final model selected in the inner cross-validation loop that was tested on the test set in the outer cross-validation loop. The weighted F_1-scores for the five outer test folds were 0.49, 0.49, 0.47, 0.46, and 0.47. The scores are relatively close to each other, meaning that the classifier generates stable results. The mean weighted F_1-score over the five iterations was 0.48 (SD 0.01). This is the estimated generalization performance, the performance that can be expected when the final model is applied to new data sets in the future.

4.4. Discussion

4.4.1. Principal results

This study aims to improve the intake procedure of a web-based mental health therapy provider by using multiclass text classification to automatically screen textual answers on open questions from an intake questionnaire for a range of different mental health disorders. The resulting classification model turned out to be especially effective in screening for Eating, correctly identifying 82% of the patients with an eating disorder. This is comparable to binary classifiers in previous studies; for example, for PTSD (80% correct, performance score for the SVM model based on unigrams; He et al., 2017) or depression (84% correct; Neuman et al., 2012). The correct classification rates for the other disorders were substantially lower; Panic (55%), PTSD (52%), Mood (50%), Somatic (50%), Anxiety (35%), and Addiction (33%), resulting in an overall accuracy of 49%. This is a reasonable score for a 7-class classification model, although not high enough to make strong and accurate referrals for all treatments.

The difference in performance is also reflected in the selected keywords, of which many are highly discriminative for Eating. For example, simple words such as "food", "binge", "weight", or "bulimia" are clearly related to eating disorders while sparsely being used in texts written by patients with other disorders. For the remaining disorders, the keywords found are more generally related to fears and feelings and occur more in all classes except for Eating and thus are less discriminative. For example "fear" and "scared" are selected as keywords for Panic, but they also have high occurrences in Anxiety and PTSD. "Sense" is a keyword for Mood, but it is also highly used in texts written by patients with somatic disorders, whereas the somatic keyword "tired" is also used often in texts written by patients with a mood disorder. As a result, the model could not accurately differentiate between mood and somatic disorders as well as between panic and anxiety disorders. None of the 50 most informative keywords was related mostly to Anxiety, for which one of the lowest classification performances was reported.

Reasons for the overlap in keywords for different disorders may be symptom overlap (in case symptoms are part of the defining symptom set of multiple disorders) and nonspecificity of defining symptoms (in case symptoms also occur

regularly in persons without the disorder), both issues resulting from definitional choices made in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5; Van Loo & Romeijn, 2015). For example, PTSD has overlapping symptom criteria with depression, generalized anxiety disorder, and panic disorder (Kessler et al., 1995). When (future) patients are asked to describe their most important symptoms (one of the three open intake questions, the answers to which were used to develop our model, see section Online questionnaire), because symptoms for several disorders overlap, it is not surprising that descriptions and thus keywords for these disorders will also overlap.

The low screening performance for Addiction could be because only a very small number of patients with addiction were present in the data set (N = 197), and as such the machine learning algorithm was provided with inadequate training data for this class. However, for Eating, few more patients were included (N = 250), and for this class the classifier performed very well. Another reason could be that patients in Addiction were found to write shorter texts; on average, the mean number of words used by patients in the Addiction class is 55.1 (SD 55.0), versus an average of 69.9 (SD 98.2) over all classes and even 76.4 (SD 72.4) for the Eating class (Table 4.2). This shows that patients with an eating disorder provide a more extensive description of their symptoms, treatment goals, and anamnesis than patients with addiction. Because of this, less information is available for Addiction than for Eating, which makes it hard for the machine learning algorithm to learn key features for this class.

The results further show that the classifier has difficulty differentiating mood from somatic disorders and panic from anxiety disorders. For mood and somatic disorders this can be explained by the fact that most patients with somatic disorders are commonly found to have an underlying mood disorder (Smith, 2006). The difficulty in distinguishing between panic and anxiety disorders could be because panic disorder is actually classified as a type of anxiety disorder in the DSM-5 (American Psychiatric Association, 2013). Despite the underlying similarity, we expected that panic disorders could be easily distinguished from anxiety disorders because of their distinctive characteristics. Although the classifier found quite a few significant keywords for Panic (e.g., "fear", "panic attack(s)", "panic"), these words also occurred often in texts written by patients with Anxiety and PTSD and thus were not discriminative enough. In contrast, none of the top 50 keywords had the highest frequency of occurrence in the Anxiety class, meaning no highly discriminative keywords were found for Anxiety. As Panic and Anxiety are closely related, merging the two classes into one would probably improve the performance of the screener. However, this would reduce the practical applicability of the screener because the goal is to refer patients to the most suitable treatment offered by the health care provider, which offers separate treatments for Panic and Anxiety.

4.4.2. Theoretical and practical contributions

First, this study extends the findings of previous research on text classification applications in mental health care in that it investigates the use of a multiclass classifier instead of a binary classifier, which is predominantly used (Abbe et al.,

2016; Wongkoblap et al., 2017). This way it is possible to screen for multiple disorders at once, without the need to make prior assumptions regarding the type of disorder a new patient signs up with. Second, this study shows an application of text mining and natural language processing applications originally developed for English text to non-English, in this case Dutch, mental health data. Although most of the scientific publications in this area focus on English data and tools (Abbe et al., 2016; Wongkoblap et al., 2017), most underlying processes and techniques are not language dependent and as such can be easily applied to non-English texts. Finally, our data set contained high-quality class labels, consisting of official clinical diagnoses made by a therapist, enabling us to compare the labels predicted by the classifier to an official "gold standard" instead of a proxy. The quality of the labels is highly important for the performance, validity, and clinical applicability of the developed model, and acquiring large, high-quality mental health text data sets is found to be challenging (Wongkoblap et al., 2017).

For the web-based mental health provider, the developed text screener provides an additional outcome score that can be used as input for the automatically generated indicative diagnosis and for the formal diagnostic interview by the therapist. Although the overall performance of the classifier still needs to be improved, the classifier was able to distinguish eating disorders very well. As an eating disorder is currently not reported as a separate scale in the DIPP questionnaire (which reports on anxiety, depression, distress, and somatization), the text screener provides additional information that was not available from the multiple-choice questions.

This study further shows how text mining, specifically text classification, can add value to current (web-based) mental health care practice because it can be used for more efficient screening, intake, or treatment referral. As described previously, mental health problems often remain undiagnosed and untreated. This can partly be attributed to the fact that most people are only seen by primary care providers, who do not always recognize mental health conditions because of comorbidity between physical and psychological diseases. Magruder et al. (2017) therefore propose that primary care clinicians should receive more training on the recognition of these conditions. However, even after being diagnosed, patients often remain untreated because of the scarcity of health care resources. To scale up the mental health workforce, the World Health Organization World Health Organization (2008) has proposed to shift care giving to mental health workers with lower qualifications or even lay helpers under the supervision of highly qualified health workers (Magruder et al., 2017). An alternative way of reducing the workload for mental health workers is to increase the use of modern technologies in screening, providing treatment, and monitoring treatment outcomes. Instead of (or in addition to) extra training for primary care providers, an automatic screening tool could also aid in the recognition of mental health problems, and instead of shifting care to lower-qualified or lay helpers, mental health providers could be supported by modern technology. The automatic screener described in this study should be seen as an example of this.

4.4.3. Limitations

An important limitation of our classifier is that it is not capable of dealing with comorbidity. Comorbidity is an important issue; 45% of the patients with psychiatric disorders are reported to meet the criteria for two or more disorders within the same year (Van Loo & Romeijn, 2015). As stated earlier, it is not unusual for patients with somatic disorders to have an underlying mood disorder (Smith, 2006), whereas mood disorders are commonly found to co-occur with anxiety disorders (Van Loo & Romeijn, 2015). Substance use disorders are also often found to co-occur with other mental health disorders; for drug use disorders in particular, high associations with anxiety (especially panic disorder) and affective (mood) disorders have been reported (Conway et al., 2006; Regier et al., 1990; Torrens et al., 2011). The main limitation of this study is that although the multiclass classifier can screen for multiple disorders at once, it does not take into account the possibility that a patient can have a combination of multiple disorders simultaneously (comorbidity). This may explain why the screener did not prove to be very capable when it came to distinguishing between some disorders, which indicates the need for a multi-label classifier that can screen for combinations of disorders instead of only a single disorder.

Another limitation may be the fact that we used a blind tool to develop the automatic screening model. Some might state that to develop a model, at least some insight into the input data is required to actively monitor the development process. However, the tool was tested and applied in a previous study by the authors and in a master's thesis (Smalbergher, 2017) in which the process and outcomes were confirmed. This tool enabled us to work on sensitive information without any insight into the textual content, on a local computer, and without the need to send the information over the internet for processing and analysis, thereby reducing not only the risk of privacy issues, but also the risk of possible confirmation bias because of prior knowledge. However, by using a tool, one is limited by the choice of models and parameters made beforehand, during the development of the tool. Adding to, or changing, the tool's settings based on new insights is quite laborious, because this requires developing, updating, and installing a new version. Therefore, we chose to use a common and proven classifier and analytic approach (Wongkoblap et al., 2017).

Yet another limitation could be the definition of the classes and class imbalance. The classes used in this study are defined by the specific diagnoses for which treatment is offered by the mental health clinic Interapy, instead of symptomatology. The performance of the classifier might be improved by grouping together comorbid disorders or disorders with overlapping symptoms (e.g., combine somatic and mood disorders or panic and anxiety disorders). However, because this would decrease the practical usability of the screener, we chose to keep these classes separate. Model performance may also be influenced by class (im)balance, that is, the extent to which the texts are evenly distributed across the classes. The classes Addiction and Eating were strongly underrepresented in our data set, and despite the use of class weights and stratified samples, performance for the Addiction class was especially poor. In contrast, the highest performance was reported for the Eating

class; therefore, it seems that as long as the text content is discriminative enough, even small samples may provide enough information to make strong predictions.

4.4.4. Future research

Future research should focus first of all on improving the overall performance of the classifier. The current screener does not show a high enough performance for all classes, which might be solved by trying alternative classification algorithms or machine learning strategies, such as a multilabel strategy to deal with comorbidity. In addition to adopting a multilabel approach, exploring a multistage learning system also seems a useful next step. Multistage models (e.g., cascade classifiers) use a staged decision process in which the output of a model (the first stage) is used as the input for a successive model (the second stage), and so on. Multistage models are widely used in medical practice, and physicians use this approach for the stepwise exclusion of possible diagnoses (Bennasar et al., 2014). Several studies show that multistage classifiers outperform the single-stage classifiers generally used in supervised multiclass classification tasks; for example, in the prediction of liver fibrosis degree (Hashem et al., 2012) and in distinguishing among levels of dementia (Bennasar et al., 2014). For our screener it could be useful to first classify the disorders into more general groups of (possibly) overlapping disorders, grouping Anxiety, Panic, and PTSD in one class and Mood and Somatic symptom disorders in another while keeping Eating and Substance abuse disorders separate, followed by a more specialized classification model to distinguish among the specific disorders within the groups. This prevents the best predictable class (in our case, Eating) from dominating the machine learning process. In addition, because one of the problems was finding (enough) discriminative keywords for some of the disorders, adding additional open questions to the web-based intake procedure to collect more text data may be helpful. Adjusting the questions by focusing less on symptoms (which are found to overlap for some disorders) and focusing instead on aspects possibly more defining for each disorder may also lead to more discriminative keywords and consequently better models.

Second, further uses of text mining and machine learning in mental health care practice should be explored. Text mining can be (and is) used for many more activities during and after treatment; for example, in analyzing patient-physician or patient-carer communication (Wallace et al., 2013) or in evaluating treatments by capturing patients' opinions from online comments (Greaves et al., 2013). In addition, text mining can also be used to assess factors and processes underlying recovery of, for example, patients with an eating disorder (Keski-Rahkonen & Tozzi, 2005). A new application for text mining in e-mental health practice could be to use it as a tool to support therapists by offering suggestions for patient-specific feedback. The current CCBT process as used in this study consists of sequential homework assignments covering common CBT interventions. On the basis of the content of these assignments, therapists offer standardized feedback and instructions, including motivational techniques, adapted to the needs and situation of the patient (Ruwaard et al., 2012). It would be interesting to examine whether we could use text mining to automatically highlight sections in the assignments that

require attention or that may indicate a positive or negative change in behavior.

4.4.5. Conclusions

This study showed that automatic text classification can improve the current web-based intake and referral procedure of a Dutch mental health clinic, by providing an additional outcome score to be used as input for the indicative referral advice and the formal diagnostic interview. Automatically generating an additional indicator based on the textual input may lead to a more efficient and standardized intake process, saving time and resources because the text no longer needs to be processed and interpreted by the therapist. As such, automatic text screening could be a step in the right direction for solving patient, systemic, and provider factors underlying the underdetection of mental health disorders and underuse of available mental health treatments (K. A. Collins et al., 2004). The overall complaint-discriminating quality of the screener still has to be improved, but the good detection performance with regard to eating disorders in this study (and with regard to PTSD and depression in other studies) shows that text-based screening is a promising technique for psychiatry. This paper contains multiple recommendations for research paths that could improve this complaint-discriminating quality of text screeners (e.g., using stratified analysis techniques when symptoms overlap complaints). Altogether, the technique is getting closer to implementation in general practice, where it definitely could be of great value. Especially in areas around the world with a limited number of mental health care workers, automatic text classification could be helpful. It could save time that is now spent on screening and assessment of patients, time that could be used for counseling and treatment.

4.5. References

Abbe, A., Grouin, C., Zweigenbaum, P., & Falissard, B. (2016). Text mining applications in psychiatry: A systematic literature review. *International Journal of Methods in Psychiatric Research*, *25*, 86–100. https://doi.org/10.1002/mpr.1481

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*.

Ben-Hur, A., & Weston, J. (2010). A user's guide to support vector machines. In O. Carugo & F. Eisenhaber (Eds.), *Data mining techniques for the life sciences* (pp. 223–239). Humana Press.

Bennasar, M., Setchi, R., Hicks, Y., & Bayer, A. (2014). Cascade classification for diagnosing dementia. *Proceedings of the 2014 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 2535–2540.

Berger, T., Caspar, F., Richardson, R., Kneubühler, B., Sutter, D., & Andersson, G. (2011). Internet-based treatment of social phobia: A randomized controlled trial comparing unguided with two types of guided self-help. *Behaviour Research and Therapy*, *49*(3), 158–169. https://doi.org/10.1016/j.brat.2010.12.007

Berger, T., Hämmerli, K., Gubser, N., Andersson, G., & Caspar, F. (2011). Internet-based treatment of depression: A randomized controlled trial comparing guided with unguided self-help. *Cognitive behaviour therapy*, *40*(4), 251–266. https://doi.org/10.1080/16506073.2011.616531

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'reilly Media, Inc.

Bourla, A., Mouchabac, S., El Hage, W., & Ferreri, F. (2018). e-PTSD: An overview on how new technologies can improve prediction and assessment of Posttraumatic Stress Disorder (PTSD). *European Journal of Psychotraumatology*, *9*(sup1), 1424448. https://doi.org/10.1080/20008198.2018.1424448

Campbell, R., & Pennebaker, J. (2003). The secret life of pronouns: Flexibility in writing style and physical health. *Psychological Science*, *14*(1), 60–65. https://doi.org/10.1111/1467-9280.01419

Collins, K. A., Westra, H. A., Dozois, D. J. A., & Burns, D. D. (2004). Gaps in accessing treatment for anxiety and depression: Challenges for the delivery of care. *Clinical Psychology Review*, *24*(5), 583–616. https://doi.org/10.1016/j.cpr.2004.06.001

Collins, P. Y., Patel, V., Joestl, S. S., March, D., Insel, T. R., & Daar, A. S. (2011). Grand challenges in global mental health. *Nature*, *475*(7354), 27–30. https://doi.org/10.1038/475027a

Conway, K. P., Compton, W., Stinson, F. S., & Grant, B. F. (2006). Lifetime comorbidity of DSM-IV mood and anxiety disorders and specific drug use disorders: Results from the national epidemiologic survey on alcohol and related conditions. *Journal of Clinical Psychiatry*, *67*(2), 247–257.

Cuijpers, P., Donker, T., Van Straten, A., Li, J., & Andersson, G. (2010). Is guided self-help as effective as face-to-face psychotherapy for depression and anxiety disorders? A systematic review and meta-analysis of comparative outcome

studies. *Psychological Medicine*, *40*(12), 1943–1957. https://doi.org/10.1017/S0033291710000772

Duan, K., Keerthi, S., & Poo, A. (2003). Evaluation of simple performance measures for tuning SVM hyper parameters. *Neurocomputing*, *51*, 41–59. https://doi.org/10.1016/S0925-2312(02)00601-X

Emmelkamp, P. M. G. (2005). Technological innovations in clinical assessment and psychotherapy. *Psychotherapy and Psychosomatics*, *74*(6), 336–343. https://doi.org/10.1159/000087780

Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, *9*, 1871–1874. http://www.jmlr.org/

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, *3*, 1289–1305. http://www.jmlr.org/

Frewen, P., Schmahl, C., & Olff, M. (2017). Interdisciplinary approaches to understand traumatic stress as a public health problem. *European Journal of Psychotraumatology*, *8:sup5*, 1441582. https://doi.org/10.1080/20008198.2018.1441582

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*, *44*, 1761–1776. https://doi.org/10.1016/j.patcog.2011.01.017

Global Burden of Disease Collaborative Network. (2017). *Global burden of disease study 2016 (gbd 2016) results [Data file]*. http://ghdx.healthdata.org/gbd-results-tool

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Use of sentiment analysis for capturing patient experience from free-text comments posted online. *Journal of Medical Internet Research*, *15*(11). https://doi.org/10.2196/jmir.2721

Hashem, A. M., Rasmy, M. E. M., Wahba, K. M., & Shaker, O. G. (2012). Single stage and multistage classification models for the prediction of liver fibrosis degree in patients with chronic hepatitis c infection. *Computer Methods and Programs in Biomedicine*, *105*(3), 194–209. https://doi.org/10.1016/j.cmpb.2011.10.005

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer Science & Business Media.

He, Q., Veldkamp, B. P., Glas, C. A. W., & de Vries, T. (2017). Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining [Advance online publication]. *Assessment*. https://doi.org/10.1177/1073191115602551

Interapy. (2015). *Digitale Indicatiehulp Psychische Problemen: Inhoudelijke Dossier.* [Digital Indication Aid for Mental Health Problems: Content File.] Unpublished report.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Proceedings of the 10th European Conference on Machine Learning*, 137–142. https://doi.org/10.1007/BFb0026683

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall.

Kaltenthaler, E., Brazier, J., De Nigris, E., Tumur, I., Ferriter, M., Beverley, C., Parry, G., Rooney, G., & Sutcliffe, P. (2006). Computerised cognitive behaviour therapy for depression and anxiety update: A systematic review and economic evaluation. *Health Technology Assessment*, *10*(33), 1–70.

Keski-Rahkonen, A., & Tozzi, F. (2005). The process of recovery in eating disorder sufferers' own words: An internet-based study. *International Journal of Eating Disorders*, *37*(S1), S80–S86. https://doi.org/10.1002/eat.20123

Kessler, R. C., Sonnega, A., Bromet, E., Hughes, M., & Nelson, C. B. (1995). Posttraumatic stress disorder in the national comorbidity survey. *Archives of General Psychiatry*, *52*(12), 1048–1060. https://doi.org/10.1001/archpsyc.1995.03950240066012

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence*, *14*, 1137–1143. http://dl.acm.org

Kohn, R., Saxena, S., Levav, I., & Saraceno, B. (2004). The treatment gap in mental health care. *Bulletin of the World health Organization*, *82*, 858–866.

Linden, M., & Schermuly-Haupt, M.-L. (2014). Definition, assessment and rate of psychotherapy side effects. *World Psychiatry*, *13*(3), 306.

Magruder, K. M., McLaughlin, K. A., & Elmore Borbon, D. L. (2017). Trauma is a public health issue. *European Journal of Psychotraumatology*, *8:1*, 1375338. https://doi.org/10.1080/20008198.2017.1375338

Neuman, Y., Cohen, Y., Assaf, D., & Kedma, G. (2012). Proactive screening for depression through metaphorical and automatic text analysis. *Artificial Intelligence in Medicine*, *56*(1), 19–25. https://doi.org/10.1016/j.artmed.2012.06.001

Olff, M. (2015). Mobile mental health: A challenging research agenda. *European Journal of Psychotraumatology*, *6*(1), 27882. https://doi.org/10.3402/ejpt.v6.27882

Perkins, J. (2014). *Python 3 text processing with NLTK 3 cookbook*. Packt Publishing Ltd.

Porter, M. (2001). Snowball stemmer [Stemming algorithm]. http://snowballstem.org/

Regier, D. A., Farmer, M. E., Rae, D. S., Locke, B. Z., Keith, S. J., Judd, L. L., & Goodwin, F. K. (1990). Comorbidity of mental disorders with alcohol and other drug abuse: Results from the epidemiologic catchment area (ECA)

study. *Journal of American Medical Association*, *264*(19), 2511–2518. https://doi.org/10.1001/jama.1990.03450190043026

Ruwaard, J., Lange, A., Schrieken, B., Dolan, C. V., & Emmelkamp, P. (2012). The effectiveness of online cognitive behavioral treatment in routine clinical practice. *PLoS one*, *7*(7). https://doi.org/10.1371/journal.pone.0040089

Shen, D., Sun, J. T., Yang, Q., & Chen, Z. (2006). Text classification improved through multigram models. *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, 672–681. https://doi.org/10.1145/1183614.1183710

Smalbergher, I. (2017). *Learning analytics towards identifying the processes of higher order thinking in online discussions of MOOCs* (Master's thesis). http://essay.utwente.nl/74246/

Smith, M. (2006). Psychiatric disorders. In C. Sproat, G. Burke, & M. McGurk (Eds.), *Essential human disease for dentists* (pp. 217–233). Churchill Livingstone. https://doi.org/10.1016/B978-0-443-10098-7.50017-8

Spek, V., Cuijpers, P., Nyklíček, I., Riper, H., Keyzer, J., & Pop, V. (2007). Internet-based cognitive behaviour therapy for symptoms of depression and anxiety: A meta-analysis. *Psychological Medicine*, *37*(3), 319–328. https://doi.org/10.1017/S0033291706008944

Terluin, B. (1996). De vierdimensionale klachtenlijst (4DKL). Een vragenlijst voor het meten van distress, depressie, angst en somatisatie [The four-dimensional symptom questionnaire (4DSQ). A questionnaire for measuring distress, depression, anxiety and somatization]. *Huisarts Wet*, *39*(12), 538–547.

Terluin, B., Van Rhenen, W., Schaufeli, W. B., & De Haan, M. (2004). The four-dimensional symptom questionnaire (4DSQ): Measuring distress and other mental health problems in a working population. *Work & Stress*, *18*(3), 187–207. https://doi.org/10.1080/0267837042000297535

Terluin, B., van Marwijk, H., Adèr, H., de Vet, H., Penninx, B., Hermens, M., van Boeijen, C., van Balkom, A., van der Klink, J., & Stalman, W. (2006). The four-dimensional symptom questionnaire (4dsq) a questionnaire to measure distress, depression, anxiety, and somatization. *BMC Psychiatry*, *6*, 34.

Torrens, M., Gilchrist, G., & Domingo-Salvany, A. (2011). Psychiatric comorbidity in illicit drug users: Substance-induced versus independent disorders. *Drug and Alcohol Dependence*, *113*(2), 147–156. https://doi.org/10.1016/j.drugalcdep.2010.07.013

Van Bebber, J., Wigman, J. T. W., Wunderink, L., Tendeiro, J. N., Wichers, M., Broeksteeg, J., Schrieken, B., Sytema, S., Terluin, B., & Meijer, R. R. (2017). Identifying levels of general distress in first line mental health services: Can GP- and eHealth clients' scores be meaningfully compared? *BMC Psychiatry*, *17*(1), 382. https://doi.org/10.1186/s12888-017-1552-3

Van Loo, H. M., & Romeijn, J. W. (2015). Psychiatric comorbidity: Fact or artifact? *Theoretical Medicine and Bioethics*, *36*, 41–60. https://doi.org/10.1007/s11017-015-9321-0

Vapnik, V. (1995). *The nature of statistical learning*. Wiley.

Vos, T., Allen, C., Arora, M., Barber, R. M., Bhutta, Z. A., Brown, A., Carter, A., Casey, D. C., Charlson, F. J., Chen, A. Z., et al. (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: A systematic analysis for the global burden of disease study 2015. *The Lancet*, *388*(10053), 1545–1602. https://doi.org/10.1016/S0140-6736(16)31678-6

Wallace, B. C., Trikalinos, T. A., Laws, M. B., Wilson, I. B., & Charniak, E. (2013). A generative joint, additive, sequential model of topics and speech acts in patient-doctor communication. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1765–1775. https://aclweb.org/

Wongkoblap, A., Vadillo, M. A., & Curcin, V. (2017). Researching mental health disorders in the era of social media: Systematic review. *Journal of medical Internet research*, *19*(6), e228.

World Health Organization. (2008). Task shifting: Rational redistribution of tasks among health workforce teams: Global recommendations and guidelines.

World Health Organization. (2015). The European mental health action plan 2013-2020. *Copenhagen: World Health Organization*.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval*, *1*, 69–90. https://doi.org/10.1023/A:1009982220290

Appendix 4-A
Translated DIPP questions

This appendix provides the translated Digitale Indicatiehulp Psychische Problemen (DIPP; Digital Indication Aid for Mental Health Problems) questionnaire used for the web-based intake of new patients. The DIPP questionnaire was originally developed and validated in Dutch (Interapy, 2015). The DIPP questionnaire starts with the Dutch version of the Four-Dimensional Symptom Questionnaire (4DSQ; Terluin, 1996; Terluin et al., 2004; Terluin et al., 2006), followed by additional questions regarding current symptoms, treatment goals, anamnesis, psychosis risk, substance use, and medication.

4DSQ questions

First the 50 questions of the 4DSQ are completed. The specific questions can be found in previous publications (Terluin, 1996; Terluin et al., 2004; Terluin et al., 2006).

Additional DIPP questions

Open questions used for text screening

Can you briefly describe your main symptom(s)?

What would you like to achieve with a treatment?

Have there been any events (such as a divorce, loss of job, or accident) that, in your opinion, affect your current symptoms, and if so, what are they?

Multiple-choice questions

Have there been any previous times in your life when you had similar symptoms?
- Yes
- No

Have you recovered from these complaints in the meantime?
- Yes
- No
- Does not apply

How long have you been on sick leave because of your current symptoms?
- I am not on sick leave; I do not have a paid job
- 1 week
- 2 weeks
- 3 weeks
- 4 weeks
- More than 4 weeks

How many hours are you actually working per week now?
I am currently working ... hours a week

How many hours of work per week are laid down in your employment contract?
... hours a week

Do you avoid daily, necessary activities or situations because of your symptoms (e.g., shopping, travelling by public transport, visiting friends)?
- Yes
- No

For how long have you been suffering from your current symptoms?
- One month or less
- 2 months
- 3 months
- 4 months
- 5 months
- 6 months
- More than 6 months

Have you previously had treatment for the same symptoms (e.g., by a psychiatrist, psychologist, general practice based mental health nurse specialist (POH-GGZ), or general practitioner)?
- Yes
- No

Was there a period after the treatment in which you were free of symptoms?
- Yes
- No

Psychosis risk
Have you experienced any of the following in the past five years?

I sometimes get messages from voices in or near my head.
- Not at all
- A little
- Quite
- Certainly
- Very much

I sometimes hear voices that other people cannot hear.
- Not at all
- A little
- Quite
- Certainly

- Very much

Substance use
Do you use alcohol?
- No
- Not any more, quit
- Yes

If you quit, since when? …

Do you use soft drugs?
- No
- Not any more, quit
- Yes

If you quit, since when? …

Do you use hard drugs?
- No
- Not any more, quit
- Yes

If you quit, since when? …

Medication
Participants are asked about medication use;
- Drug name
- Dose
- First use

Date of birth
What is your date of birth? (dd-mm-yyyy)

Sex
What is your gender?
- Female
- Male

Telephone number
Required to make a callback appointment.

Schedule an appointment
When all questions have been answered, there are two options to schedule an appointment for the indication interview; this depends on the practice/organization that applies the DIPP:

A. the patient can choose from a number of day times offered in the application;
B. the patient provides his telephone number and is called by the assistant or secretary for an appointment.

5

Recognizing hotspots in Brief Eclectic Psychotherapy for PTSD by text and audio mining

This chapter was published as: Wiegersma, S., Nijdam, M.J., Van Hessen, A.J., Truong, K.P., Veldkamp, B.P., & Olff, M. (2020). Recognizing hotspots in Brief Eclectic Psychotherapy for PTSD by text and audio mining, European Journal of Psychotraumatology. *European Journal of Psychotraumatology, 11*(1). https://doi.org/10.1080/20008198.2020.1726672

Abstract

Identifying and addressing hotspots is a key element of imaginal exposure in Brief Eclectic Psychotherapy for PTSD (BEPP). Research shows that treatment effectiveness is associated with focusing on these hotspots and that hotspot frequency and characteristics may serve as indicators for treatment success. This study aims to develop a model to automatically recognize hotspots based on text and speech features, which might be an efficient way to track patient progress and predict treatment efficacy. A multimodal supervised classification model was developed based on analog tape recordings and transcripts of imaginal exposure sessions of ten successful and ten non-successful treatment completers. Data mining and machine learning techniques were used to extract and select text (e.g., words and word combinations) and speech (e.g., speech rate, pauses between words) features that distinguish between "hotspot" (N = 37) and "non-hotspot" (N = 45) phases during exposure sessions. The developed model resulted in a high training performance (mean F_1-score = 0.76) but a low testing performance (mean F_1-score = 0.52). This shows that the selected text and speech features could clearly distinguish between hotspots and non-hotspots in the current data set, but will probably not recognize hotspots from new input data very well. In order to improve the recognition of new hotspots, the described methodology should be applied to a larger, higher quality (digitally recorded) data set. As such this study should be seen mainly as a proof of concept, demonstrating the possible application and contribution of automatic text and audio analysis to therapy process research in posttraumatic stress disorder (PTSD) and mental health research in general.

5.1. Introduction

Posttraumatic stress disorder (PTSD) is a mental health disorder that can develop after experiencing or witnessing a traumatic event (American Psychiatric Association, 2013). The lifetime prevalence rate of PTSD in the general population is 7.4% (De Vries & Olff, 2009; Kessler et al., 2017). Several effective treatments for PTSD exist (Bisson et al., 2019), examples of which are trauma-focused cognitive behavioral therapy (CBT; Ehlers & Clark, 2000) and eye movement desensitization and reprocessing (EMDR; Shapiro, 2001). Of all effective psychotherapies, one of the ingredients they have in common is exposure to trauma (Olff et al., in press; Schnyder et al., 2015). Despite its efficacy, there is still a considerable proportion of patients that does not (sufficiently) respond to this form of trauma-focused therapy. For example, in their meta-analysis of psychotherapy for PTSD, Bradley et al. (2005) report mean improvement rates of 37.6% and 47.4%, among CBT intent-to-treat patients and treatment completers respectively.

Grey et al. (2002) argue that the effectiveness of PTSD treatment can significantly improve by focusing on hotspots. This is in line with the results of Nijdam et al. (2013), who showed that hotspots were more frequently addressed in successful than in non-successful treatments. Hotspots, the moments of traumatic experiences with the highest emotional impact, have been an important topic of research in the past decades. For example, Ehlers et al. (2004) and Ehlers et al. (2005) found that imaginal exposure during trauma-focused CBT should focus on addressing and changing the meaning of hotspots as this could lead to greater PTSD symptom reduction. The importance of hotspots in psychotherapy was also highlighted in earlier studies that argued that hotspots need to be addressed to ensure habituation (Richards & Lovell, 1999) or to identify deeper meanings (Ehlers & Clark, 2000).

A form of trauma-focused CBT that focuses on the identification and addressing of hotspots is Brief Eclectic Psychotherapy for PTSD (BEPP; Gersons et al., 2000). Through imaginal exposure, the patient is led slowly through the traumatic situation until the worst moment (the hotspot) is reached (Grey & Holmes, 2008). Hotspots are addressed by encouraging the patient to describe and remember the exact details of the most frightening or emotional moment, for example by asking about sounds, smells, weather, or surroundings. By helping the patient to remember the details, cues to new aspects and details of the event can come to mind, enabling the patient to relive the situation as vividly as possible (Gersons et al., 2011). When the hotspot is sufficiently covered, the trauma narrative can be continued until (over the course of several exposure sessions) all the hotspots have been addressed. The imaginal exposure phase is completed when all hotspots are addressed and the emotions associated with the traumatic event are leveled down sufficiently (Nijdam et al., 2013).

With regard to the content of hotspots, previous studies focused on the presence of emotions (Grey et al., 2001; Grey et al., 2002) and cognitions (Grey & Holmes, 2008; Holmes et al., 2005), which showed that especially anxiety, helplessness, horror, anger, sadness, shame, and guilt frequently occurred in hotspots. In addition, Grey et al. (2002) found that hotspots are characterized by subtle

textual changes, which may guide the therapist in the identification of emotional hotspots. An example of a study assessing textual differences within trauma narratives is that of Jelinek et al. (2010), who studied the organization and content of the "worst moments" of traumatic memories by analyzing the degree of disorganization, emotions, and speaking style. They found that these moments showed different characteristics with regard to organization than the rest of the narrative.

To obtain a deeper knowledge and understanding of trauma treatment and specifically hotspots, more in-depth, large scale analysis of treatment and hotspot content is required. Until now, treatment content has mainly been studied by manually coding the occurrence of a predefined set of characteristics within therapy session recordings or transcripts retrospectively. Due to the time-consuming nature of such analyses, most of these studies focus on one specific construct, such as text cohesion (Foa et al., 1995), complexity (Amir et al., 1998), or dissociation (Zoellner et al., 2002). It is suggested that future studies should focus on assessing the relationship between multiple constructs underlying traumatic narratives instead of studying every construct separately (Amir et al., 1998).

An effective way to study multiple constructs and variables at once is to analyze treatment sessions using automatic text and audio analyses. Text analysis is frequently used in PTSD research, as word use and linguistic features proved to be indicative of people's mental, social, and sometimes physical state, and their defensive operations (Nelson & Horowitz, 2010; Pennebaker et al., 2003). Word counts have been used to study trauma narrative content in relation to PTSD symptom severity (e.g., Jelinek et al., 2010; Pennebaker, 1993) and insight in the linguistic elements present within trauma narratives could lead to improved PTSD treatment (Alvarez-Conrad et al., 2001). For example, specific linguistic features such as cognitive processing words (Alvarez-Conrad et al., 2001; D'Andrea et al., 2012; Pennebaker et al., 2001), emotion words (Alvarez-Conrad et al., 2001; Pennebaker et al., 2001), words related to insight (Pennebaker et al., 2001), reflection (D'Andrea et al., 2012), causation (Boals & Klein, 2005), and affection and death (Alvarez-Conrad et al., 2001) have been used to predict improvements in post-treatment PTSD symptoms, perceived physical health, and personal functioning. Because mood and emotions are found to influence speaking behavior and speech sound characteristics, audio signal analysis is regularly applied in psychiatric studies as well, for example to predict recovery time in depression (Kuny & Stassen, 1993) or to recognize psychosis development in high-risk youths (Bedi et al., 2015).

Text and speech features can be used to identify and study specific concepts on a large scale, in a transparent and uniform fashion, over a long period of time. For the automatic recognition or prediction of pre-defined concepts, supervised classification is generally used. Supervised classification is a data mining application in which objects (e.g., texts or audio signals) are assigned to a set of predefined class labels using a classification model based on labeled training samples (Bird et al., 2009). Supervised classification based on text features has been used for example to screen forum posts for PTSD (He et al., 2012) or to predict treatment adherence for schizophrenia patients (Howes et al., 2012), whereas speech features have been used to classify distress in PTSD patients (Van Den Broek et al., 2009).

Though most studies use either text or audio analysis, Schuller et al. (2005) and Forbes-Riley and Litman (2004) found that models based on multimodal feature sets outperformed models based on either acoustic or linguistic features alone (e.g., in emotion classification), as multimodal sets provide a broader and more complete picture of one's (emotional) state (Bhaskar et al., 2015).

This study aims to develop a multimodal supervised classification model to automatically recognize hotspots based on text and speech features extracted from tape recordings and transcripts of imaginal exposure sessions of successful and non-successful treatment completers. Automatic hotspot recognition can provide clinicians with insight in the occurrence and characteristics of hotspots during their treatments, which may assist them in offering a more effective intervention. We hypothesized that a combination of text and speech features extracted from patient speech could be used to develop a supervised classification model to automatically distinguish between hotspot and non-hotspot phases during imaginal exposure sessions. Based on the formal hotspot characteristics and previous research on hotspots and CBT sessions, we identified nine constructs (affect, emotions, cognitions, dissociation, avoidance, cohesion, organization, fragmentation, and complexity, further described in section 5.2) that we expected to differ between hotspots and non-hotspots. Each construct was operationalized through a number of text and speech characteristics that were captured using a large range of features extracted from CBT session transcripts and recordings.

5.2. Methods

5.2.1. Sample and data set

We used data of patients undergoing Brief Eclectic Psychotherapy for PTSD (BEPP; Gersons et al., 2000). To develop the hotspot classification model, an existing expert-annotated data set consisting of imaginal exposure session recordings was used in which hotspots and their characteristics were coded. This data set consisted of analog cassette tape recordings of 45 PTSD patients and was collected for a previous study by Nijdam et al. (2013), who investigated differences in hotspots between successful and unsucessful BEPP trauma-focused psychotherapies. They analyzed session recordings in which imaginal exposure was present for 20 of the 45 patients (the ten most and the ten least successful treatment completers). The sample consisted of twelve female and eight male adults with a mean age of 39.60 (SD 10.98) and different ethnic backgrounds (mainly Dutch, N = 15, but also Indonesian, Surinamese, Aruban, and Bosnian). The types of trauma the patients experienced included assault (N = 13), disaster (N = 2), sexual assault (N = 1), accident (N = 1), war-related (N = 1), and other (N = 2).

Nijdam et al. (2013) coded the frequency of hotspots, their characteristics (interrater reliability K = 0.86), emotions (interrater reliability K = 0.81), and cognitions (interrater reliability K = 0.85) for 102 recordings based on the Hotspot Identification Manual, an adaptation of the Hotspots Manual of Holmes and Grey (2002), developed by Nijdam and colleagues to enable retrospective coding based on audio recordings. Of the 102 coded sessions, recordings of insufficient quality for tran-

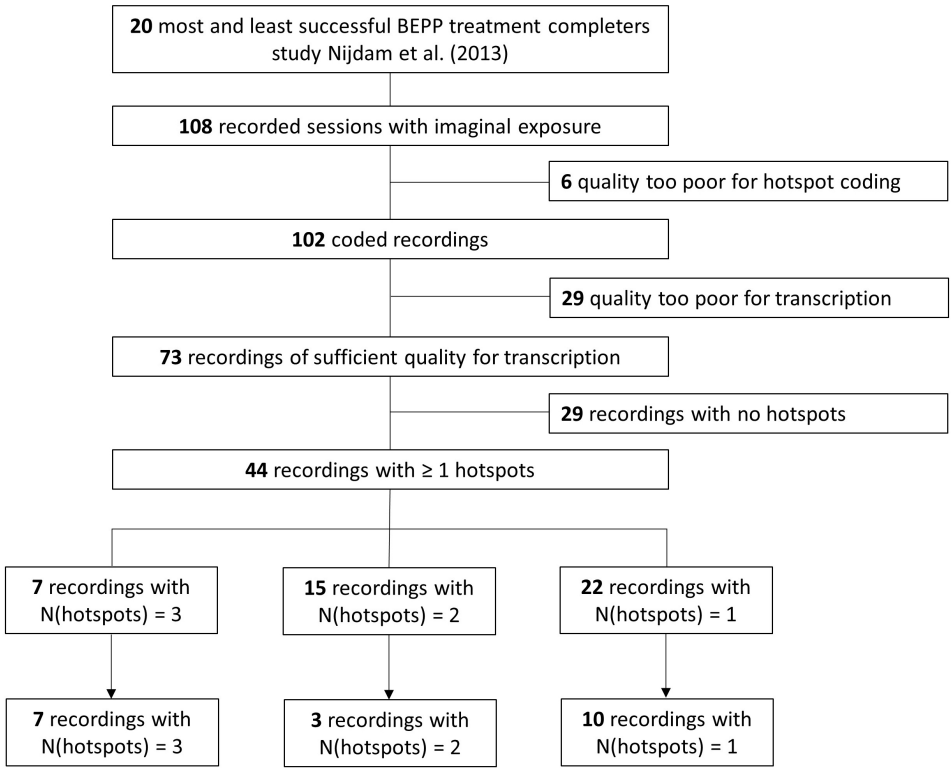


Figure 5.1: Data selection chart for available session recordings

scription (mainly due to heavy background noise, N = 29) or that did not contain any hotspots (N = 29) were excluded in the present study. From the remaining 44 recordings one session was selected per patient. This was the session in which the most hotspots occurred. In case there were multiple sessions with the same number of hotspots, the session occurring earliest in treatment was used. In total the twenty selected sessions contained 37 hotspots; seven recordings with three hotspots, three recordings with two hotspots, and ten recordings with one hotspot (see Figure 5.1).

Data preparation

The data consisted of tape recordings (mono channel) of complete imaginal exposure sessions, which were converted to WAV format (16-bit, 16 kHz, mono) using the digital audio editor Audacity® version 2.0.5 (Audacity Team, 2013). The recordings were over ten years old at the time of digitization, which negatively influenced the sound quality. Each recording contained a complete imaginal exposure session consisting of four elements (Gersons et al., 2011):

- Discussion: discussion of the previous session, the course of the PTSD symp-

toms, and the structure and content of the current session.

- Relaxation exercises: repeatedly tensing and relaxing muscle groups to enable the patient to focus on the traumatic event and go back to the situation.
- Exposure: for the first exposure session, the patient is brought back to the day of the traumatic event and is asked to give a detailed account of the situation prior to the event and the event itself. In subsequent sessions, exposure starts where it left off in the previous session.
- Discussion: discussion of the exposure experienced so far and explanation of the content and structure of the following session.

Since we were only interested in the imaginal exposure phase, the initial discussion of the previous session, relaxation, and concluding discussion were removed, leaving only the exposure phase, usually about 15-20 minutes per recording, for analysis.

Because the audio was of poor quality and transcriptions needed to be as detailed as possible, automatic speech recognition (ASR) was not applicable. The recordings were therefore transcribed and annotated by the first author, who was blind to therapeutic outcome. The transcriptions are verbatim, meaning that every recorded word, including unfinished words (stammering), non-fluencies (e.g., uh, hmm), and forms of backchanneling (e.g., uhhu, ok), was transcribed. Background noise was removed only if necessary for transcription, using the noise reduction function implemented in Audacity® version 2.0.5. However, for some sessions small parts of the speech could still not be transcribed due to the amount of noise, heavy emotions, or weakness of the speech signal. These parts are coded as "inaudible", including start and end time. The exact start and end time of each hotspot were coded by the first two authors.

The transcriptions were then converted to the (C)XML file format for annotating transcriptions to enable parsing (easily separating patient from therapist speech and hotspot from non-hotspot phases for the text analysis) and to link the transcribed text to the digital audio recordings. Linking of text and audio data was done using forced alignment within the WebMaus Pipeline version 2.25 (Kisler et al., 2017), including the Chunker function by Poerner and Schiel (2016). The resulting TextGrid files were then complemented with interval tiers; connected sequences of labeled intervals annotating hotspots, speaker turns, and silences, using Praat version 6.0.4.3 (Boersma & Weenink, 2019). This way the transcriptions and recordings were converted to input formats suitable for the multimodal classification pipeline.

Identifying hotspots

The Hotspots Manual of Holmes and Grey (2002), and succeeding research on hotspots by Holmes et al. (2005) and Grey and Holmes (2008), was used to identify hotspots: 1) the moment is defined by the patient as the "worst moment"; 2) the moment was identified as a hotspot in a previous session; 3) an audible change in affect; 4) the patient changes from present to past tense; 5) the patient changes

from first to third person; 6) the patient is "whizzing through"; 7) the patient cannot remember details of the moment; 8) the patient is dissociating; or 9) the moment is mentioned by the patient to correspond to an intrusion.

5.2.2. Operational constructs for automatic recognition

We distinguished nine constructs underlying hotspots that could be used in their automatic recognition. Five of these are based on the formal hotspot characteristics; affect, emotions, cognitions, dissociation, and avoidance. The remaining four (cohesion, organization, fragmentation, and complexity) were selected based on previous research on CBT sessions. Although until now, except for organization, these additional constructs were mainly studied with regard to complete trauma narratives and not to specific parts such as hotspots, we expected them to be useful for automatic hotspot recognition as they do play a part in the emotional processing of traumatic events (Amir et al., 1998).

Each construct is operationalized through variables that can be measured based on combinations of either text, speech, or text and speech features. Since the aim is to recognize hotspots automatically, we only used those variables that could be measured based on automatically extracted (i.e., without the need for manual coding) text and speech features. The features used to capture each construct are described in sections Text feature extraction and Speech feature extraction. More feature details, including examples and equations, can be found in Appendix 5-A and Appendix 5-B. The operationalization of each construct and the related features are shown schematically in Figure 5.2 and elaborated upon in Appendix 5-C.

Affect According to Grey et al. (2002) a visible change in affect (e.g., bursting into tears, turning red, shaking, or sweating) is the most obvious way to identify a hotspot. When working with audio files, audible cues can be used instead of visible cues, as in Nijdam et al. (2013), who showed that change in affect remains a strong identifier even without the visible aspect. Juslin and Scherer (2005) define affect as *"a general, umbrella term that subsumes a variety of phenomena such as emotion, stress, mood, interpersonal stance, and affective personality traits"* [p. 69].

Emotions Emotion is one of the affective phenomena listed by Juslin and Scherer (2005), and as such the constructs affect and emotions are closely related. Holmes et al. (2005) distinguished 11 emotion categories based on emotion words that occurred during hotspots: fear, helplessness, anger, sadness, surprise, disgust, dissociation, happiness, shame, guilt, and horror. Of these, especially anxiety, helplessness, and horror are deemed important, as these were specified explicitly under PTSD criterion A2 of the DSM-4-TR (American Psychiatric Association, 2013), although this criterion was removed from the most recent version, the DSM-5 (American Psychiatric Association, 2000). We also expected higher occurrences of the emotions anger, sadness, shame, disgust, and guilt, as these were found to be often related to hotspots (Grey et al., 2001; Grey et al., 2002; Holmes et al., 2005).

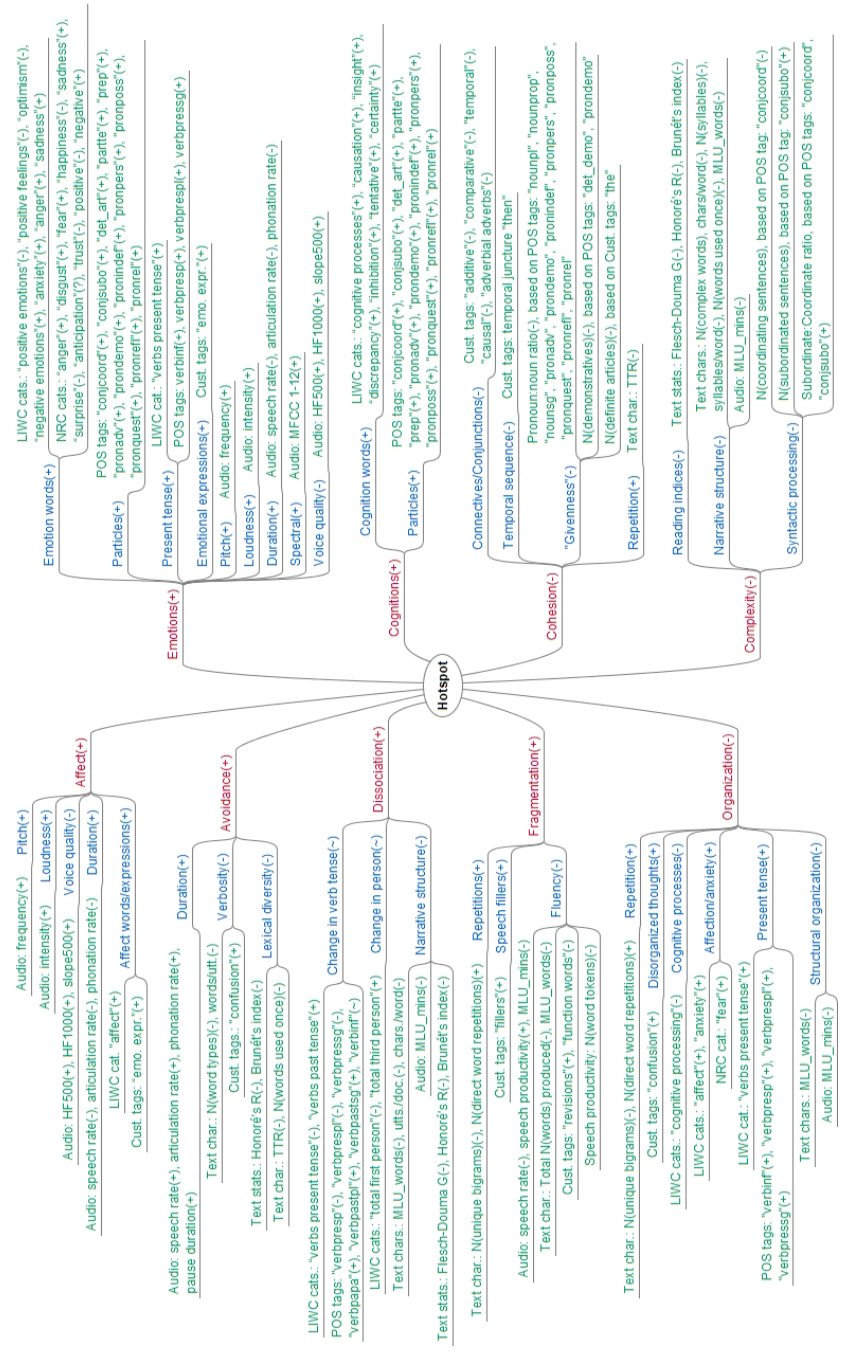


Figure 5.2: Operationalization scheme for constructs underlying hotspots (red), related variables (blue), and extracted features (green). For each node is indicated whether it is expected to increase(+), decrease(-), change in both directions(~), or either direction(?).

Cognitions In addition to emotion categories, Holmes et al. (2005) distinguished seven cognitive themes that can characterize hotspots: uncertain threat, general threat of injury and death, control and reasoning, consequences, abandonment, esteem, and cognitive avoidance. Cognitive themes of psychological threat (sense of self) were found to appear more in hotspots than those of physical threat (physical integrity) (Grey & Holmes, 2008; Holmes et al., 2005).

Dissociation Hotspots are also identified by changes in speaking style. During imaginal exposure, patients are asked to describe the past event as if it were happening now, in the first person present tense. Patients may dissociate during hotspots by changing from present to past tense or from first to third person (Grey et al., 2002). This altered or unreal perception of the traumatic event may indicate that peritraumatic dissociation occurred during or directly after the traumatic experience.

Avoidance Other hotspot characteristics related to speaking style described by Grey et al. (2002) are "whizzing through" (rushing through the main event giving minimal details, while extensively describing the buildup and aftermath) and the patient declaring he or she is unable to remember details of the moment. These characteristics reflect (non-conscious) avoidance.

Cohesion Narrative cohesion focuses on the occurrence of explicit cues within the text that enable the reader (or listener) to make connections within or between sentences or clauses (Crossley et al., 2016). Previous studies found cohesion to be related to the level of intrusive symptoms in children (O'Kearny et al., 2007) and trauma-related avoidance (O'Kearney et al., 2011), which both are hotspot characteristics.

Organization Trauma survivors with PTSD are found to produce more disorganized trauma narratives than trauma survivors without PTSD (Halligan et al., 2003; Jones et al., 2007). The (dis)organization of the "worst moments" (hotspots) in traumatic memories was previously studied based on text features by Jelinek et al. (2010).

Fragmentation Foa et al. (1995) suggest that trauma memories are more fragmented (i.e., lacking flow) for trauma survivors with PTSD, because information could not be adequately processed and encoded under stressful conditions. They found a significant correlation between fragmentation and PTSD symptoms over treatment.

Complexity Amir et al. (1998) found that narrative complexity correlated negatively with PTSD severity three months after the trauma. They found that patients who wrote more simplistic narratives showed more severe PTSD than patients who

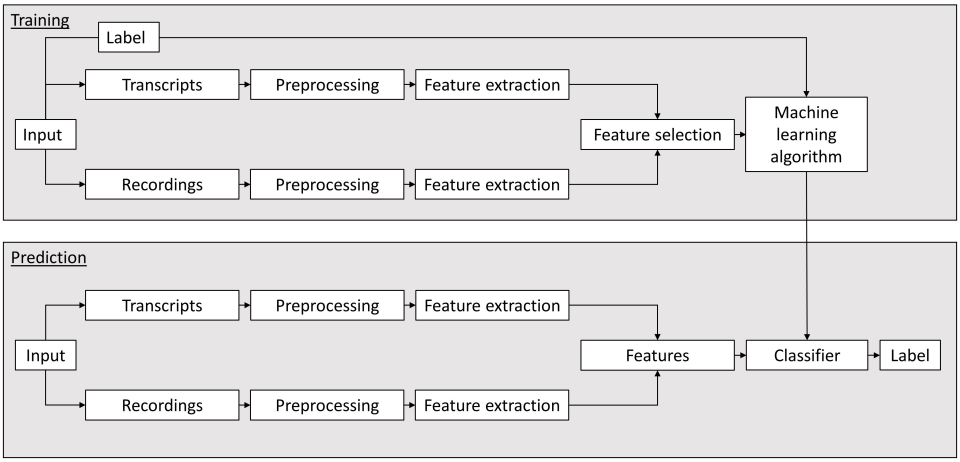


Figure 5.3: Multimodal supervised classification pipeline

wrote more complex narratives. However, later studies concluded that found effects could also be due to differences in writing skill and cognitive ability (see Gray & Lombardo, 2001). Complexity may relate to the hotspot characteristic "whizzing through", due to which hotspot moments are described in a more simplistic fashion and in less detail. Also, hotspot moments may be narrated in a more fragmented way due to changes in affect.

5.2.3. Classification pipeline

The development of a new classification model involves two phases; a training phase and a prediction phase. In the training phase, information is extracted from each object following a range of preprocessing and feature extraction steps, resulting in labeled feature sets. A machine learning algorithm uses those labeled feature sets to learn and select the most discriminative text and speech features for the "hotspot" versus the "non-hotspot" phases. In the prediction phase, the classifier uses those features to identify hotspots from new imaginal exposure session recordings and transcripts (for more on the development of classification models, see Chapter 3). This sequence of steps, in which the output of each step is the input for the next, is called a pipeline (see Figure 5.3).

The preparation, preprocessing and feature extraction steps were done separately for text and speech features because they require different techniques. Feature selection and machine learning were applied to the combined, multimodal feature sets. Text preprocessing and feature extraction was done in Python 3.7.2 (Python Software Foundation, 2019) using the Natural Language Toolkit (NLTK 3.4; NLTK Project, 2019) and Python's Textstat package (version 0.5.4; Bansal & Aggarwal, 2018), and in LIWC using the Dutch LIWC dictionary and the NRC emotion lexicon. Audio preprocessing and feature extraction was done using Audacity® version 2.0.5, WebMaus version 2.25, and Praat version 6.0.4.3. Conversion of the

text transcripts from plain text files to parsable and linkable file formats was done using custom XML and CXML converters developed by one of the authors (available upon request). For feature selection and machine learning, the Scikit-learn library (Pedregosa et al., 2011) version 0.20.2 was used.

Preprocessing

The text and audio analysis focused on patient speech only. The textual input for the classification pipeline consisted of plain text files containing the transcribed, anonymized patient speech cut into "hotspot" and "non-hotspot" segments (parts in the exposure phase preceding or following a hotspot). In total the transcripts were split into 37 hotspot segments and 45 non-hotspot segments. To analyze the text on word level, separate words were extracted from the transcripts using the word tokenizer for Dutch implemented in NLTK (see Perkins, 2014, for more on tokenization). All words were normalized by removing punctuation, accents, and capital letters. For the N-gram extraction, each word except for stop words was stemmed (reduced to its base form, see Jurafsky & Martin, 2009, for more on stemming) using a standard Dutch Snowball stemmer included in NLTK (Porter, 2001). For the tagger-based feature extraction and the overall text characteristics the unstemmed input text was used.

For the audio analysis, the prepared TextGrid files (see Data preparation) were directly processed in Praat, selecting the audio signals for patient-speech only and distinguishing between hotspot and non-hotspot phases within the annotated interval tiers. In line with Jurafsky and Martin (2009), we used utterances instead of sentences because we work with a corpus of transcribed speech that does not contain punctuation such as original text corpora. Utterances, which can be words, phrases or clauses, were identified based on Tanaka et al. (2014), in which utterances are separated based on a pause in speech longer than one second.

Text feature extraction

Text features capture what is being said, focusing on the textual content. Text content can be examined on word or phrase level by extracting unigrams, N-grams, or N-multigrams (single words, phrases, or variable-length word combinations). With small samples, frequencies of individual words or phrases may be too low to recognize specific patterns. In that case it is useful to analyze words belonging to particular grammatical or lexical categories by assigning labels (tags) to each word using parts-of-speech (POS), lexicon-based, or custom taggers.

In general, grammatical POS tags such as personal pronouns and verb tense are thought to give information about one's (temporal) focus and psychological distance towards a situation or event, which may provide cues on thought processes, priorities, and intentions (Tausczik & Pennebaker, 2010). Tags regarding verb tense are also considered useful in assessing memory (dis)organization and time perspective (Jelinek et al., 2010). Previous studies in which POS tags were used showed that tags such as first-person singular pronouns correlated positively with psychological distress (Rude et al., 2004; Wolf et al., 2007). In addition, trauma survivors that were sensitive to developing posttraumatic stress symptoms were found to use

more first-person plural than first-person singular pronouns (Chung & Pennebaker, 2007; Stone & Pennebaker, 2002).

A widely used lexicon-based tagger is LIWC, which assigns words to categories related to linguistic elements, emotions, and cognitive processes, and counts their relative frequencies. Since hotspots are the most emotionally distressing moments of trauma (Nijdam et al., 2013), special attention was paid to the emotions present in the transcripts. Although LIWC extracts several emotion categories (anxiety, anger, and sadness), more extensive insight in the emotions was gained using a General Purpose Emotion Lexicon (GPEL), which is considered to significantly improve emotion classification (Aman & Szpakowicz, 2007).

Finally, text characteristics and statistics were extracted to analyze textual differences on the general level. Previous studies showed that these characteristics can be used to detect emotions (Lee & Narayanan, 2005) or as indicators for physical symptoms and discomfort (Alvarez-Conrad et al., 2001).

The text features were extracted over the complete hotspot or non-hotspot phase, extracting all text features for each separate hotspot and non-hotspot segment. To prevent bias towards longer text documents, the extracted N-grams were weighted by normalized term frequency (tf; occurrence counts normalized by document length, see more in Forman, 2003) or term frequency-inverse document frequency (tf-idf; see more in Jurafsky & Martin, 2009), which are the most commonly used feature weights. The occurrence frequencies returned by the taggers were normalized by document length. A detailed description of all used text features, their relation to the operational constructs, and the extraction process can be found in Appendix 5-A.

Speech feature extraction

In addition to what is being said, which is captured by the text features, it is of interest how things are said, since one's manner of speaking can convey signs of emotions or stress (Lefter et al., 2011; K. R. Scherer, 2003). Some emotions, especially emotions that are high in arousal, such as anger and fear, can be better identified from spoken than from written data (e.g., Truong & Raaijmakers, 2008).

The study of speech sounds is called phonetics. Phonetic studies can focus on how sounds are produced (articulatory phonetics), how sounds are perceived (auditory phonetics), or how sounds are transmitted (acoustic phonetics) (Ashby, 2013). The latter concentrates specifically on the acoustic characteristics (or physical properties) of speech, such as frequency, amplitude, and duration, which can be objectively measured by analyzing acoustic waveforms. A waveform is a graphical representation of a sound wave, in which the variation in air pressure (y-axis) involved with the production of sound is plotted over time (x-axis) (Jurafsky & Martin, 2009). It is generally assumed that one's affective state is reflected by objectively measurable voice cues. As such, acoustic phonetics are considered the most promising phonetic features in examining affect and emotion (Juslin & Scherer, 2005).

Lefter et al. (2011) divide acoustic features into prosodic, spectral, and voice quality features. Studies in which the identification of emotions or affective state

Table 5.1: Feature overview

Feature	Description
N-grams	Text representation schemes such as the bag-of-words model for unigrams (single words) or language-model based schemes like N-grams or N-multigrams.
POS tags	Grammatical tags that classify words in their "parts-of-speech" and assign a label (tag) from a collection of tags (the tagset)[a].
LIWC categories	Lexicon-based tags captured by LIWC[b], which categorizes words as linguistic elements, emotions, and cognitive processes.
NRC emotion categories	Eight emotions and two sentiment categories captured using the general purpose NRC emotion lexicon[c].
Custom tags	Custom tags are used to tag words or word patterns (e.g., specific expressions) in the transcripts that met a specified set of words or phrases.
Text characteristics	General descriptive features that capture information on the overall text structure and general characteristics.
Pitch	Perceived pitch is objectively measured by its acoustic correlate, fundamental frequency (F0)[d].
Loudness	Perceived loudness is gauged by speech intensity, which objectively measures the energy in the acoustic signal.
Duration	Duration covers the temporal aspects of speech, which are tempo (speaking rate) and pause.
Spectral features	Frequency based features that represent the different frequencies (called "spectrum") that together make up the acoustic waveform[e].
Voice quality features	Perceived voice quality is measured by high-frequency energy (HF); the relative proportion of energy in an acoustic signal above versus below a specific frequency, and formant frequencies[d].
Turn statistics	General overall speech features that gauge language strength (poverty of speech) and structural organization[f].

Note. More details are provided in Appendix 5-A and Appendix 5-B.
[a] Bird et al. (2009)
[b] Linguistic Inquiry and Word Count program, Pennebaker et al. (2001)
[c] NRC emotion lexicon Mohammad and Turney (2010, 2013)
[d] Juslin and Scherer (2005)
[e] Jurafsky and Martin (2009)
[f] Orimaye et al. (2014)

plays a role mostly depend on prosodic features. Prosody refers to a collection of acoustic features that concern intonation-related (pitch), loudness-related (intensity), and tempo-related (e.g., durational aspects, speaking rate) features (Jurafsky & Martin, 2009). This can closely contribute to meaning and may reveal information normally not captured by textual features, such as emotional state or attitude (Wilson & Wharton, 2006).

Prosodic features generally cover speech units larger than one segment, such as syllables, words, or speaker turns, and are therefore also termed suprasegmentals (Jurafsky & Martin, 2009). The suprasegmentals pitch, loudness, and duration (tempo and pause) are among the most used features in the phonetic study of prosody (see e.g., the prosodic frameworks of Ladd & Cutler, 1983; Roach, 2000; Schoentgen, 2006). Several recent clinical studies used suprasegmental features for the diagnosis of a range of psychological disorders (S. Scherer et al., 2013), or specific disorders such as PTSD (Vergyri et al., 2015) and dementia (Fraser et al., 2014; Jarrold et al., 2014). Other purposes for which suprasegmentals have been used include identifying indicators for PTSD therapy progress (Van Den Broek et al., 2009) and assessing depression severity during therapeutic intervention (Lamers et al., 2014).

In addition to prosodic features, spectral features such as Mel-frequency cepstral coefficients (MFCCs) are commonly used in emotion detection as these are affected by emotional arousal (Lefter et al., 2011). Voice quality features such as high-frequency energy (HF) are found to be strongly related to emotions as well. Apart from neutral, voice qualities can be for example breathy, creaky, harsh, tense, or whispery. Finally, overall speaker turn statistics (e.g., turn length, the number of utterances per turn) were extracted as these can gauge language strength (poverty of speech) and structural organization (Orimaye et al., 2014).

The audio data was analyzed based on the prosodic features pitch, loudness, and duration, which are the most commonly used voice cues (Juslin & Scherer, 2005), acoustic parameters related to spectral and voice quality features, and turn statistics. The prosodic, voice quality, and general features were extracted at speaker turn and utterance level, the spectral features at frame level. In three segments not all speech features could be extracted at patient level because these segments contained no or only one voiced segment, due to which no SDs could be calculated for the concerning speech features. For these missing values, overall averages for the concerning classes (hotspot or non-hotspot) were imputed. More information on the used speech features, their relation to the operational constructs, and the extraction process is given in Appendix 5-B.

Feature union

Table 5.1 shows all extracted text and speech features. These features consist of a mixture of scales and quantities (e.g., normalized term and category frequencies, overall text statistics, mean amplitude values, and duration measures). Feature rescaling was done to make sure all input features have the same scale. This is preferred for many machine learning applications, to prevent features measured in greater numeric ranges from dominating features measured in smaller ranges. As

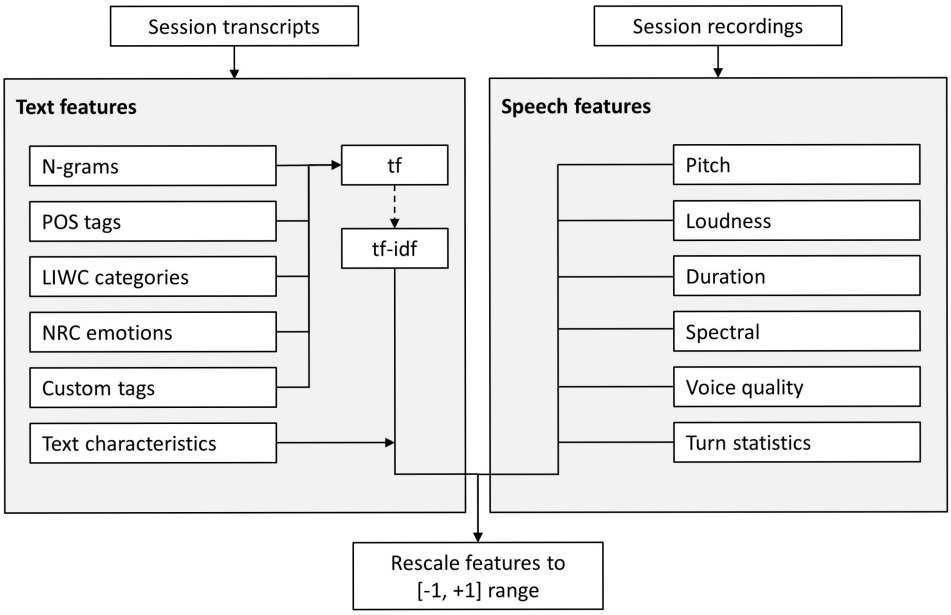


Figure 5.4: Rescaling process applied to extracted text and speech features before feature selection.

such, each text and speech feature was rescaled to the [-1, +1] range, as proposed by Hsu et al. (2003), so that each feature's maximal absolute value is equal to one (see Figure 5.4). This same scaling method is later applied to rescale the features in the test set.

Feature selection

The most informative features are selected using Pearson's chi-squared (χ^2) test, an effective feature selection metric (Yang & Pedersen, 1997) often used in text classification tasks. A more thorough explanation of χ^2 feature selection can be found in Oakes et al. (2001) or Manning et al. (2008). The χ^2-test compares the observed and expected feature occurrences in the hotspot versus non-hotspot phases. All features are then ranked based on their χ^2-scores and the k features with the highest χ^2-scores are selected for the final classification model (see Chapter 3 for a complete description of the process).

Excluded features Some state that stop words should not be included in the classification model, because these words do not add to the meaning of text (Jurafsky & Martin, 2009; Perkins, 2014). Other studies found that stop words such as particles and pronouns may indicate health improvements (Campbell & Pennebaker, 2003). Since we expected particles and pronouns to be related to the construct fragmentation, we think stop words should not simply be excluded without further investigation, even if some (such as particles and pronouns) are also captured by the POS tagger.

Table 5.2: Confusion matrix to assess model performance

	Predicted class	
True class	Positive (C_{HS})	Negative (C_{nHS})
Positive (C_{HS})	**True positive** (tp)	False negative (fn)
Negative (C_{nHS})	False positive (fp)	**True negative** (tn)

Note. Comparison of true (rows) and predicted (columns) class labels for the positive (hotspot) class C_{HS} and the negative (non-hotspot) class C_{n-HS}. The values on the diagonal (in boldface) show the correctly predicted class labels.

To avoid needlessly large feature sets, other words that were considered for exclusion were words that only occur in very few documents (Joachims, 1998). This was assessed through minimal document frequency; the minimal number of different training documents a word occurs in.

Machine learning algorithm

The extracted text and speech feature sets were used to train a support vector machine (SVM; Vapnik, 1995). SVMs are found to be among the best performing, most robust classification algorithms that can deal well with high-dimensional or imbalanced data sets (Joachims, 1998). We used the "C-Support Vector Classifier" (SVC) with a linear kernel, implemented in Scikit-learn's LIBSVM library (Chang & Lin, 2011). Two hyperparameters needed to be set; the kernel parameter γ, which we set to linear as is commonly done in text classification tasks, and the regularization parameter C, for which we compared different values in the parameter grid search.

Our classification task was a two-class problem; we wanted to distinguish hotspot phases from non-hotspot phases based on patient speech, defining hotspot phases as the positive class. To compensate for possible class imbalance we balanced class weights to be inversely proportional to the class sizes within the total data set, as in King and Zeng (2001).

Classification performance In the training phase the most informative features were extracted and selected for the final classification model. In the prediction phase, the occurrences of those selected features were used to predict a class label for each new input file. The model's classification performance was measured by comparing the true (known) labels of each input feature set with the predicted label for that feature set. Labels were predicted by applying the decision function resulting from the training phase to the segments present in the test set. The segments were given a positive label ("hotspot") if the decision function resulted in a value > 0, and a negative label ("non-hotspot") otherwise (see Alpaydin, 2004, for an extensive description of the decision function and optimization problems involved when using SVMs).

The instances in the true and predicted classes can be included respectively in

Table 5.3: Performance metrics and functions

Metric (M)	Description	Function
Accuracy	Proportion of correctly classified segments	$\frac{tp + tn}{tp + fn + fp + tn}$
Precision	Proportion of correctly identified positive segments	$\frac{tp}{tp + fp}$
Recall	Proportion of positive segments identified	$\frac{tp}{tp + fn}$
F_1-score	Harmonic mean of precision and recall	$2 \times \frac{precision \times recall}{precision + recall}$

Note. tp = true positives for each class, where true and predicted label are both positive. tn = true negatives for each class, where true and predicted label are both negative. fp = false positives for each class, where true label is negative but predicted label is positive. fn = false negatives for each class, where true label is positive but predicted label is negative.

the rows and columns of a confusion matrix, as displayed in Table 5.2. The cells on the diagonal contain the number of correctly predicted labels (true positives and true negatives), the errors (false positives and false negatives) are in the off-diagonal cells (Bird et al., 2009). We used the correct and false predictions to calculate the classification performance metrics accuracy, precision, recall, and F_1-score (see Table 5.3 for definitions) for the positive class. Of these, accuracy and F_1-score are the most commonly used in the evaluation of supervised classification models, although the F_1-score is the most suitable to deal with possible class imbalance. We will report all performance scores for each class and the weighted average over both classes, in which the performance scores of both classes are macro-averaged (Yang, 1997) and weighted by class size.

Analytical strategy

We adopted a nested k-fold cross-validation (CV) strategy, iterating over alternating subsets of data (folds) to train, validate, and test the model in order to prevent model evaluation bias. In the inner loop, a 10-fold CV grid search was conducted on the training set, calculating training performance on the validation set to find the optimal combination of (hyper) parameter settings. In the outer loop, the selected model was trained on the complete development set (consisting of the training plus the validation set), calculating the testing performance on the held-out test set to evaluate model generalizability. We will report on the selected parameters and features for the model with the highest testing performance, which was selected as the final hotspot classification model.

Because each session was split in several hotspot and non-hotspot segments, the data set contained multiple labeled segments per patient. To prevent the machine learning algorithm from learning patient specific features instead of class specific features, we used Scikit-learn's group-K-fold sampling strategy in both cross-validation loops. This strategy splits the folds in such a way that data of the same

patient will not simultaneously occur in the training as well as the test set.

Parameter grid search To find the best performing combination of parameter settings and features, an exhaustive grid search guided by the F_1-score was conducted in which all possible parameter combinations (within the set ranges) were fitted on the data set. The following parameters and parameter values were compared:

- Stop word removal: because there is no clear consensus on stop word removal, we included this as a parameter in the grid search. Stop words are either included or excluded using the Dutch stop word list from the NLTK library. This list includes 101 words, an overview can be found in Appendix 5-A.
- Minimal document frequency: we compared the effect of only including N-grams that occurred in at least one, two, or three separate training segments.
- Representation schemes: we compared four N-gram representation schemes: unigrams (1,1), bigrams (2,2), trigrams (3,3), and 3-multigrams (1,3).
- Term weights: we compared weighting textual content features by tf versus $tf\text{-}idf$.
- Select k best features: we compared different cut-off points (k) for the number of features to be included in the model based on the χ^2 feature selection metric. We compared values in the range 10-500 (increasing with 20 features each time) and all available features.
- Regularization parameter C: the values 1, 2, 3, 100, and 1,000 were compared.

To compare the performance of text features with that of speech features and text and speech features combined, the complete model development pipeline, including nested k-fold cross-validation and exhaustive grid search, was run three times. This resulted in three trained and tested models; one text only model, one speech only model, and one multimodal model. The model with the highest training performance was selected as the final model.

5.3. Results

5.3.1. Sample characteristics

In total, the selected recordings contained around 6.5 hours of imaginal exposure speech, of which over 2 hours of "hotspot" speech (mean hotspot length $\approx$ 3.5 minutes) and over 4 hours of "non-hotspot" speech (mean non-hotspot length $\approx$ 5.5 minutes). Of the hotspot speech 70% is uttered by the patient, for the non-hotspot speech this is 78%. On average, the non-hotspot segments contain almost twice as many speaker turns and utterances as the hotspot segments, although the high SDs show there are large differences between segments. The number of

word types and tokens show that patients use more unique words in non-hotspots than in hotspots, and that patient speech has a higher pitch in hotspots than in non-hotspots. A summary of the main hotspot and non-hotspot characteristics is given in Table 5.4.

Table 5.4: Summary of characteristics hotspots, non-hotspots, and total sample

Characteristics	Hotspots (N = 37)	Non-hotspots (N = 45)	Total (N = 82)
Record length[a]	02:14:03	04:18:06	06:32:09
Mean duration[a]	00:03:37	00:05:44	00:04:46
Speaker turns, M(SD)	24.43(21.52)	40.13(45.38)	33.05(37.23)
Utterances, M(SD)	27.22(16.71)	47.22(34.43)	38.20(29.47)
Word tokens, M(SD)	259.62(187.90)	546.69(478.98)	417.16(401.21)
Word types, M(SD)	104.11(44.79)	170.47(95.47)	140.52(83.35)
Type:Token Ratio, M(SD)	0.47(0.12)	0.42(0.16)	0.44(0.14)
Words per turn, M(SD)	16.49(13.99)	20.27(17.14)	18.57(15.82)
Word length, M(SD)	3.91(0.21)	3.95(0.17)	3.93(0.19)
Honoré's R, M(SD)	606.31(105.95)	636.88(142.64)	623.09(127.58)
Flesch-Douma G, M(SD)	110.44(6.36)	107.46(7.73)	108.80(7.26)
Brunét's index, M(SD)	12.40(1.73)	13.14(2.59)	12.80(2.26)
Patient speech length[a]	1:33:52	03:21:05	04:54:58
Sounding[a]	00:44:17	01:58:21	02:42:39
Mean duration[a]	00:01:11	00:02:37	00:01:59
Silent[a]	00:49:35	01:22:43	02:12:18
Mean duration[a]	00:01:20	00:01:50	00:01:36
Pitch, M(SD)	253.24(67.86)	231.00(61.61)	241.03(65.06)
Intensity, M(SD)	60.06(5.06)	59.20(6.12)	59.58(5.65)
Speech rate, M(SD)	1.35(0.68)	1.57(0.71)	1.47(0.70)
Articulation rate, M(SD)	3.65(0.74)	3.56(0.60)	3.60(0.66)
Phonation rate, M(SD)	0.37(0.18)	0.44(0.19)	0.41(0.18)
Speech productivity, M(SD)	1.41(1.65)	1.03(0.92)	1.20(1.30)

Note. Except for the number of speaker turns and record length, all characteristics take into account patient speech only.

[a] hr:min:sec

Validation splits

The total data set consisted of data of 20 patients. In the outer loop of the nested k-fold cross-validation process, the data set was iteratively split into ten development and test sets. The development sets consisted of the hotspot and non-hotspot segments of 18 patients (90% of the total sample), and the test sets of the remaining 10% (two patients). An exhaustive grid search was conducted on the development

set in the inner loop, during which the development set again was iteratively split into ten training and test sets, respectively consisting of 90% (16 or 17 patients) against 10% (two or one patients) of the development data.

5.3.2. Model comparison

We developed three different models; the first model was based only on text features, the second model used only speech features, and the third (multimodal) model consisted of text and speech features combined. This section reports the mean training performance of each model. The mean testing performance for all three models is discussed under Generalizability.

Text features only

The model based exclusively on text features was trained using N-grams, N-multigrams, the 96 lexicon-based, POS, or custom tags, and the general text characteristics included in Appendix 5-A. The exhaustive grid search resulted in a mean training F_1-score of 0.75 (SD 0.03) for the hotspot class. This is a good classification performance, and the low SD shows that the grid search results are stable with little variation over the different folds. The model with the highest testing performance resulted in a reasonable precision (0.60), perfect recall (1.00), high F_1-score (0.75), and a high classification accuracy (0.75). This model consisted of N-multigrams ranging from one to three words weighted by the $tf\text{-}idf$ scheme. Among the most informative features for the hotspot class are words and word combinations such as "neck", "terrible", and "no no no". The best text model was based on only 10 N-multigrams; general text features, lexicon-based features and POS tags were not among the most informative features selected by the grid search.

Speech features only

For the speech feature only model, 111 extracted speech features (see Appendix 5-B for an overview) were compared in the exhaustive CV grid search. The mean training F_1-score resulting from the exhaustive grid search was 0.62 (SD 0.03) for the hotspot class. This is a reasonable performance score, although lower than that of the text only model. Like the text only model, the low SD points to stable grid search results over the folds in the inner loop. The model with the highest testing results was based on ten speech features selected by the grid search and had a good precision (0.75), recall (0.75), and F_1-score (0.75), and an overall classification accuracy of 0.75. The five most informative hotspot (marked by *) and non-hotspot features for this model are displayed in the index graph in Figure 5.5. This graph shows the change in each feature for the consecutive hotspot and non-hotspot segments compared to the base value of that feature at the start of the exposure session.

Text and speech features combined

When using both text and speech features, the mean training F_1-score was 0.76 (SD 0.04) for the hotspot class. As for the speech only model, the combined model with the highest testing performance had a good precision (0.75), recall (0.75), and

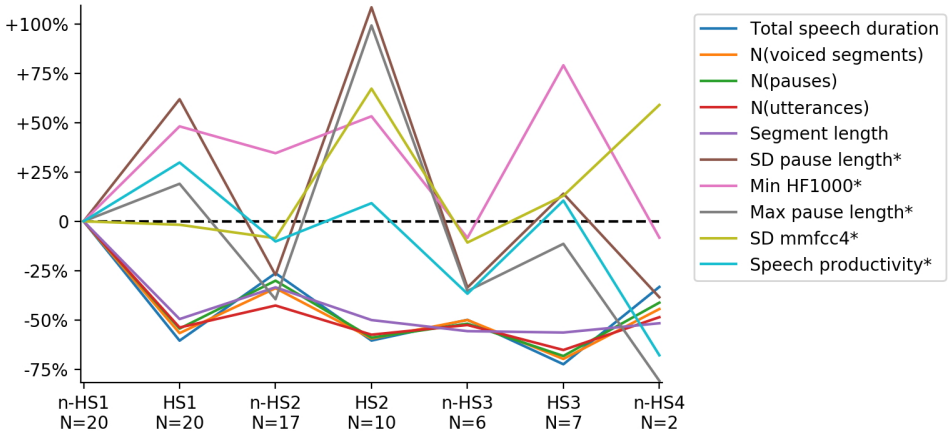


Figure 5.5: Five most informative speech features for hotspots (*) and non-hotspots

F_1-score of 0.75. The overall training accuracy of the multimodal model was slightly better than for the best text only and speech only models, namely 0.78.

5.3.3. Final model

The multimodal model was selected as the final model because this resulted in the highest training F_1-score for the hotspot class and overall accuracy. This model consisted of 310 text and speech features, where the text features were $tf\text{-}idf$ weighted trigrams that occurred in at least two different segments in the training set.

Most informative features

Of the fifty most informative features, three are speech features, seven are LIWC features, two are features extracted through the NRC emotion lexicon, one is a POS tag, one is a custom tag, one is a text statistic, and the remaining 35 are trigrams. To illustrate the occurrences of different feature types in both classes, Table 5.5 shows a selection of 25 highly informative features that were included in the model.

Confusion matrix

The confusion matrix in Table 5.6 shows the number of correctly versus erroneously predicted labels for the hotspots and non-hotspots present in the test set. This shows that on average the model labeled three of the four hotspots correctly, and four of the five non-hotspots. It incorrectly labeled one hotspot as a non-hotspot and vice versa.

Generalizability

The model generalization is the average testing performance over all test sets in the outer loop of the nested k-fold CV. This shows how well a model trained and

Table 5.5: Selection of most informative features of the multimodal classifier

Feature	χ^2	P	Hotspots	Non-hotspots
Nee nee nee (no no no)[a]	23.347	0.127	**4**	1
Angst euh euh (fear uh uh)[a]	23.060	0.129	**2**	0
War euh war (were uh were)[a]	22.071	0.137	0	**5**
Category "Disgust"[b]	21.840	0.139	**0.97**	0.44
Category "Death"[c]	21.099	0.146	**0.23**	0.04
Pijn helemal nik (pain absolutely nothing)[a]	20.692	0.150	**2**	0
Weg vlucht euh (away flight uh)[a]	20.692	0.150	**2**	0
Zeg euh euh (say uh uh)[a]	20.408	0.153	0	**11**
Emotional expressions[d]	18.663	0.172	**8.02**	1.71
Category "Negative emotions"[c]	17.905	0.181	**2.30**	1.22
Category "Interrogative pronoun"[e]	17.879	0.181	0.00	**0.04**
Category "Anger"[c]	17.803	0.182	**0.46**	0.19
Absolute word count (word tokens)[e]	17.498	0.186	245.85	**521.12**
Bang dod gan (afraid to die)[a]	17.443	0.187	**3**	0
Category "Sadness"[c,*]	17.192	0.190	**0.69**	0.23
Euh soort euh (uh sort uh)[a]	17.138	0.190	0	**4**
Zeg euh kom (say uh come)[a]	17.003	0.192	**2**	0
Ging ging wer (went went again)[a]	16.500	0.199	0	**2**
Category "Anxiety"[c]	16.249	0.202	**0.77**	0.37
Category "Sadness"[b,*]	15.045	0.220	**1.80**	1.05
Number of voiced units[f]	15.043	0.220	7.58	**72.05**
Category "Eating"[c]	15.038	0.220	0.05	**0.17**
Number of silent units[f]	14.569	0.227	5.79	**66.95**
Total duration of speech[f]	14.543	0.228	8.68	**47.72**
Category "Swear words"[c]	14.388	0.230	**0.12**	0.02

Note. 25 of the 50 most informative features, based on χ^2 ranking. The first column shows a selection of high ranked features. *N*-grams are Dutch and stemmed (hence might seem misspelled; e.g.,"dood" is stemmed to "dod", and "gaan" to "gan"), with unstemmed English translations in parentheses. The remaining columns show occurrence counts and means for both classes. Values for the class with the highest occurrence are in boldface.

[*] Sadness is listed twice: the first is the LIWC category and the second is the NRC emotion.

[a] *N*-gram of 3 consecutive words.

[b] Emotion feature extracted using the NRC emotion lexicon.

[c] LIWC feature extracted using the LIWC dictionary.

[d] Emotional expressions extracted using custom tagger.

[e] Text statistic extracted using Python's TextStat package.

[f] Speech feature extracted using Praat.

Table 5.6: Confusion matrix to assess model performance

	Predicted class	
True class	Hotspot	Non-hotspot
Hotspot	**3**	1
Non-hotspot	1	**4**

Note. Comparison of true (rows) and predicted (columns) class labels for the hotspot and the non-hotspot class. The values on the diagonal (in boldface) show the correctly predicted class labels.

validated on the labeled input data predicts the correct output for new, future data (Alpaydin, 2004). The testing performance for the final (multimodal) model was lower than the training performance (see Table 5.7), which means the developed model will not generalize well to new data. This was also the case for the best performing text only and speech only models. However, the models in which is made use of speech features (the speech only and text and speech features combined) seem to be slightly more robust than the model based only on text features. Since the text only model was based on only ten N-multigrams, it could be that the selected features for the text model were too specific.

5.4. Discussion

The aim of this study was to examine if it was possible to automatically recognize hotspots in patients undergoing a trauma-focused treatment for PTSD. We hypothesized that a combination of text and speech features extracted from recorded and transcribed patient speech could be used to develop a supervised classification model to automatically distinguish between hotspot and non-hotspot phases during imaginal exposure sessions. Based on the formal hotspot characteristics and previous research on hotspots and CBT sessions, we identified nine constructs that we expected to differ between hotspots and non-hotspots. We expected that hotspots would contain more affect, avoidance, dissociation, fragmentation, emotions, and cognitions, and less organization, cohesion, and complexity. These nine constructs were operationalized through a number of text and speech characteristics that were captured using a large range of features extracted from CBT session transcripts and recordings, as shown in Figure 5.2.

The results showed that text and speech features related to these constructs could indeed be used to train a stable model to distinguish between hotspots and non-hotspots within the current data set. The models consisting of text features alone or text and speech features combined resulted in the highest training performance. The training performance of models based on speech features alone was lower. However, clear fluctuations in speech features over the hotspot and non-hotspot segments were found. The high training performance shows that we were able to develop a model based on text and speech features that could classify the

Table 5.7: Mean testing performance

Class	Precision	Recall	F_1-score	Accuracy	N(segments) in test set
			Text features only		
Hotspots	0.443	0.675	0.530		4
Non-hotspots	0.652	0.435	0.469		5
Weighted average/Total(N)	0.568	0.546	0.501	0.545	9
			Speech features only		
Hotspots	0.543	0.592	0.534		4
Non-hotspots	0.603	0.560	0.553		5
Weighted average/Total(N)	0.586	0.565	0.543	0.566	9
			Multimodal (text and speech features)		
Hotspots	0.464	0.617	0.525		4
Non-hotspots	0.594	0.495	0.512		5
Weighted average/Total(N)	0.543	0.556	0.522	0.555	9

Note. Per class and average performance scores for the final models.

hotspot and non-hotspot segments included in the current data set very well.

The feature overview in Table 5.5 shows that many of the selected features are related to the construct Emotions (e.g., emotion categories disgust, anger, sadness, and anxiety, as well as audible emotional expressions such as sniffing and sighing). This was in line with our expectations, as hotspots are considered the most emotional moments in trauma (Nijdam et al., 2013) and emotions are found to occur more frequently in hotspot than in non-hotspot phases (Holmes et al., 2005). Moreover, the strong, clearly distinguishable dictionary-based features and audible cues that were used to capture emotions may have benefited their recognition.

Table 5.5 further shows that the LIWC category Sadness was slightly more discriminative than Sadness captured using the NRC lexicon. This could be because the Dutch LIWC dictionary is validated (Zijlstra et al., 2004), whereas the NRC categories were simply converted to Dutch using Google Translate. However, both Sadness categories were discriminative enough for inclusion in the final model. The added value of the NRC dictionary is mainly in the fact that it distinguishes more emotional categories than LIWC, such as the category Disgust, which is also included in the model. Despite this extended range of emotions, two emotions

defined by Holmes and colleagues as characterizing for hotspots, namely guilt and horror (the latter of which was also an explicit PTSD criterion of the DSM-4-TR; American Psychiatric Association, 2013), were not covered by the lexicons used. Expanding the emotion lexicon with dictionaries for guilt and horror might improve classification performance.

Psychological theories explaining the working mechanisms underlying PTSD treatments (see Nijdam & Wittmann, 2015), state that trauma memories are represented differently than ordinary memories (e.g., lacking spatial or temporal context, or inadequately integrated with broader memories). As exposure aims to re-encode and restructure the trauma memory in such a way that it no longer evokes the feeling of current threat, successful treatment should result in more integrated, cohesive, and less fragmented trauma narratives, indicating adequate processing of the trauma (Brewin et al., 1996; Ehlers & Clark, 2000; Foa & Rothbaum, 1998). However, only a few features related to organization, cohesion, or fragmentation were included in the model, for example the use of interrogative pronouns (related to Cohesion), the absolute word count and the frequent presence of the speech filler 'uh' in the selected N-grams (indicators of Fragmentation), the number of voiced and silent units, and the total duration of speech (to capture Avoidance). This could be because for some features, changes in opposite directions may be indicative of different hotspot related constructs (e.g., an increased speech rate is related to Avoidance, whereas a decreased speech rate may indicate Emotions). This may reduce these features' discriminative power. Another reason could be that some hotspot characteristics based on which we defined the set of constructs and features to be extracted, did not occur (frequently) in our data set. For example the change from first to third person, which is a clear identifyer for hotspots, did not take place in any of the sessions.

The low testing performance shows that the selected model does not generalize well to new data sets. Since we tried to fit a complex model with a large number of parameters to a small data set, the low testing performance most likely indicates overfitting (also called overtraining). This means that the selected model has not only learned the underlying structure but also the noise present in the training data (Alpaydin, 2004). Another reason for overfitting could be that the noisy audio data impeded accurate extraction of speech features.

Several studies have shown that emotions and mood influence speaking behavior and speech sound characteristics (Kuny & Stassen, 1993; K. R. Scherer et al., 2003). As acoustic features can be used in detecting conditions in which changes in speech are common (Fraser et al., 2014), one could also expect these features to detect moments in which changes in speech occur, such as hotspots. Therapy session recordings and transcripts hold a lot of information. Text and audio analysis can help to extract and process this information in a structured, efficient, and reproducible way. Moreover, the collection and analysis of text and audio data can be considered to be non-, or at least less, obtrusive than for example questionnaire-based research or biosignal analysis (which requires sensors to be attached to a patient; Van Den Broek et al., 2009). Given that lots of therapy data may already be recorded and processed as part of the standard treatment procedure, for thera-

pist training and ongoing research, or as part of e-health interventions (e.g., Bourla et al., 2018; Olff, 2015; Rizzo & Shilling, 2017; Wild et al., 2016), it is worth exploring how these available data can be made of further value.

It should be noted that most studies on emotion classification and vocal affect expression are based on clean, artificial data in which emotions are portrayed by actors (Juslin & Scherer, 2005) in simple and short utterances (Cowie et al., 2001). The data used in the current study contains raw, authentic emotions embedded in a broader context, from people with different backgrounds who experienced different types of trauma, which is more in line with the real world. As such, our data set can be considered highly ecologically valid and valuable not only for psychiatric research and practice but also for studies on speech sounds and emotion recognition (Van Den Broek et al., 2009). However, this strong point is also a huge limitation. Although reusing existing data sets seems efficient and durable, it also introduces challenges. The biggest challenge is the background noise due to simple recording equipment and the transitory nature of analog recordings, which reduced the recording quality over the years. Due to this it was not possible to use automatic speech recognition and session content needed to be transcribed manually, which remained impossible for small parts of the recordings even after noise reduction.

Another limitation is a methodological one. Because we had such a small data set, we chose not to waste any information by holding out a part of the data for model testing and validation. Instead we used nested cross-validated grid search, a standard tool included in Scikit-learn. This tool does not provide the option to remove keywords with an occurrence frequency of lower than five in the training set (which is suggested in some studies, e.g., Manning & Schütze, 1999, to ensure reliability of the χ^2 calculation).

Despite these limitations we developed a hotspot classification model with high training performance, meaning that the model could clearly distinguish between the hotspots and non-hotspots present in our data set. However, the low testing performance indicates that the model will have difficulty recognizing hotspots from new input data. This is probably due to the application of a complex training strategy using many different features on a relatively small, low quality, but ecologically valid data set. Another resaon could be that the patient characteristics and trauma types present in our data set may have influenced speech characteristics and word use, and as such the features included in the model. This should be studied in more detail on a larger data set. The techniques used lend themselves well to application on larger data sets, and current audio recording equipment makes it easier to collect and process high-quality audio data which can be transcribed automatically using automatic speech recognition. This way, much larger sets of therapy session transcripts and recordings can be generated. Because this study only used text and speech features that could be automatically extracted it is very easy to train and test a new hotspot recognition model on new data using the same constructs, which we expect to improve model generalizability.

Although model performance needs to be improved, this type of research has the potential to advance theories about effective treatment elements in the context of trauma treatment. The automatic recognition of hotspots may aid in the comparison

of hotspot characteristics for different patient groups, trauma types, or dropouts to investigate potential mediators of treatment success as suggested by Nijdam et al. (2013). In addition, clinicians can gain more insight in the occurrence and characteristics of hotspots and the way hotspots are addressed, which might assist them in offering a more effective intervention to patients that otherwise would not respond sufficiently to treatment (Nijdam & Wittmann, 2015).

Because of the low generalizability, the current study should merely be seen as a proof of concept, showing the technical and practical feasibility and possibilities of text and audio mining for research on trauma treatment processes and mental health research in general. Future research should focus on applying this method to larger, higher quality data sets before more general conclusions can be drawn. Still we want to emphasize the added value and potential of the used methods and data for future research. For clinical practice, in the future this work may benefit the patient because these types of models can provide the therapist with (direct) automated feedback, which allows for more precise and unobtrusive monitoring of treatment progress.

5.5. References

Alpaydin, E. (2004). *Introduction to machine learning*. MIT Press.

Alvarez-Conrad, J., Zoellner, L., & Foa, E. (2001). Linguistic predictors of trauma pathology and physical health. *Applied Cognitive Psychology*, *15*, 159–170. https://doi.org/10.1002/acp.839

Aman, S., & Szpakowicz, S. (2007). Identifying expressions of emotion in text. In V. Matoušek & P. Matner (Eds.), *Text, speech and dialogue. TSD 2007. Lecture notes in Computer Science, vol 4629* (pp. 196–205). Springer. https://doi.org/10.1007/978-3-540-74628-7_27

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders, Text Revision (4th edition)*.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*.

Amir, N., Stafford, J., Freshman, M., & Foa, E. (1998). Relationship between trauma narratives and trauma pathology. *Journal of Traumatic Stress*, *11*(2), 385–392. https://doi.org/10.1023/A:1024415523495

Ashby, P. (2013). *Understanding phonetics*. Routledge.

Audacity Team. (2013). Audacity(R) software is copyright (c) 1999-2016 Audacity Team. The name Audacity(R) is a registered trademark of Dominic Mazzoni. http://audacityteam.org/

Bansal, S., & Aggarwal, C. (2018). Textstat [Python package]. https://pypi.org/project/textstat/

Bedi, G., Carrillo, F., Cecchi, G., Slezak, D., Sigman, M., Mota, N., Ribeiro, S., Javitt, D., Copelli, M., & Corcoran, C. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *Schizophrenia*, *1*(1), 15030. https://doi.org/10.1038/npjschz.2015.30

Bekkerman, R., & Allan, J. (2003). *Using bigrams in text categorization* (Report IR-408). Center for Intelligent Information Retrieval, UMass. Amherst, MA, University of Massachusetts. http://ist.psu.edu

Bhaskar, J., Sruthi, K., & Nedungadi, P. (2015). Hybrid approach for emotion classification of audio conversation based on text and speech mining. *Procedia Computer Science*, *46*, 635–643. https://doi.org/10.1016/j.procs.2015.02.112

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'reilly Media, Inc.

Bisson, J. I., Berliner, L., Cloitre, M., Forbes, D., Jensen, T. K., Lewis, C., Monson, C. M., Olff, M., Pilling, S., Riggs, D. S., et al. (2019). The international society for traumatic stress studies new guidelines for the prevention and treatment of posttraumatic stress disorder: Methodology and development process. *Journal of Traumatic Stress*. https://doi.org/10.1002/jts.22421

Boals, A., & Klein, K. (2005). Word use in emotional narratives about failed romantic relationships and subsequent mental health. *Journal of Language and Social Psychology*, *24*(3), 252–268. https://doi.org/10.1177/0261927X05278386

Boersma, P., & Weenink, D. (2019). Praat: doing phonetics by computer [Computer program]. http://www.praat.org/

Bormuth, J. R. (1969). *Development of standards of readability: Toward a rational criterion of passage performance* (tech. rep.). University of Chicago. Chicago.

Bourla, A., Mouchabac, S., El Hage, W., & Ferreri, F. (2018). e-PTSD: An overview on how new technologies can improve prediction and assessment of Posttraumatic Stress Disorder (PTSD). *European Journal of Psychotraumatology*, *9*(sup1), 1424448. https://doi.org/10.1080/20008198.2018.1424448

Bradley, R., Greene, J., Russ, E., Dutra, L., & Westen, D. (2005). A multidimensional meta-analysis of psychotherapy for PTSD. *American Journal of Psychiatry*, (162), 214–227. https://doi.org/10.1176/appi.ajp.162.2.214

Brewin, C., Dalgleish, T., & Joseph, S. (1996). A dual representation theory of posttraumatic stress disorder. *Psychological Review*, *103*(4), 670. https://doi.org/10.1037//0033-295X.103.4.670

Campbell, R., & Pennebaker, J. (2003). The secret life of pronouns: Flexibility in writing style and physical health. *Psychological Science*, *14*(1), 60–65. https://doi.org/10.1111/1467-9280.01419

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 1–39. https://doi.org/10.1145/1961189.1961199

Chung, C., & Pennebaker, J. (2007). The Psychological Functions of Function Words. In K. Fiedler (Ed.), *Social communication* (pp. 343–359). Psychology Press.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, (January).

Croisile, B., Ska, B., Brabant, M., Duchene, A., Lepage, Y., Aimard, G., & Trillet, M. (1996). Comparative study of oral and written picture description in patients with Alzheimer's disease. *Brain and Language*, *53*(1), 1–19. https://doi.org/10.1006/brln.1996.0033

Crossley, S., Kyle, K., & McNamara, D. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*(4), 1227–1237. https://doi.org/10.3758/s13428-015-0651-7

D'Andrea, W., Chiu, P., Casas, B., & Deldin, P. (2012). Linguistic predictors of posttraumatic stress disorder symptoms following 11 September 2001. *Applied Cognitive Psychology*, *26*(2), 316–323. https://doi.org/10.1002/acp.1830

De Lira, J., Ortiz, K., Campanha, A., Bertolucci, P., & Minett, T. (2011). Microlinguistic aspects of the oral narrative in patients with Alzheimer's disease. *International Psychogeriatrics*, *23*(3), 404–412. https://doi.org/10.1017/S1041610210001092

De Vries, G., & Olff, M. (2009). The lifetime prevalence of traumatic events and posttraumatic stress disorder in the Netherlands. *Journal of Traumatic Stress*, *22*(4), 259–267. https://doi.org/10.1002/jts.20429

Douma, W. (1960). De leesbaarheid van landbouwbladen: Een onderzoek naar en een toepassing van leesbaarheidsformules.

Ehlers, A., & Clark, D. (2000). A cognitive model of chronic post-traumatic stress disorder. *Behaviour Research and Therapy*, *38*(4), 319–45. https://doi.org/10.1016/S0005-7967(99)00123-0

Ehlers, A., Clark, D., Hackmann, A., McManus, F., & Fennell, M. (2005). Cognitive therapy for post-traumatic stress disorder: Development and evaluation. *Behaviour Research and Therapy*, *43*(4), 413–431. https://doi.org/10.1016/j.brat.2004.03.006

Ehlers, A., Hackmann, A., & Michael, T. (2004). Intrusive re-experiencing in post-traumatic stress disorder: Phenomenology, theory, and therapy. *Memory*, *12*(4), 403–415. https://doi.org/10.1080/09658210444000025

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221. https://doi.org/10.1037/h0057532

Foa, E., Molnar, C., & Cashman, L. (1995). Change in rape narratives during exposure therapy for PTSD. *Journal of Traumatic Stress*, *8*(4), 675–90.

Foa, E., & Rothbaum, B. (1998). *Treating the trauma of rape: Cognitive-behavioral therapy for PTSD*. Guilford Press.

Forbes-Riley, K., & Litman, D. (2004). Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, *3*, 1289–1305. http://www.jmlr.org/

Fraser, K., Hirst, G., Graham, N., Meltzer, J., Black, S., & Rochon, E. (2014). Comparison of different feature sets for identification of variants in progressive aphasia. *Workshop on Computational Linguistics and Clinical Psychology: From linguistic signal to clinical reality*, 17–26. https://doi.org/10.3115/v1/W14-3203

Gersons, B., Carlier, I., Lamberts, R., & Van Der Kolk, B. (2000). Randomized clinical trial of brief eclectic psychotherapy for police officers with posttraumatic stress disorder. *Journal of Traumatic Stress*, *13*(2), 333–347. https://doi.org/10.1023/A:1007793803627

Gersons, B., Meewisse, M., Nijdam, M., & Olff, M. (2011). Protocol - Brief eclectic psychotherapy for Posttraumatic Stress Disorder (BEP).

Gil, T., Calev, A., Greenberg, D., Kugelmass, S., & Lerer, B. (1990). Cognitive functioning in post-traumatic stress disorder. *Journal of Traumatic Stress*, *3*(1), 29–45. https://doi.org/10.1002/jts.2490030104

Gray, M., & Lombardo, T. (2001). Complexity of trauma narratives as an index of fragmented memory in PTSD: A critical analysis. *Applied Cognitive Psychology*, *15*(7 SPEC. ISS.), 171–186. https://doi.org/10.1002/acp.840

Grey, N., & Holmes, E. (2008). "Hotspots" in trauma memories in the treatment of post-traumatic stress disorder: A replication. *Memory*, *16*(7), 788–796. https://doi.org/10.1080/09658210802266446

Grey, N., Holmes, E., & Brewin, C. (2001). Peritraumatic emotional "hotspots" in memory. *Behavioural and Cognitive Psychotherapy*, *29*(03), 367–372. https://doi.org/10.1017/S1352465801003095

Grey, N., Young, K., & Holmes, E. (2002). Cognitive restructuring within reliving: A treatment for peritraumatic emotional "hotspots" in posttraumatic stress disorder. *Behavioural and Cognitive Psychotherapy*, *30*(1), 37–56. https://doi.org/10.1017/S1352465802001054

Halligan, S., Michael, T., Clark, D., & Ehlers, A. (2003). Posttraumatic stress disorder following assault: The role of cognitive processing, trauma memory, and appraisals. *Journal of Consulting and Clinical Psychology*, *71*(3), 419–431. https://doi.org/10.1037/0022-006X.71.3.419

He, Q., Veldkamp, B., & De Vries, T. (2012). Screening for posttraumatic stress disorder using verbal features in self narratives: A text mining approach. *Psychiatry Research*, *198*(3), 441–447. https://doi.org/10.1016/j.psychres.2012.01.032

Hellawell, S., & Brewin, C. (2004). A comparison of flashbacks and ordinary autobiographical memories of trauma: Content and language. *Behaviour Research and Therapy*, *42*(1), 1–12. https://doi.org/10.1016/S0005-7967(03)00088-3

Holmes, E., & Grey, N. (2002). *Hotspots manual: Third revision*. Camden; Islington Mental Health; Social Care Trust, The Traumatic Stress Clinic.

Holmes, E., Grey, N., & Young, K. (2005). Intrusive images and "hotspots" of trauma memories in posttraumatic stress disorder: An exploratory investigation of emotions and cognitive themes. *Journal of Behavior Therapy and Experimental Psychiatry*, *36*(1 SPEC. ISS.), 3–17. https://doi.org/10.1016/j.jbtep.2004.11.002

Howes, C., Purver, M., McCabe, R., Healey, P. G. T., & Lavelle, M. (2012). Predicting adherence to treatment for schizophrenia from dialogue transcripts. *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 79–83. http://dl.acm.org/citation.cfm?id=2392814

Hsu, C., Chang, C., & Lin, C. (2003). *A practical guide to support vector classification. [Technical Report]*. https://www.csie.ntu.edu.tw/~cjlin/

Iliou, T., & Anagnostopoulos, C. (2010). SVM - MLP - PNN classifiers on speech emotion recognition field - A comparative study. *5th International Conference on Digital Telecommunications, ICDT 2010*, 1–6. https://doi.org/10.1109/ICDT.2010.8

Jacewicz, E., Fox, R., O'Neill, C., & Salmons, J. (2009). Articulation rate across dialect, age, and gender. *Language Variation and Change*, *21*(2), 233–256. https://doi.org/10.1017/S0954394509990093

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M., & Ogar, J. (2014). Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 27–37. https://doi.org/10.3115/v1/W14-3204

Jelinek, L., Stockbauer, C., Randjbar, S., Kellner, M., Ehring, T., & Moritz, S. (2010). Characteristics and organization of the worst moment of trauma memories in posttraumatic stress disorder. *Behaviour Research and Therapy*, *48*(7), 680–685. https://doi.org/10.1016/j.brat.2010.03.014

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Proceedings of the 10th European Conference on Machine Learning*, 137–142. https://doi.org/10.1007/BFb0026683

Jones, C., Harvey, A., & Brewin, C. (2007). The organisation and content of trauma memories in survivors of road traffic accidents. *Behaviour Research and Therapy*, *45*(1), 151–162. https://doi.org/10.1016/j.brat.2006.02.004

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall.

Juslin, P. N., & Scherer, K. R. (2005). Vocal expression of affect. In J. Harrigan, R. Rosenthal, & K. Scherer (Eds.), *The new handbook of methods in nonverbal behavior research*. Oxford University Press.

Kessler, R. C., Aguilar-Gaxiola, S., Alonso, J., Benjet, C., Bromet, E. J., Cardoso, G., Degenhardt, L., De Girolamo, G., Dinolova, R. V., Ferry, F., et al. (2017). Trauma and PTSD in the WHO world mental health surveys. *European Journal of Psychotraumatology*, *8*(sup5), 1353383. https://doi.org/10.1080/20008198.2017.1353383

Kincaid, J. P., Fishburne Jr, R., Rogers, R., & Chissom, B. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Institute for Simulation and Training, University of Central Florida*, *56*. https://stars.library.ucf.edu/istlibrary/56

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *45*, 326–347. https://doi.org/10.1016/j.csl.2017.01.005

Kuny, S., & Stassen, H. (1993). Speaking behavior and voice sound characteristics in depressive patients during recovery. *Journal of Psychiatric Research*, *27*(3), 289–307. https://doi.org/10.1016/0022-3956(93)90040-9

Ladd, D., & Cutler, A. (1983). Introduction. Models and measurements in the study of prosody. In A. Cutler & D. Ladd (Eds.), *Prosody: Models and measurements. Springer Series in Language and Communication* (pp. 1–10). Springer. https://doi.org/10.1007/978-3-642-69103-4_1

Lamers, S., De Jong, F., Truong, K., Steunenberg, B., & Westerhof, G. (2014). Applying prosodic speech features in mental health care: An exploratory study in a life-review intervention for depression. *Workshop on Computational Linguistics and Clinical Psychology*, 61–68. https://doi.org/10.3115/v1/W14-3208

Lapp, D. (2006). *The pyshics of music and musical instruments*. Wright Center for Science Education Tufts University. http://kellerphysics.com/acoustics/Lapp.pdf

Lee, C., & Narayanan, S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, *13*(2), 293–303. http://ieeexplore.ieee.org/xpls/abs%7B%5C_%7Dall.jsp?arnumber=1395974

Lefter, I., Rothkrantz, L., Van Leeuwen, D., & Wiggers, P. (2011). Automatic stress detection in emergency (telephone) calls. *International Journal of Intelligent Defence Support Systems*, *4*(2), 148. https://doi.org/10.1504/IJIDSS.2011.039547

Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. *Proceedings of the International Symposium on Music Information Retrieval (ISMIR) 2000*, 1–11.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Mohammad, S., & Turney, P. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, Los Angeles, California*, 26–34. https://dl.acm.org/citation.cfm?id=1860635

Mohammad, S., & Turney, P. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, *29*(3), 436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Nelson, K., & Horowitz, L. (2010). Narrative structure in recounted sad memories. *Discourse Processes*, *31*(3), 307–324. https://doi.org/10.1207/S15326950dp31-3_5

Nijdam, M., Baas, M., Olff, M., & Gersons, B. (2013). Hotspots in trauma memories and their relationship to successful trauma-focused psychotherapy: A pilot study. *Journal of Traumatic Stress*, *26*, 38–44. https://doi.org/10.1002/jts.21771

Nijdam, M., & Wittmann, L. (2015). Psychological and social theories of PTSD. In U. Schnyder & M. Cloitre (Eds.), *Evidence based treatments for trauma-related psychological disorders: A practical guide for clinicians* (pp. 41–61). Springer International Publishing Switzerland. https://doi.org/10.1007/978-3-319-07109-1_3

NLTK Project. (2019). Copyright © 2019 NLTK Project. http://www.nltk.org/

Oakes, M., Gaaizauskas, H., Rand Fowkes, Jonsson, A., Wan, V., & Beaulieu, M. (2001). A method based on the chi-square test for document classification. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 440–441. https://doi.org/10.1145/383952.384080

O'Kearney, R., Hunt, A., & Wallace, N. (2011). Integration and organization of trauma memories and posttraumatic symptoms. *Journal of Traumatic Stress*, *24*(6), 716–725. https://doi.org/10.1002/jts.20690

O'Kearny, R., Speyer, J., & Kenardy, J. (2007). Children's Narrative Memory for Accidents and their Post-traumatic Distress. *Applied Cognitive Psychology*, *21*(7), 821–838. https://doi.org/10.1002/acp.1294

Olff, M. (2015). Mobile mental health: A challenging research agenda. *European Journal of Psychotraumatology*, *6*(1), 27882. https://doi.org/10.3402/ejpt.v6.27882

Olff, M., Monson, C., Riggs, D., Lee, C., Ehlers, A., & Forbes, D. (in press). Psychological treatments for adults with PTSD: core and common elements of effective treatment. In D. Forbes, C. Monson, L. Berliner, & J. Bisson (Eds.), *Effective treatments for PTSD*. Third edition.

Open Leercentrum. (n.d.). *Voegwoorden overzicht* [Accessed: 2017-04-24]. http://www.openleercentrum.com/Nederlands/Staatsexamen/STEX%5C%201/stex%5C%20I%5C%20schrijven/voegwoorden.doc

Orimaye, S., Wong, J., & Golden, K. (2014). Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From linguistic signal to clinical reality*, 78–87. https://doi.org/10.3115/v1/W14-3210

Park, H., Rogalski, Y., Rodriguez, A., Zlatar, Z., Benjamin, M., Harnish, S., Bennett, J., Rosenbek, J., Crosson, B., & Reilly, J. (2011). Perceptual cues used by listeners to discriminate fluent from nonfluent narrative discourse. *Aphasiology*, *25*(9), 998–1015. https://doi.org/10.1080/02687038.2011.570770

Pedregosa, G., Fand Varoquaux, Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. T., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. http://www.jmlr.org/

Pennebaker, J. (1993). Putting stress into words: Health, linguistic, and therapeutic implications. *Behaviour Research and Therapy*, *31*(6), 539–548. https://doi.org/10.1016/0005-7967(93)90105-4

Pennebaker, J., Francis, M., & Booth, R. (2001). *Linguistic inquiry and word count: LIWC 2001 [Software]*. Lawrence Erlbaum Associates.

Pennebaker, J., Mehl, M., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*, 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Perkins, J. (2014). *Python 3 text processing with NLTK 3 cookbook*. Packt Publishing Ltd.

Pillemer, D., Desrochers, A., & Ebanks, C. (1998). Autobiographical memory: Theoretical and applied perspectives. In C. Thompson, D. Herrmann, J. Read, G. Payne, & M. Toglia (Eds.), *Remembering the past in the present: Verb tense shifts in autobiographical memory narratives*. Lawrence Erlbaum Associates.

Poerner, N., & Schiel, F. (2016). An automatic chunk segmentation tool for long transcribed speech recordings. *Proceedings of the Phonetics & Phonology Conference*, 6–8.

Porter, M. (2001). Snowball stemmer [Stemming algorithm]. http://snowballstem.org/

Python Software Foundation. (2019). Copyright © 2001-2019 Python Software Foundation; All Rights Reserved. https://www.python.org/

Richards, D., & Lovell, K. (1999). Post-traumatic stress disorders: Concepts and therapy. In W. Yule (Ed.), *Behavioural and cognitive behavioural interventions in the treatment of PTSD* (pp. 239–266). Chichester: Wiley.

Rizzo, A., & Shilling, R. (2017). Clinical virtual reality tools to advance the prevention, assessment, and treatment of PTSD. *European Journal of Psychotraumatology*, *8*(sup5), 1414560. https://doi.org/10.1080/20008198.2017.1414560

Roach, P. (2000). Techniques for the phonetic description of emotional speech. *ITRW on Speech and Emotion*, 53–59.

Römisch, S., Leban, E., Habermas, T., & Döll-Hentschker, S. (2014). Evaluation, immersion, and fragmentation in emotion narratives from traumatized and nontraumatized women. *Psychological Trauma: Theory, Research, Practice, and Policy*, *6*, 465–472. https://doi.org/10.1037/a0035169

Rude, S., Gortner, E., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, *18*(8), 1121–1133. https://doi.org/10.1080/02699930441000030

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1), 227–256. https://doi.org/10.1016/S0167-6393(02)00084-5

Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003). Vocal expression of emotion. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Series in affective science. handbook of affective sciences* (pp. 433–456). Oxford University Press.

Scherer, S., Stratou, G., Lucas, G., Mahmoud, M., Boberg, J., Gratch, J., Rizzo, A., & Morency, L. (2013). Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing*, *32*(10), 648–658. https://doi.org/10.1016/j.imavis.2014.06.001

Schmid, H. (1999). Improvements in part-of-speech tagging with an application to German. In *Natural language processing using very large corpora* (pp. 13–25). Springer.

Schnyder, U., Ehlers, A., Elbert, T., Foa, E. B., Gersons, B. P. R., Resick, P. A., Shapiro, F., & Cloitre, M. (2015). Psychotherapies for PTSD: What do they have in common? *European Journal of Psychotraumatology*, *6*(1), 28186. https://doi.org/10.3402/ejpt.v6.28186

Schoentgen, J. (2006). Vocal cues of disordered voices: An overview. *Acta Acustica United with Acustica*, *92*(5), 667–680.

Schuller, B., Villar, R., Rigoll, G., & Lang, M. (2005). Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition. *Proceedings (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing*, 325–328.

Shapiro, F. (2001). *Eye movement desensitization and reprocessing: Basic principles, protocols and procedures*. Guilford Press.

Shaw, R., Harvey, J., Nelson, K., Gunary, R., Kruk, H., & Steiner, H. (2001). Linguistic analysis to assess medically related posttraumatic stress symptoms. *Psychosomatics*, *42*(1), 35–40. https://doi.org/10.1176/appi.psy.42.1.35

Shen, D., Sun, J. T., Yang, Q., & Chen, Z. (2006). Text classification improved through multigram models. *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, 672–681. https://doi.org/10.1145/1183614.1183710

Shen, P., Changjun, Z., & Chen, X. (2011). Speech emotion recognition using support vector machine. *International Conference on Electronic & Mechanical Engineering and Information Technology, IEEE*, 621–625.

Shriberg, E. (2001). To 'errrr'is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, *31*(1), 153–169. https://doi.org/10.1017/S0025100301001128

Stone, L., & Pennebaker, J. (2002). Trauma in real time: Talking and avoiding online conversations about the death of Princess Diana. *Basic and Applied Social Psychology*, *78712*, 1–36. https://doi.org/10.1207/S15324834BASP2403_1

Tan, C. M., Wang, Y. F., & Lee, C. D. (2002). The use of bigrams to enhance text categorization. *Information Processing and Management*, *38*(4), 529–546. https://doi.org/10.1016/S0306-4573(01)00045-0

Tanaka, H., Sakti, S., Neubig, G., Toda, T., & Nakamura, S. (2014). Linguistic and acoustic features for automatic identification of autism spectrum disorders in children's narrative. *Proceedings of the workshop on Computational Linguistics and Clinical Psychology: From linguistic signal to clinical reality*, 88–96.

Tausczik, Y., & Pennebaker, J. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Truong, K., & Raaijmakers, S. (2008). Automatic recognition of spontaneous emotions in speech using acoustic and lexical features. In A. Popescu-Belis & R. Stiefelhagen (Eds.), *Machine Learning for Multimodal Interaction. MLMI 2008. Lecture Notes in Computer Science, vol 5237* (pp. 161–172). Springer-Verlag.

Uddo, M., Vasterling, J., Brailey, K., & Sutker, P. (1993). Memory and attention in combat-related post-traumatic stress disorder (PTSD). *Journal of Psychopathology and Behavioral Assessment*, *15*(1), 43–52. https://doi.org/10.1007/BF00964322

Van Den Broek, E., Van Der Sluis, F., & Dijkstra, T. (2009). Therapy Progress Indicator (TPI): Combining speech parameters and the subjective unit of distress. *Proceedings of the 3rd international conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009*, 2–7. https://doi.org/10.1109/ACII.2009.5349554

Van Miltenburg, E. (2015). Dutch tagger. Retrieved November 5, 2017, from https://github.com/evanmiltenburg/Dutch-tagger

Van Wijk, C., & Kempen, G. (1980). Funktiewoorden. een inventarisatie voor het Nederlands - An inventory of Dutch function words. *International Journal of Applied Linguistics*, *47*, 53–68. https://doi.org/10.1075/itl.47.05van

Vapnik, V. (1995). *The nature of statistical learning*. Wiley.

Vergyri, D., Knoth, B., Shriberg, E., Mitra, V., Mclaren, M., Ferrer, L., Garcia, P., & Marmar, C. (2015). Speech-based assessment of PTSD in a military population using diverse feature classes, 3729–3733.

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, *48*(9), 1162–1181. https://doi.org/10.1016/j.specom.2006.04.003

Wild, J., Warnock-Parkes, E., Grey, N., Stott, R., Wiedemann, M., Canvin, L., Rankin, H., Shepherd, E., Forkert, A., Clark, D. M., et al. (2016). Internet-delivered cognitive therapy for PTSD: A development pilot series. *European Journal of Psychotraumatology*, *7*(1), 31019. https://doi.org/10.3402/ejpt.v7.31019

Wilson, D., & Wharton, T. (2006). Relevance and prosody. *Journal of Pragmatics*, *38*(10), 1559–1579. https://doi.org/10.1016/j.pragma.2005.04.012

Wolf, M., Sedway, J., Bulik, C., & Kordy, H. (2007). Linguistic analyses of natural written language: Unobtrusive assessment of cognitive style in eating disorders. *International Journal of Eating Disorders*, *40*(8), 711–717. https://doi.org/10.1002/eat.20445

Yang, Y. (1997). An evaluation of statistical approaches to text categorization. *Information Retrieval*, *1*(1-2), 69–90. https : / / doi . org / 10 . 1023 / A : 1009982220290

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *Fourteenth International Conference on Machine Learning (ICML)*, 412–420.

Zijlstra, H., Van Meerveld, T., Van Middendorp, H., Pennebaker, J., & Geenen, R. (2004). De Nederlandse versie van de 'Linguistic and Word Count' (LIWC). Een gecomputeriseerd tekstanalyseprogramma. *Gedrag & Gezondheid*, *32*(4), 271–281.

Zoellner, L., Alvarez-Conrad, J., & Foa, E. (2002). Peritraumatic dissociative experiences, trauma narratives, and trauma pathology. *Journal of Traumatic Stress*, *15*(49), 49–57. https://doi.org/10.1023/A:1014383228149

Appendix 5-A
Extracted text features

N-grams

N-grams were extracted to analyze differences in content of hotspots versus non-hotspots. N-grams are for example bigrams (sequences of two words) or trigrams (sequences of three words), whereas N-multigrams consist of variable-length sequences of maximum N words (D. Shen et al., 2006). Contrary to unigrams, N-grams and N-multigrams can take into account the relationship between consecutive words and word context, which can be valuable when analyzing words with multiple meanings or when the relationship between consecutive words changes the meaning of a phrase, for example in case of negation (Bekkerman & Allan, 2003; D. Shen et al., 2006; Tan et al., 2002). Unigrams, bigrams, trigrams, and N-multigrams of maximum three words were extracted and weighted using the CountVectorizer implemented in Scikit-learn.

Table 5.A.1: N-grams

Feature	Description	Construct
Unigrams	Single words	Content
N-grams	Short phrases of N consecutive words (max N was set to 3)	Content
N-multigrams	Variable-length sequences of max N words (max N was set to 3)	Content

Parts-of-speech (POS) tags

Many different POS tagsets exist, but almost every tagset contains the 12 universal grammatical tags, which are verbs, common and proper nouns, pronouns, adjectives, adverbs, pre- and postpositions, conjunctions, determiners, cardinal numbers, participles, "other", and punctuation (Perkins, 2014). A POS tagger is generally trained on a training corpus that consists of POS tagged words; tokens of the format (word, tag). We used a pre-trained Perceptron tagger for Dutch by Van Miltenburg (2015) based on the NLCOW14 corpus, which was tagged using the stochastic TreeTagger by Schmid (1999), available in Python 3. We extracted 24 POS tags that were expected to relate to the prespecified operational constructs.

LIWC categories

Of the 66 categories included in LIWC, 23 were expected to relate to the prespecified operational constructs. Another 11 were included as features because of their expected relation to hotspot content (e.g., categories related to perceptual processes, assent, and negation). The occurrence frequencies for these categories were determined using the validated Dutch dictionary developed by Zijlstra et al. (2004).

Table 5.A.2: Parts-of-speech (POS) tags

Tag	Example (Dutch)	Construct
Conjcoord (coordinate conjunction)	and (en)	Complexity, Emotions, Cognitions
Conjsubo (subordinate conjunction)	as (als)	Complexity, Emotions, Cognitions
Det_art (determiner article)	the (de, het)	Emotions, Cognitions
Det_demo (demonstrative determiner)	those (die)	Cohesion
Nounpl (common noun plural)	humans (mensen)	Cohesion
Nounprop (proper noun)	Sudan (Soedan)	Cohesion
Nounsg (common noun singular)	human (mens)	Cohesion
Partte (particle)	to (te)	Emotions, Cognitions
Prep (preposition)	at (aan)	Emotions, Cognitions
Pronadv (adverbial pronoun)	with that (er, daarmee)	Cohesion, Emotions, Cognitions
Prondemo (demonstrative pronoun)	self (zelf)	Cohesion, Emotions, Cognitions
Pronindef (indefinite pronoun)	some (sommigen)	Cohesion, Emotions, Cognitions
Pronpers (personal pronoun)	he (hij)	Cohesion, Emotions, Cognitions
Pronposs (possessive pronoun)	his, mine (zijn, mijn)	Cohesion, Emotions, Cognitions
Pronquest (interrogative pronoun)	who, what (wie, wat)	Cohesion, Emotions, Cognitions
Pronrefl (reflexive pronoun)	"X"-self, each other (zich, elkaar)	Cohesion, Emotions, Cognitions
Pronrel (relative pronoun)	what (wat)	Cohesion, Emotions, Cognitions
Verbinf (verb infinitive)	to do (doen)	Emotions, Dissociation, Organization
Verbpapa (verb past participle)	painted (geschilderd)	Dissociation
Verbpastpl (verb past tense plural)	could (konden)	Dissociation
Verbpastsg (verb past singular)	dived (dook)	Dissociation
Verbpresp (verb present participle)	laughing (lachend)	Emotions, Dissociation, Organization
Verbprespl (verb present tense plural)	sit (zitten)	Emotions, Dissociation, Organization
Verbpressg (verb present tense singular)	sit (zit)	Emotions, Dissociation, Organization

Note. POS tag overview as published in the Dutch tagset documentation for the TreeTagger Tool developed by Helmut Schmid, Institute for Computational Linguistics, University of Stuttgart. Retrieved from http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/. Examples adopted from Sketch Engine; https://www.sketchengine.eu/dutch-treetagger/.

NRC emotion lexicon

To captue emotions we used the open source NRC word-emotion association lexicon (also known as EmoLex) developed by Mohammad and Turney (2010, 2013). This is a hand-coded lexicon originally annotated for English and translated for over twenty languages using Google Translate (July, 2015) based on the assumption that affective norms are stable across languages despite possible cultural differences. The Dutch NRC emotion lexicon contains associations for 7,850 words. Despite possible errors the lexicon may contain due to incorrect or transliteral translations, we expected features extracted using the NRC emotion lexicon to complement the LIWC emotion features, because it covers emotion categories not included in LIWC (e.g., disgust, trust, anticipation, and surprise).

Custom tags

We specified several (parts of) words and word patterns that we expected to relate to the prespecified constructs. Counts for all words or phrases that matched these specific patterns were returned by the custom tagger.

Text characteristics and statistics

General text characteristics are for example the total number of words used (text length), the number of unique unigrams or N-grams (lexical diversity), number of complex words (words of six or more characters; Tausczik & Pennebaker, 2010),

Table 5.A.3: LIWC categories

Category	Example (Dutch)	Construct
Total 1st person	I, we (ik, wij)	Dissociation
Total 2nd person	you, your (jij, jouw)	Dissociation
Total 3rd person	their, she (hun, zij)	Dissociation
Negations	no, never (nee, nooit)	Content
Assent	agree, yes (eens, ja)	Content
Affect words (emotional processes)	happy, sad (blij, verdrietig)	Affect, Organization
Positive emotions	thankful, brave (dankbaar, dapper)	Emotions
Positive feelings	fun, love (plezier, liefde)	Emotions
Optimism	proud, willpower (trots, wilskracht)	Emotions
Negative emotions	hurt, hostile (gekwetst, vijandig)	Emotions
Anxiety	nervous, worried (nerveus, bezorgd)	Emotions, Organization
Anger	hate, threat (haat, dreiging)	Emotions
Sadness	crying, grief (huilen, rouw)	Emotions
Cognitive processes	cause, know (oorzaak, weten)	Cognitions, Organization
Causation	because, effect (omdat, effect)	Cognitions
Insight	think, consider (denk, overwegen)	Cognitions
Discrepancy	should, could (zouden, kunnen)	Cognitions
Inhibition	constrain, stop (beperken, stoppen)	Cognitions
Tentative	maybe, perhaps (misschien, wellicht)	Cognitions
Certainty	always, never (altijd, nooit)	Cognitions
Perceptual processes	observing, feel (observeren, voelen)	Dissociation
Time	end, until (eind, totdat)	Dissociation
Verbs in past tense	went, ran (ging, rende)	Dissociation
Verbs in present tense	is, does (is, doet)	Emotions, Dissociation, Organization
Verbs in future tense	will, going to (zal, gaan)	Dissociation
Religion	pray, honour (bidden, eren)	Content
Death	bury, kill (begraven, doden)	Content
Physical	ill, faint (ziek, flauwvallen)	Content
Body	vital, cramp (vitaal, kramp)	Content
Sexual	flirt, kiss (flirten, kussen)	Content
Ingestion	drink, hungry (drinken, honger)	Content
Sleep	nightmare, awake (nachtmerrie, wakker)	Content
Groom	shower, wash (douchen, wassen)	Content
Swear words		Content

Note. LIWC categories and examples translated from Zijlstra et al. (2004).

Table 5.A.4: NRC emotion lexicon

Category	Example (Dutch)	Construct
Anger	Crunch, harassing (knarsen, storend)	Emotions
Disgust	Dank, decompose (vochtig, ontleden)	Emotions
Fear	Crouch, hesitation (hurken, aarzeling)	Emotions, Organization
Happiness/Joy	Pleased, praise (tevreden, lof)	Emotions
Sadness	Homesick, pity (heimwee, jammer)	Emotions
Surprise	Incident, pop (incident, knal)	Emotions
Anticipation	Hurry, importance (haasten, belang)	Emotions
Trust	Personal, stable (persoonlijk, stabile)	Emotions
Positive sentiments	Amiable, learn (beminnelijk, leren)	Emotions
Negative sentiments	Chilly, suffer (kil, lijden)	Emotions

Note. Emotion categories and examples derived from Dutch NRC emotion lexicon file.

number of repeated words and bigrams, revisions, speaker turns and utterances, and statistical measures such as reading ease and grade level indices to examine language strength.

Except for the number of complex words, which was extracted using the LIWC

Table 5.A.5: Custom tags

Tag	Example (Dutch)	Construct
Emotional expressions	sniff, sob, cry, sigh, cough (snif, snik, huil, zucht, kuch)	Affect, Emotions
Additive connectives/ Conjunctions	and, also, in addition, besides, not only…but also, moreover, further (en, ook, daarbij, daarnaast, niet alleen…maar ook, verder, voorts)	Cohesion
Comparative connectives/ conjunctions	**Comparison:** like, as if, except (zoals, alsof, behalve) **Contradiction:** (even) though, although, despite, in spite of, without ((al)hoewel, ofschoon, ondanks dat, zonder dat) **Condition:** if, in case, provided, unless (als, indien, mits, tenzij) **Between sentences:** and, or, but, neither, however, nor (en, of, maar, doch, edoch, noch)	Cohesion
Temporal connectives/ conjunctions	**Time:** when, if, while, once, before, for, now, then, after, afterwards, before (wanneer, als, terwijl, zodra, voordat, voor, nu, toen, nadat, nadien, vooraleer) **Duration:** as long as, until, since, as, according as (zolang als, totdat, sinds, sedert, naarmate, naargelang)	Cohesion
Causal connectives/ conjunctions	Cause/effect/reason/purpose words: because, so, so...that, whereby, for, for that, therefore, that, since, if...then, by, in case (doordat, zodat, zo…dat, waardoor, omdat, opdat, daarom, dat, aangezien, als…dan, door, in geval)	Cohesion
Adverbial adverbs	**Connecting:** moreover, likewise, nor, also, besides, even, therewith (bovendien, eveneens, evenmin, ook, tevens, zelfs, daarbij) **Contradicting:** on the other hand, nevertheless, nonetheless, however, though, on the contrary, meanwhile, yet, now (daarentegen, des(al)niettemin, desondanks, echter, evenwel, integendeel, intussen, nochtans, niettemin, nu, toch) **Consequential:** consequently, therefore, thus, hence, because of (bijgevolg, derhalve, deswege, dus, dientengevolge) **Other:** at least, after all, by the way, besides, yet (althans, immers, overigens, trouwens, toch)	Cohesion
Temporal juncture	"then" (dan)	Cohesion
Definite articles	"the" (de, het)	Cohesion
Confusion	don't know, don't get it, don't understand, don't remember (weet (het) niet, snap(te) (het) niet, begrij(ee)p (het) niet, herinner (me) niet, niet herinneren)	Avoidance, Organization
Speech fillers	uh, hmm, hmm-m, so, like, but, anyway, well (dus, ofzo, enzo, zeg maar, soort van, oke, he, weet je, toch, nou ja)	Fragmentation
Revisions	**Fragments:** -word, word-	Fragmentation
Function words	Function word list for Dutch (Van Wijk & Kempen, 1980)	Fragmentation

Note. Connectives and conjunctions derived from grammar overviews by Dutch language course providers (Open Leercentrum, n.d.).

tool, and Honoré's R and Brunét's index, which were calculated separately as in Fraser et al. (2014), all general and statistical text features were extracted using Python's Textstat package. As readability index we only used the Dutch Flesch-Douma measure G (Douma, 1960), an adaptation of the English Flesch reading ease index (FRE; Flesch, 1948), because the Bormuth Grade Level (Zoellner et al.,

2002) uses a standard list of familiar words in English for which no Dutch translation is available.

Table 5.A.6: Text characteristics and statistics

Characteristic/Statistic	Definition/Function	Construct
N(words)	Total number of words used (word tokens)	Fragmentation
N(unique words)	Number of unique words used (word types)	Avoidance
Type:Token Ratio (TTR)	$N(word\ types)/N(word\ tokens)$	Avoidance, Cohesion
Words used once	Words that occur only once in the text	Avoidance, Complexity
N(characters)	Total per phase	
Mean N(characters) per word	Mean word length in characters	Dissociation, Complexity
N(complex words)	Words of > 6 characters	Complexity
N(syllables)	Total syllables per phase	Complexity
Mean N(syllables per word)	Mean word length in syllables	Complexity
Repetition	Number of immediate word repetitions	Organization, Fragmentation
N(unique bigrams)	Number of unique bigrams used	Organization, Fragmentation
Pronoun:Noun ratio (PNR)	$N(pronouns)/N(nouns)$	Cohesion
Subordinate:coordinate ratio	$N(conjsubo)/N(conjcoord)$	Complexity
Dutch Flesch-Douma G	$207 - \frac{0.93 \times N(word\ tokens)/N(utterances)}{77 \times N(syllables)/N(word\ tokens)}$	Dissociation, Complexity
Honoré's R	$\frac{100 \log(N(word\ tokens))}{1-(N(words\ used\ once)/N(word\ types))}$	Avoidance, Complexity
Brunét's index	$N(word\ tokens)^{N(word\ types)^{-0.165}}$	Avoidance, Complexity

Note. Extracted using Python's TextStat package and LIWC.

Stop words

Table 5.A.7: Stop word list

Dutch stop words	English translation
'de', 'en', 'van', 'ik', 'te', 'dat', 'die', 'in', 'een', 'hij', 'het', 'niet', 'zijn', 'is', 'was', 'op', 'aan', 'met', 'als', 'voor', 'had', 'er', 'maar', 'om', 'hem', 'dan', 'zou', 'of', 'wat', 'mijn', 'men', 'dit', 'zo', 'door', 'over', 'ze', 'zich', 'bij', 'ook', 'tot', 'je', 'mij', 'uit', 'der', 'daar', 'haar', 'naar', 'heb', 'hoe', 'heeft', 'hebben', 'deze', 'u', 'want', 'nog', 'zal', 'me', 'zij', 'nu', 'ge', 'geen', 'omdat', 'iets', 'worden', 'toch', 'al', 'waren', 'veel', 'meer', 'doen', 'toen', 'moet', 'ben', 'zonder', 'kan', 'hun', 'dus', 'alles', 'onder', 'ja', 'eens', 'hier', 'wie', 'werd', 'altijd', 'doch', 'wordt', 'wezen', 'kunnen', 'ons', 'zelf', 'tegen', 'na', 'reeds', 'wil', 'kon', 'niets', 'uw', 'iemand', 'geweest', 'andere'	'the', 'and', 'of', 'I', 'too', 'that', 'this', 'in', 'a'/'an', 'he', 'it', 'not', 'to be', 'is', 'was', 'on', 'at', 'with', 'if', 'for', 'had', 'there', 'but', 'to', 'hem', 'then', 'would', 'or', 'what', 'mine', 'one', 'this', 'so', 'through', 'over', 'they', 'them', 'with', 'too', 'until', 'you', 'me', 'from', 'there', 'her', 'to', 'have', 'how', 'has', 'to have', 'these', 'you', 'because', 'still', 'will', 'me', 'they', 'now', 'no', 'because', 'something', 'to become', 'still', 'already', 'were', 'many', 'more', 'to do', 'then', 'have to', 'am', 'without', 'can', 'their', 'so', 'all', 'under', 'yes', 'once', 'here', 'who', 'was', 'always', 'but', 'will be', 'went', 'could', 'us', 'self', 'against', 'after', 'already', 'want to', 'could', 'nothing', 'your', 'someone', 'has been', 'other'

Note. Adapted from NLTK.

Appendix 5-B
Extracted speech features

Pitch

The pitch is measured by the fundamental frequency (F0). The fundamental frequency is the lowest frequency of the waveform. Sounds with higher frequency are generally perceived as having a higher pitch (Jurafsky & Martin, 2009). For each patient utterance we extracted statistics related to the mean pitch (m_pitch) and the standard deviation of pitch (s_pitch). M_pitch is the mean pitch measured when the patient is speaking. For each patient utterance in a hotspot or non-hotspot segment, the mean pitch is measured and averaged over all patient utterances in that hotspot or non-hotspot. S_pitch is the standard deviation of pitch measured when the patient is speaking. For each patient utterance in a hotspot or non-hotspot segment, the standard deviation of pitch is measured and averaged over all patient utterances in that hotspot or non-hotspot. For both m_pitch and s_pitch, the mean, variance, min, max, and range are calculated over all patient utterances in the segment (hotspot or non-hotspot phase in the session) in order to obtain one value per related statistic per segment.

Table 5.B.1: Pitch

Feature	Parameters	Construct
m_pitch	Mean, SD, min, max, range	Affect, Emotions
s_pitch	Mean, SD, min, max, range	Affect, Emotions

Note. Extracted using Praat version 6.0.4.3.

Loudness

The intensity is correlated with a sound wave's amplitude; the maximum vertical displacement from rest (silence) to the top (crest) or bottom (trough) of the wave, which is expressed in decibels (dB) (Lapp, 2006). In general, sounds with higher amplitudes are perceived as being louder (Jurafsky & Martin, 2009). Equal to the extraction of the pitch features, we extracted $m\_intensity$ and $s\_intensity$ and calculated their mean, variance, min, max, and range over all patient utterances in the hotspot and non-hotspot segments, resulting in one value per statistic per segment.

Duration

Duration covers tempo and pause. Tempo refers to the speaking rate, which is measured as overall duration (e.g., sound length in (mili)seconds or total duration of speaking time (as in Fraser et al., 2014; Lamers et al., 2014), or as units per duration (e.g., words or syllables per second or minute) (Juslin & Scherer, 2005).

We measured speech tempo for the entire audio fragment including pauses

Table 5.B.2: Loudness

Feature	Parameters	Construct
m_intensity	Mean, SD, min, max, range	Affect, Emotions
s_intensity	Mean, SD, min, max, range	Affect, Emotions

Note. Extracted using Praat version 6.0.4.3.

(speech rate) and for the spoken parts only, excluding pauses and hesitations (articulation rate) (Jacewicz et al., 2009). Similarly, pauses can be silent or voiced (Roach, 2000). Voiced pauses were covered by the lexical feature "speech-fillers", silent pauses were captured automatically using Praat's silence detection function, with the minimal silence duration set at 500 ms, as in Lamers et al. (2014). We extracted mean, SD, min, max, and rate for the duration of silences (pauses) and speaking time in Praat. Based on these values, we calculated phonation rate, speech productivity, and variables related to speech tempo.

Table 5.B.3: Duration

Feature	Parameters/Function	Construct
Speech rate (incl pauses)	• Words per minute • Syllables per minute • Praat: mean, SD, min, max, range	Affect, Emotions, Fragmentation, Avoidance
Articulation rate (excl pauses)	• Words per voiced minute • Syllables per voiced minute	Affect, Emotions, Avoidance
Phonation rate	$\frac{N(voiced\ minutes)}{N(total\ minutes)}$	Affect, Emotions, Avoidance
Speech productivity (pause:speech ratio)[a]	$\frac{N(silent\ minutes)}{N(voiced\ minutes)}$	Fragmentation
MLU (mean length utterance)	• MLU_words • MLU_mins	Dissociation, Organization, Fragmentation, Complexity, Avoidance
Silent (pause duration)	Mean, SD, max, n, rate, sum	Avoidance
Sounding (speech duration)	Mean, SD, max, n, rate, sum	Avoidance

Note. Extracted using Praat version 6.0.4.3.
[a] Lamers et al. (2014).

Spectral features

Spectral features are frequency based features that represent the different frequencies (called "spectrum") that together make up the acoustic waveform (Jurafsky & Martin, 2009). These features were extracted at frame level, over frames with a window length of 0.015 s and time steps of 0.005 s. We extracted 12 Mel-frequency cepstral coefficients (MFCCs) and calculated mean and variance over all frames. The MFCCs jointly form a mel-frequency cepstrum, which represents a sound's short-term power spectrum (Iliou & Anagnostopoulos, 2010), see Logan (2000) for more on MFCC features.

Table 5.B.4: Spectral features

Feature	Parameters	Construct
m_MFCC_{1-12}	Mean, SD	Emotions
s_MFCC_{1-12}	Mean, SD	Emotions

Note. Extracted using Praat version 6.0.4.3.

Voice quality features

Perceived voice quality is measured by high-frequency energy (HF), which is the relative proportion of energy in an acoustic signal above versus below a specific frequency, and formant frequencies (Juslin & Scherer, 2005). We used a common cut-off frequency of 500 Hz for the high-frequency energy, extracting mean and variance for HF 500. For the formant frequencies, we extracted the mean and precision of the first formant (F1), as commonly used.

Table 5.B.5: Voice quality features

Feature	Parameters	Construct
HF 500	Mean, SD, min, max	Affect, Emotions
HF 1000	Mean, SD, min, max	Affect, Emotions
Slope 500	Mean, SD, min, max	Affect, Emotions

Note. Extracted using Praat version 6.0.4.3.

Turn statistics

Turn statistics are general, overall speech features for each hotspot and non-hotspot phase, such as the number of speaker turns, turn length, and the number of utterances.

Table 5.B.6: Turn statistics

Feature	Parameters	Construct
N(speaker turns)	Total number of speaker turns	General
Turn length	Mean length of speaker turn (in words and minutes)	Complexity
N(utterances)	Total number of patient utterances, split by silences > 1 sec	Dissociation

Note. Extracted using Praat version 6.0.4.3.

Appendix 5-C
Operationalization of hotspot constructs

Affect
To capture the construct affect, we adopted voice cues commonly used in studies of vocal affect, which are pitch, loudness, voice quality, and duration (Juslin & Scherer, 2005). In addition, we used lexicon-based features (LIWC) to assess the occurrence of affect words and custom tags for the occurrence of audible emotional expressions (e.g., sniffing, sighing).

Emotions
We assessed emotions through the use of emotion words, captured through lexicon-based features related to emotion (LIWC and NRC emotion lexicon), and audible emotional expressions. Although the LIWC and NRC categories do not completely cover the emotions found to relate most to hotspots (e.g., guilt and horror are not included in either of the lexicons, see Appendix 5-A, we still expect the available emotion categories to provide additional information on the emotions present in hotspot moments. Emotions can additionally be represented by other textual features, such as an increased use of the present tense (Hellawell & Brewin, 2004; Pillemer et al., 1998) and particles (Pennebaker et al., 2003), which we respectively measured through lexicon-based features and POS tags related to verb tense and particles (e.g., pronouns, articles, prepositions, conjunctives).

Apart from text features, speech features can also be expected to differ among emotions. For example, fundamental frequency and voice intensity (related to pitch and loudness, respectively) are found to be higher for the emotions anger, fear, and stress, and lower for sadness (Juslin & Scherer, 2005). We adopted prosodic features related to pitch, loudness, and duration, and spectral and voice quality features, as these are used in several studies related to emotion, such as the phonetic description of emotional speech (Roach, 2000), emotion detection (Cowie et al., 2001; P. Shen et al., 2011; Ververidis & Kotropoulos, 2006), and the measurement of emotional distress (Van Den Broek et al., 2009).

Cognitions
We operationalized cognitive themes through lexicon-based features (LIWC) related to cognitive processes. As for emotions, not all cognitive themes as defined by Holmes et al. (2005) are covered by the cognitive categories included in LIWC. Still we expect to gain extra information from the lexicon-based features that are available. For example, the categories "causation" and "insight" might relate to the cognitive theme consequences, and the categories "tentative" and "inhibition" to the theme uncertain threat (see overview of extracted cognitions in Appendix 5-A). Moreover, as for emotional state, the POS tag "particle" can be indicative of one's cognitive style (Pennebaker et al., 2003).

Dissociation

We adopted lexicon-based features (LIWC) and POS tags to capture the change in personal pronouns and verb tense associated with dissociation. Following Zoellner et al. (2002), who studied indications for peritraumatic dissociation in trauma narratives, general text characteristics related to narrative structure (characters per word, words per sentence, total number of sentences, and several reading indices) were also used.

Avoidance

We operationalized avoidance through audio statistics related to duration (tempo and pauses), text statistics related to the extensity of descriptions (verbosity) and lexical diversity (also termed vocabulary richness) such as Type:Token Ratio (TTR), Honoré's R, Brunét's index, as in Fraser et al. (2014), and custom tags that indicate confusion.

Cohesion

We operationalized cohesion through custom tags concerning the use of connectives and conjunctions (as in O'Kearny et al., 2007; O'Kearney et al., 2011), and the temporal juncture "then" to measure the temporal sequence of spoken clauses (based on Shaw et al., 2001). According to Shaw and colleagues, use of this temporal juncture by PTSD patients indicates that the patient is closer to re-experiencing a narrated memory with high emotional involvement. Following Crossley et al. (2016), we also used the pronoun:noun ratio (PNR, calculated based on POS tags for nouns and pronouns), the occurrence of demonstratives (captured using POS tags), and definite articles (captured using a customized tag set) to gauge the amount of information given in the text (referred to as "givenness"). Finally, the general text statistic Type:Token Ratio (TTR), an indicator of word repetition across a text, was adopted to assess overall text cohesion.

Organization

Jelinek et al. (2010) studied (dis)organization by counting the number of words indicative of cognitive processes, words related to affection and anxiety, and words in present tense (captured through lexicon-based text features and POS tags). They also used unfinished thoughts (based on Foa et al., 1995) and the "total disorganization score" introduced by Halligan et al. (2003), which is calculated based on the occurrence of repetitions, disorganized thoughts and organized thoughts.

Repetitions are captured by counting the number of direct word repetitions (Croisile et al., 1996; De Lira et al., 2011) and the number of unique bigrams, which is indicative of repeated bigram patterns (Orimaye et al., 2014). Disorganized thoughts, which consist of utterances implying confusion such as "I don't remember" or "I don't know" (Foa et al., 1995) are captured through custom tags, and structural organization of sentences is measured by the Mean Length of Utterance (MLU, as in Orimaye et al., 2014).

Fragmentation

Previous studies assessed fragmentation by coding repetitions, unfinished thoughts, and speech fillers (Foa et al., 1995; Römisch et al., 2014). Of these, we included repetitions (captured as for the construct Organization) and speech fillers (or filled pauses, Fraser et al., 2014, captured using custom tags for e.g., "uh" or "hmm"), since these could be automatically extracted from the data.

Another commonly used indicator for fragmentation is (dis)fluency, because this is a direct and homogeneous measure (Römisch et al., 2014). Speech fluency was found to be inversely related to PTSD symptoms (e.g., Gil et al., 1990; Uddo et al., 1993). Examples of speech disfluencies are repetitions, repairs, filled pauses, and false starts (Shriberg, 2001). To measure speech fluency we used the speech features speech rate and speech productivity and the text feature audible struggle, which were found by Park et al. (2011) to be the most discriminative features for fluency. To capture audible struggle we used custom tags for revisions (based on Croisile et al., 1996; De Lira et al., 2011; Orimaye et al., 2014). Revisions are moments in which the patient retraces and corrects a preceding error, which is extracted from speech transcripts by counting transcribed fragments. Fragments in this context are words that are broken off in the middle. In speech transcripts, fragments are generally represented using '-', e.g., word- or -word (Jurafsky & Martin, 2009).

Finally we used the total number of words produced (as in Fraser et al., 2014), because fragmented speech may be characterized by the use of short, less meaningful, or fragmented phrases and single words, and the total number of function words. Function words are the words that give meaning to a text (Orimaye et al., 2014). Their occurrence was counted using a standard Dutch function words list (first published by Van Wijk & Kempen, 1980).

Complexity

We operationalized complexity through text characteristics related to reading indices, narrative structure, and syntactic processing complexity. Although the use of readability indices to capture text comprehensibility is not undisputed, many different reading indices exist and are used in scientific studies. Amir et al. (1998) for example used the Flesch Reading Ease Index (FRE; Flesch, 1948) and the Flesch-Kincaid Grade Level (FKGL; Kincaid et al., 1975) to capture narrative articulation (i.e., comprehensibility, complexity), whereas Zoellner et al. (2002) used the Bormuth Readability Index (Bormuth, 1969). To gauge narrative structure and syntactic processing complexity, we used general text characteristics such as mean word and sentence length, number of syllables and complex words, and the number and ratio of coordinated and subordinated conjunctions (captured through POS tags, see De Lira et al., 2011; Fraser et al., 2014). Finally, the total number of utterances and the mean number of words per utterance were also adopted as measures for language strength and verbosity (as in Orimaye et al., 2014).

6

Exploring “Letters from the Future” by visualizing narrative structure

This chapter was published as: Wiegersma, S., Sools, A.M., & Veldkamp, B.P. (2016). Exploring “Letters from the Future” by Visualizing Narrative Structure. *Proceedings of the 7th Workshop on Computational Models of Narrative (CMN 2016), 53*, 5:1-5:18. https://doi.org/10.4230/OASIcs.CMN.2016.5

Abstract

Whereas natural language processing is used to automatically extract textual and structural features from narratives, visualizing these features can help to explore patterns and shifts in text content and structure. This study shows how data visualization can be used to explore differences in narrative styles. Streamgraphs were developed for different types of “Letters from the Future”, an online mental health promotion instrument. The visualizations showed differences between as well as within the different letter types, providing directions for future research in both the visualization of narrative structure and in the field of narrative psychology. The method presented here is not limited to “Letters from the Future”, the current object of study, but can in fact be used to explore any digital or digitalized textual source, like books, speech transcripts, or email conversations.

6.1. Introduction

In this study "Letters from the Future", a narrative-based instrument used by Sools and Mooren (2012) to investigate the human capacity to imagine the future, is studied using a combination of quantitative analysis, natural language processing and text visualization methods. Traditionally, in narrative psychology, qualitative methods for analysing narrative content and structure are predominantly based on hand-coded data. The underlying structure of a narrative is represented for example by defining clusters or counting word frequencies. A widely used approach in narrative studies is componential analysis, which focusses on identifying and examining the structural elements that narratives consist of. The narrative framework of Labov and Waletzky (1967), who originally divided narratives into five structural units (orientation, complication, resolution, evaluation, and coda), is a prime example of the componential approach.

The many features and feature combinations that can potentially be extracted from narratives can quickly result in an overwhelming quantity of data to be processed and interpreted. In addition, as a consequence of the growing popularity of e-mental health interventions, more and more digital narrative data become available for analysis. Processing and interpreting all these data by hand is a tremendous, if not impossible, task. However, the growing availability of digital narrative data also generates new opportunities. There now are sufficient narrative data available to scale up the study of narratives by applying natural language processing (NLP) methods. In NLP, computers are used to process and manipulate natural language (Chowdhury, 2005; Liddy, 2001), "natural language" being any spoken or written language used by humans in everyday life (Bird et al., 2009). The main benefits of NLP over the manual processing of narratives is that it is far less time consuming and less error prone than human coders. Moreover, NLP enables researchers to process and compare large data sets or very detailed textual data.

A recent systematic literature review on text mining applications in psychiatry (Abbe et al., 2015) showed that the use of NLP and text mining methods is still in its infancy in the fields of psychology and psychiatry. NLP applications have also only recently found their ways in the field of humanities. From the humanities perspective, computational narratology (Mani, 2014) can be described as a methodological instrument to develop narratological theories, enabling researchers to extend and test their models on larger text corpora and to specify and apply concepts and models automatically and thus more consistently (Meister & Matthews, 2003). As described by Mani (2014), in computational narratology narratives and narrative structures are explored using computation and information processing methods.

Bradley and Rockwell (1994) state that an efficient approach to explore the underlying mathematical structure of narratives is text visualization. The mathematical structure is generally captured using first-, second- and third-order statistics like word frequencies, clustering, and natural language algorithms (Wise et al., 1995). Visually representing this structure enables researchers to reveal and interpret differences and relationships within and between text documents that would have been difficult, or even impossible, to identify solely from the texts or from tables of numerical data extracted from these texts (Bradley & Rockwell, 1994; Valéry et al.,

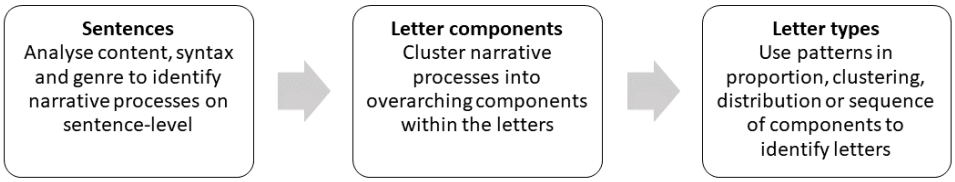


Figure 6.1: Schematic overview of procedure

1999; Wise et al., 1995). Contrary to graphs, visualizations are generally used as an exploratory tool to explore and analyze the data and not to present study results to the public (Valéry et al., 1999).

6.1.1. Letters from the Future

This study uses computational narratology to explore "Letters from the Future", an online narrative-based mental health promotion instrument developed by Sools and Mooren (2012). The instrument is adapted from an earlier exercise by Bohlmeijer (2007), in which storytelling groups are used to enhance mental health. Using a web-based tool, participants are asked to write a letter from a particular situation and moment in the future to someone in the present. Sools et al. (2015) studied the human capacity to imagine the future by hand-coding narrative processes within each individual letter on sentence-level (see procedure in Figure 6.1). They clustered these narrative processes into five overarching components which were then used to identify six different letter types; 1) Imagining and evaluating the futured past; 2) Imagining and orienting to the futured present and futured past; 3) Expressive imagining of the futured present; 4) Reminiscing and evaluating the past without imagination of the future; 5) Intentional orientation with expression of emotions; and 6) Advisory letters about current practical and moral concerns.

The letter types were defined based on a comparative analysis of the following elements: 1) the dominant narrative process (imagining, evaluating, orienting, expressing emotions or engaging in dialogue); 2) the use of certain grammatical elements like past, past imperfect, present and future tense, modals ("would", "could", "should"), intentional time ("hope", "wish", "want"), or the imperative ("go!", "remember!"); 3) the presence and clearness of a path between present and future; and 4) the level of detail of the imagination. Table 6.1 gives an overview of the six letter types and the corresponding structures found by Sools et al. (2015).

As shown in Table 6.1, Sools et al. (2015) found a clear distribution and sequence of narrative processes and grammatical elements for half of the letters (letter types one, two, and four). However, these structures are not always uniformly applicable to all letters of the corresponding type. For example, type one letters generally consist of five elements, but the order of these elements can differ; letters can start either with narrative imagination of the desired future, anticipated reminiscence of the future past or an evaluative part preceding narrative imagination. The same goes for letters of type two, about which Sools et al. (2015) write: *"The orienting function could be prominent from the first sentence, in letters starting with goal-*

setting or value orienting phrases rather than with a situation (but the order could be reversed as well).” [p.19]. Another remark on type two letters was the finding that hope, a prominent feature in those letters, could occur either at the beginning or end of a letter.

The current study is a response to the suggestion of Sools and Mooren (2012) that more in-depth insight into how and why narrative futuring works can be gained by combining traditional qualitative with quantitative methods. The principal aim is to gain a more detailed understanding of the differences in letter content, specifically the distribution (sequential order) and proportion of narrative processes and grammatical elements, both within and between the letter types. This study first addresses the two esential topics in the development of text visualizations: capturing the mathematical structure of the narratives using an existing NLP package, and visualizing these structures in such a way they can be used to study differences both within and between the different types of letters. The developed text visualizations are then compared to previous findings based on qualitative methods by Sools et al. (2015) and linked to the existing narrative framework of Labov and Waletzky (1967). The new insights from this study can be used not only to confirm the previous findings of Sools et al. (2015) but also to develop new theories and hypotheses regarding the human capacity to imagine the future.

6.2. Methods

6.2.1. Data set

An existing data set of 492 “Letters from the future” collected for a previous study by the Storylab, the Dutch expert centre for narrative psychology and mental health promotion at the University of Twente, was used (see Sools & Mooren, 2012; Sools et al., 2015, for more information on the data collection process). Informed consent to reuse these letters for on-going research by the Storylab was obtained. The letters were written by a relatively diverse, mainly Dutch (70%) and German (27%) participants. The letters were manually categorized into six categories by three independent raters (interrater reliability score = 0.672). Table 6.2 shows an overview of the number of letters and mean text length per category. In the current study only Dutch letters that were clearly categorized in one of the six letter types are used, resulting in a data set of 351 letters.

6.2.2. Preprocessing

Salutations, recipient and sender names, location and dates at the beginning and end of the letters are removed. This is done because these elements are considered non-informative and may cause difficulties when splitting and concatenating the letters into segments, distorting the results of subsequent analyses and visualizations. After that the narratives are split into equally sized segments, for which word frequencies can be plotted along the horizontal axis. In a previous study by Clark (2008), document streamgraphs were created for the book “Tom Sawyer” by splitting the text into ten segments. Although using ten segments is suitable for long text documents like books, the narratives used in the current study are much

Table 6.1: Letter structure and characteristics

	Imagining/experiencing a future situation[a]	**Generic letter[b]**
Retrospective evaluation[c]	Type 1, structure: • Narrative imagination of desired future situation (present tense) • Anticipated reminiscence of the future past (past tense) • Conclusion/insight from evaluated experiences and/or • Worldly wisdom (self-praising remarks) • Comments on implications for the future (moral advice/future promises)	Type 4, structure: Equal to structure of type 1. Recounted/evaluated period in past instead of futured past, presented as current concern taking place before moment of writing.
Prospective orientation[d]	Type 2, structure: • Statement about present position in life (present/past perfect tense) • Imaginary goals/purposes • Description of how to realize these objectives	Type 5, no clear structure: No clear action orientation or path from present to future. Sometimes written from future instead of present. Much use of intentional time (hope/wish), future tense (shall/will) and hesitation.
Present-oriented[e]	Type 3, no clear structure: No orientation/evaluation or path from present to future. Sometimes conclusions are drawn. Contains sensory details (hopes, wishes, gratitude and self-appraising). Imagined future described mainly in present-tense.	Type 6, no clear structure: Consists mainly of general insights/conclusions, generic (existential/moral) advices or worldly wisdoms. No path to origination of conclusions or insights.

Note.
[a] Extended core with imaginative component, information on events, places, persons, experience
[b] No/limited imaginative components. Possibly global descriptions of future situations at end of letter
[c] Look back from future or present to past
[d] Look forward from present to future/from future even further ahead
[e] Focus on moment in time (present/future present) instead of period

Table 6.2: Data set characteristics

	Imagination letter	Generic letter
Retrospective evaluation	Type 1: N(letters) = 137 Mean N(words/text) = 324	Type 4: N(letters) = 19 Mean N(words) = 292
Prospective orientation	Type 2: N(letters) = 47 Mean N(words) = 303	Type 5: N(letters) = 9 Mean N(words) = 196
Present-oriented	Type 3: N(letters) = 94 Mean N(words) = 289	Type 6: N(letters) = 45 Mean N(words) = 270

shorter (see Table 6.2 for mean number of words per letter type). Therefore a smaller number of segments may be more appropriate. To decide on the number of segments to use, three different splits were made and the resulting visualizations were compared.

First, following Clark (2008), the narratives were split into ten segments, which resulted in very dynamic and detailed visualizations. However, these results were too fine-grained, making it difficult to use the visualizations for their initial purpose; to confirm previous findings and develop new hypotheses. Second, the narratives were split into three segments (representing the beginning, middle and end of the story, a structure often used in the formation and analysis of narratives (Hogan, 2006)). It was expected that the three segments would result in more interpretable visualizations revealing major trends. The resulting visualizations were however very global and flat, making it difficult to draw conclusions or gain new insights. Therefore third, based on the framework of Labov and Waletzky (1967), widely used to represent narrative information and analyze personal narratives, the narratives were split into five segments: orientation, complication, resolution, evaluation, and coda.

Although five segments may still seem too fine-grained for short narratives like the letters used in the current study, the 'narrative clause' used by Labov and Waletzky (1967) as the basic unit of narrative can be as short as one sentence. This framework therefore is very suitable (and widely used) for analysing short narratives like daily life stories or therapeutic interviews (Labov, 1997). In addition, splitting the narratives into five segments is in line with the five narrative processes used by Sools et al. (2015) to identify the different letter types and letter structures. The five segments resulted into well-interpretable visualizations, showing the same trends as the visualizations for ten segments but then for larger-grained sections more inherent in personal letters.

Since the aim is to develop visualizations per letter type, for each type the letters are split into five equal segments and concatenated in one new text file per

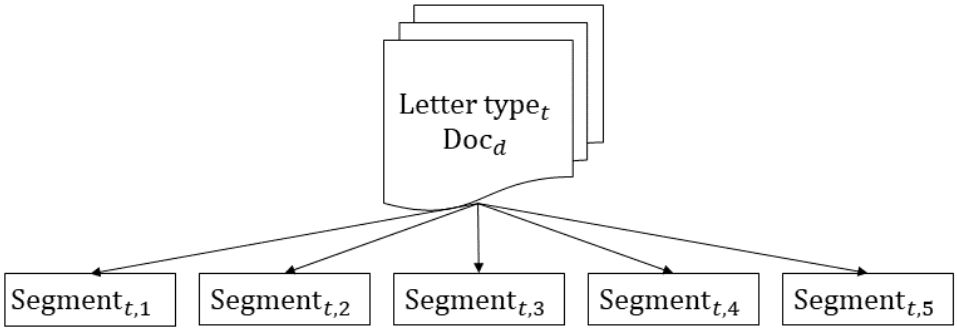


Figure 6.2: Splitting text documents into segments for each letter type

segment. This results in five new text files for each letter type, as shown in Figure 6.2. The five segments are analyzed and visualized for each letter type separately.

6.2.3. Mathematical structure

To explore the differences in letter content and structure, plotting word-frequencies within each text segment for each letter type seems appropriate. However, since plotting frequencies for all used words will probably not lead to legible and interpretable visualizations, generally a sub selection of the occurring words is included in the visualizations. Clark (2008) for example only used words starting with capital letters or only the most prominent words as series in his graph. Another way to reduce the number of series is by categorizing them into word classes, as Weber (2007) did. In the current study words are categorized hierarchically using the text analysis program Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2001). LIWC is a structured, knowledge-rich method, relying on tight structures from existing software and dictionaries. LIWC processes texts on word level, comparing each word to a dictionary files for each category. It is a validated, ready-to use efficient and effective method to study a range of cognitive, emotional and structural components in spoken and written narratives (Pennebaker et al., 2007).

In order to process Dutch texts, the Dutch LIWC dictionary developed by Zijlstra et al. (2004) was used. Contrary to the more complete English dictionary, the Dutch dictionary contains variables for the grammatical tenses past, present and future, but not for modals, intentional time or the imperative. The Dutch dictionary is based on the English LIWC dictionary (2001 version) and, as shown in Table 6.3, consists of 66 word categories divided over five dimensions. The words can be assigned to one or more categories, scoring the occurrences as percentages. The 66 LIWC categories are organized into a hierarchy of eleven main categories and 55 subcategories, which, when applied to the range of letter segments, results in a set of hierarchical additive time series.

Table 6.3: Categories Dutch LIWC dictionary (translated from Zijlstra et al., 2004)

I. Linguistic processes
1. Pronouns (I, them, our)
2. 1st person singular (I, me, mine)
3. 1st person plural (we, our, us)
4. Total 1st person (I, we, me)
5. Total 2nd person (you, your, thou)
6. Total 3rd person (they, their, she)
7. Negations (no, not, never)
8. Assent (agree, ok, yes)
9. Articles (a, an, the)
10. Prepositions (to, with, above)
11. Numbers (second, thousand)

II. Psychological processes
12. Emotional (happy, sad, down)
13. Positive emotions (happy, pleased)
14. Positive feelings (fun, love, smile)
15. Optimism (proud, passionate)
16. Negative emotions (hurt, hostile)
17. Anxiety (nervous, fearful, worried)
18. Anger (hate, annoyed, threat)
19. Sadness (grief, disappointment)
20. Cognitive (cause, know, ought)
21. Causation (because, effect, hence)
22. Insight (think, know, consider)
23. Discrepancy (should, would, could)
24. Inhibition (block, constrain, stop)
25. Tentative (maybe, perhaps, guess)
26. Certainty (always, never)
27. Perceptual (observe, heard, feeling)
28. See (view, saw, seen)
29. Hear (listen, hearing)
30. Feel (feel, touch)
31. Social (share, talk, help)
32. Communication (interview, rumour)
33. Other references (we, them, they)
34. Friends (buddy, friend, neighbour)
35. Family (daughter, husband)
36. Humans (adult, baby, boy)

III. Relativity
37. Time (end, until, season)
38. Verbs in past tense (went, ran)
39. Verbs in present tense (is, does)
40. Verbs in future tense (will, going)
41. Space (nearby, place, North)
42. Up (above, higher, top)
43. Down (deeper, lower, bottom)
44. Including (and, inclusive, too)
45. Excluding (unless, except, out)
46. Motion (approach, walk, climb)

IV. Personal concerns
47. Occupation (achieve, promote)
48. School (student, exam)
49. Work (job, career, colleague)
50. Achievement (earn, hero, win)
51. Leisure (cook, bike, movie)
52. Home (kitchen, home, garden)
53. Sports (game, fitness, work-out)
54. Television (film, video, tv)
55. Music (sing, song, guitar)
56. Money (profit, cash, owe)
57. Metaphysical (altar, church)
58. Religion (pray, honour, bless)
59. Death (bury, mourn, kill)
60. Physical (ill, faint, appetite)
61. Body (vital, thirsty, cramp)
62. Sexual (flirt, love, kiss)
63. Ingestion (drink, hungry, dish)
64. Sleep (dream, wake, sleepy)
65. Groom (shower, make-up)

V. Experimental dimensions
66. Swear words

6.2.4. Analytical procedure

As stated earlier, in this study visualizations are used to explore differences in letter content both within the letters of the same type as between different types of letters. Two separate analyses were used to find the most informative categories. First, to find which categories best visualize the differences in category occurrence *between* the letter types a one-way analysis of variance (ANOVA) is used. The ANOVA is used to determine if there are significant differences between the means of multiple groups (Maxwell et al., 2003). The mean category occurrence is calculated for each letter type by summing up the category scores for all segments and then dividing the sum by five (the total number of segments). The mean occurrences were compared using Welch's statistic (Tomarken & Serlin, 1986).

Second, to find which LIWC categories fluctuate the most *within* each letter type, the spread in category occurrence values for the segments was evaluated. The most commonly used measure of spread in a set of values is the standard deviation (SD). As low SD values indicate that all data points are close to the mean, LIWC categories with low SD values can be presumed to show little to no fluctuation in occurrence within the letter. LIWC categories with high SD values can be presumed to be highly fluctuating and thus showing more differences in occurrence within the concerning letter type. Since there are big differences in the means of the occurrence categories, to be able to compare the variation each SD is normalized with respect to its mean by: SD/mean. The resulting value is known as the coefficient of variation (CV, also known as relative standard deviation), which shows the amount of variability in relation to the mean (Lovie, 2005). A major limitation of the CV is that when the mean is very small, a small variation in the data set will already result in a large CV value (Brown, 2012). Therefore the LIWC categories with mean < 1 were excluded. Then for each letter type the ten most fluctuating LIWC categories (thus the ten categories with the highest CV scores) were selected and included in the letter type specific visualizations.

6.2.5. Text visualization design

Time series analysis, the study of changes in variables over time, can focus on one given variable, or the change of a specific variable compared to others over a certain time period. The time series consists of a sequence of measurements over a continuous, equal distanced time interval (Shumway & Stoffer, 2006). Time series are often visualized using simple line graphs, which works well when comparing a small number of series since the line graph shows direct values for each series at each time point. Another way to visualize time series are stacked graphs, where the series, represented by coloured layers, are stacked on top of each other, showing not only the individual values for each layer but also the total value at certain time points on the horizontal axis (Byron & Wattenberg, 2008). Stacked graphs are very useful to visualize hierarchical time series. However, both line and stacked graphs become illegible when using a large number of series (Byron & Wattenberg, 2008; Clark, 2008). To overcome this problem, Havre et al. (2002) created ThemeRiver, a smooth, continuous graph stacked symmetrically around the x-axis, which is situated at the centre of the graph instead of at the bottom. ThemeRivers later became

known as Streamgraphs, thanks to a popular visualization in the New York Times by Cox and Byron (2008). Streamgraphs differ mainly from ThemeRivers in the design and layout decisions (like colour, interaction or geometry) made to make the graph visually attractive and more organic. Although originally applied to music, movies (Bloch et al., 2008), and (baby) names (Wattenberg, 2005), streamgraphs have also been applied to text documents and seem very suitable for the visualization of narratives.

The smooth, continuous lines that distinguish between the layers are the main advantage of a streamgraph, since this visualizes the data in an intuitive and easily interpretable way (Byron & Wattenberg, 2008; Havre et al., 2002). Continuous data are required to generate such smooth, curving lines. However, splitting the texts into five separate segments results in a discrete data set with different values for y at the data points x_1, x_2, ..., x_5. This problem is solved by interpolating between the discrete data points as suggested by Havre et al. (2002). By using interpolation, intermediate values between data points are estimated from the neighbouring data points (de Carvalho et al., 2007). This results in smooth, continuous lines connecting the discrete data points (McGreggor, 2015). There are different interpolation methods, of which the Cubic Splines model based on third degree polynomials results in the smoothest curve fits (de Carvalho et al., 2007) and is therefore used in the current study.

There are two final notes with regard to the graph design. First, for all visualizations counts that since the layers are stacked symmetrically at the centre of the graph, the values on the y-axis are of no added value and are therefore not included in the plots, as in Havre et al. (2002) and Byron and Wattenberg (2008). Second, sequential colour palettes were used to visualize the hierarchical structure of the time series, as was done by Wattenberg and Kriss (2006). Each main category is assigned its own colour with a range colours with slightly different shades to reflect the corresponding subcategories.

6.3. Results

In this section first the results of the quantitative analysis are described, followed by the resulting visualizations.

6.3.1. Selected LIWC categories

To determine for which LIWC categories there are significant differences in mean occurrence between the different letter types, a one-way ANOVA was used. For 31 of the 66 LIWC categories (indicated by an asterisk (*) in Table 6.4) significant differences between the means ($p \leq 0.05$) were found. The CV was used to measure the amount of variability in category occurrences throughout the letters. Table 6.4 contains the mean proportion (in %) and standard deviation of each LIWC category over all segments. For each letter type, the values for the most fluctuating categories are printed in bold. These categories are selected for the visualizations in Figures 6.4, 6.5 and 6.6.

Table 6.4: Means and standard deviations for each letter type

LIWC category	Type 1 Mean (SD)	Type 2 Mean (SD)	Type 3 Mean (SD)	Type 4 Mean (SD)	Type 5 Mean (SD)	Type 6 Mean (SD)
1. Pronoun*	12.75 (0.66)	11.63 (1.14)	10.56 (0.47)	13.41 (1.11)	13.67 (2.44)	13.19 (0.82)
2. I*	5.07 (0.80)	**3.42 (1.14)**	4.84 (0.62)	6.60 (0.36)	6.95 (1.93)	**3.55 (0.69)**
3. We*	0.46 (0.07)	0.56 (0.15)	0.66 (0.18)	0.43 (0.29)	0.45 (0.73)	0.12 (0.11)
4. Self*	5.54 (0.81)	**3.99 (1.00)**	5.50 (0.71)	7.04 (0.58)	7.40 (1.32)	**3.67 (0.74)**
5. You*	5.50 (0.75)	5.93 (0.76)	**3.11 (0.84)**	4.69 (0.84)	4.01 (1.16)	8.05 (0.30)
6. Other	0.67 (0.19)	0.78 (0.26)	**1.02 (0.29)**	0.65 (0.32)	0.96 (0.51)	0.46 (0.15)
7. Negation	**1.42 (0.41)**	**1.27 (0.35)**	**1.52 (0.38)**	1.51 (0.28)	**1.42 (0.53)**	2.09 (0.31)
8. Assent	0.17 (0.05)	0.14 (0.09)	0.15 (0.06)	0.13 (0.10)	0.00 (0.00)	0.26 (0.20)
9. Article	7.37 (0.76)	7.68 (0.74)	8.21 (0.91)	7.27 (0.94)	7.18 (0.92)	7.16 (0.51)
10. Prepos.*	11.03 (0.72)	11.40 (0.76)	11.13 (0.08)	10.54 (0.95)	10.96 (2.10)	10.24 (0.44)
11. Number	**1.24 (0.60)**	**1.15 (0.63)**	**1.21 (0.58)**	**1.26 (0.64)**	**1.25 (0.82)**	0.83 (0.51)
12. Affect	4.20 (0.66)	3.72 (0.34)	3.90 (0.63)	4.02 (0.94)	4.64 (1.43)	**4.18 (0.76)**
13. Pos. emo.	**2.94 (0.62)**	**2.62 (0.49)**	**2.89 (0.50)**	**2.54 (0.64)**	3.84 (1.18)	**2.67 (0.81)**
14. Pos. feel.	0.77 (0.27)	0.55 (0.17)	0.83 (0.23)	0.52 (0.21)	0.73 (0.47)	0.56 (0.14)
15. Optimism*	0.55 (0.17)	0.56 (0.17)	0.48 (0.17)	0.50 (0.21)	**1.64 (0.54)**	0.51 (0.23)
16. Neg. emo.*	1.15 (0.09)	0.98 (0.16)	0.92 (0.19)	**1.44 (0.34)**	0.73 (0.33)	1.45 (0.14)
17. Anxiety	0.21 (0.07)	0.22 (0.11)	0.11 (0.06)	0.25 (0.12)	0.00 (0.00)	0.27 (0.10)
18. Anger	0.11 (0.03)	0.13 (0.07)	0.10 (0.04)	0.05 (0.08)	0.06 (0.13)	0.21 (0.12)
19. Sadness*	0.28 (0.07)	0.15 (0.04)	0.34 (0.11)	0.45 (0.16)	0.28 (0.20)	0.33 (0.09)
20. Cognitive*	5.72 (0.86)	5.53 (0.28)	5.11 (0.71)	5.99 (1.25)	8.14 (1.39)	7.60 (0.34)
21. Causation	0.57 (0.13)	0.63 (0.07)	0.56 (0.14)	0.52 (0.21)	0.56 (0.34)	0.72 (0.10)
22. Insight	2.10 (0.29)	1.83 (0.18)	1.71 (0.28)	2.35 (0.41)	**2.04 (0.99)**	2.90 (0.24)
23. Discrep.*	**2.33 (0.47)**	2.46 (0.24)	**2.19 (0.37)**	**2.35 (0.65)**	5.26 (0.75)	2.99 (0.24)
24. Inhibition	0.06 (0.03)	0.03 (0.04)	0.06 (0.03)	0.07 (0.10)	0.00 (0.00)	0.08 (0.04)
25. Tentative	**1.50 (0.28)**	1.57 (0.15)	1.49 (0.17)	**1.71 (0.65)**	**2.65 (1.14)**	1.72 (0.18)
26. Certainty	1.58 (0.23)	1.20 (0.14)	**1.32 (0.23)**	1.62 (0.37)	**1.53 (0.62)**	**1.60 (0.46)**
27. Senses	1.27 (0.13)	1.23 (0.16)	1.27 (0.14)	**1.55 (0.39)**	0.68 (0.51)	**1.46 (0.35)**
28. See	0.48 (0.05)	0.39 (0.11)	0.52 (0.17)	0.41 (0.15)	0.28 (0.28)	0.53 (0.12)
29. Hear*	0.44 (0.06)	0.51 (0.15)	0.46 (0.09)	0.76 (0.15)	0.23 (0.37)	0.57 (0.25)
30. Feel	0.34 (0.07)	0.34 (0.09)	0.26 (0.10)	0.38 (0.20)	0.17 (0.25)	0.35 (0.08)
31. Social*	9.61 (1.12)	10.24 (0.65)	8.04 (0.57)	8.99 (0.95)	8.47 (1.52)	11.29 (0.12)
32. Comm.*	0.81 (0.13)	0.80 (0.11)	0.77 (0.12)	1.03 (0.21)	0.23 (0.24)	1.00 (0.11)
33. Others*	6.71 (0.88)	7.33 (0.61)	4.94 (0.42)	5.97 (0.88)	5.48 (1.50)	8.74 (0.19)
34. Friends*	0.24 (0.05)	0.24 (0.05)	0.20 (0.10)	0.14 (0.10)	0.39 (0.16)	0.24 (0.10)
35. Family*	0.89 (0.05)	0.80 (0.19)	0.80 (0.09)	0.88 (0.24)	1.02 (0.33)	0.46 (0.09)
36. Humans*	0.65 (0.16)	0.56 (0.18)	0.86 (0.10)	0.54 (0.25)	0.96 (0.43)	0.69 (0.17)
37. Time	**7.10 (1.23)**	**6.61 (1.97)**	6.82 (0.81)	7.13 (1.16)	**6.04 (2.25)**	**6.73 (1.30)**
38. Past*	**4.46 (0.83)**	**3.87 (0.97)**	**2.91 (0.52)**	5.83 (1.26)	0.96 (0.42)	**2.61 (0.93)**
39. Present*	12.61 (1.23)	12.57 (0.85)	13.44 (0.92)	12.18 (1.35)	14.12 (1.11)	15.05 (0.83)
40. Future*	0.93 (0.19)	**1.21 (0.20)**	0.83 (0.17)	0.90 (0.49)	2.71 (0.82)	1.42 (0.13)
41. Space	1.91 (0.28)	**1.95 (0.44)**	1.93 (0.28)	**1.62 (0.45)**	**2.03 (0.94)**	1.40 (0.20)
42. Up	1.11 (0.16)	0.96 (0.22)	1.21 (0.17)	**1.06 (0.35)**	0.73 (0.51)	0.97 (0.19)
43. Down	0.04 (0.03)	0.02 (0.02)	0.05 (0.05)	0.05 (0.05)	0.06 (0.13)	0.02 (0.02)
44. Incl.*	8.66 (0.21)	9.13 (0.84)	8.52 (0.19)	8.12 (0.54)	10.00 (0.63)	7.71 (0.77)
45. Excl.*	3.92 (0.54)	3.16 (0.24)	**3.64 (0.66)**	**4.15 (1.00)**	**4.01 (1.54)**	4.71 (0.51)
46. Motion	**1.87 (0.33)**	**2.15 (0.39)**	1.91 (0.19)	**2.04 (0.53)**	**1.75 (0.83)**	**1.98 (0.40)**
47. Occup.*	**2.04 (0.38)**	1.84 (0.27)	**1.33 (0.39)**	0.90 (0.22)	0.96 (0.71)	**1.60 (0.28)**
48. School*	0.76 (0.23)	0.72 (0.23)	0.38 (0.10)	0.40 (0.08)	0.40 (0.47)	0.72 (0.11)
49. Job*	**1.01 (0.26)**	0.91 (0.25)	0.75 (0.30)	0.45 (0.17)	0.51 (0.36)	0.49 (0.14)
50. Achieve*	0.32 (0.09)	0.25 (0.13)	0.22 (0.11)	0.11 (0.08)	0.11 (0.15)	0.41 (0.14)
51. Leisure*	0.59 (0.15)	0.91 (0.37)	0.87 (0.21)	0.95 (0.38)	0.73 (0.74)	0.29 (0.12)
52. Home*	0.49 (0.18)	0.73 (0.30)	0.68 (0.20)	0.45 (0.25)	0.45 (0.37)	0.22 (0.13)
53. Sports*	0.05 (0.02)	0.17 (0.07)	0.14 (0.07)	0.27 (0.14)	0.23 (0.37)	0.05 (0.04)

Table 6.4: Means and standard deviations for each letter type (Continued)

LIWC category	Type 1 Mean (SD)	Type 2 Mean (SD)	Type 3 Mean (SD)	Type 4 Mean (SD)	Type 5 Mean (SD)	Type 6 Mean (SD)
54. TV	0.03 (0.03)	0.01 (0.02)	0.01 (0.01)	0.07 (0.08)	0.06 (0.13)	0.00 (0.00)
55. Music	0.02 (0.02)	0.01 (0.03)	0.04 (0.03)	0.22 (0.14)	0.06 (0.13)	0.02 (0.02)
56. Money*	0.25 (0.12)	0.32 (0.12)	0.39 (0.12)	0.07 (0.08)	0.34 (0.12)	0.27 (0.05)
57. Metaphys.	0.07 (0.01)	0.06 (0.06)	0.08 (0.03)	0.13 (0.12)	0.00 (0.00)	0.04 (0.04)
58. Religion	0.04 (0.02)	0.04 (0.04)	0.07 (0.03)	0.07 (0.08)	0.00 (0.00)	0.04 (0.04)
59. Death	0.03 (0.02)	0.03 (0.03)	0.02 (0.01)	0.05 (0.08)	0.00 (0.00)	0.00 (0.00)
60. Physical	0.66 (0.10)	0.54 (0.14)	0.83 (0.09)	0.74 (0.56)	0.73 (0.32)	0.66 (0.14)
61. Body	0.30 (0.03)	0.24 (0.06)	0.33 (0.05)	0.34 (0.39)	0.34 (0.12)	0.40 (0.09)
62. Sexual	0.06 (0.03)	0.06 (0.03)	0.07 (0.03)	0.02 (0.04)	0.06 (0.13)	0.10 (0.08)
63. Eating	0.07 (0.02)	0.07 (0.02)	0.20 (0.11)	0.11 (0.12)	0.17 (0.16)	0.05 (0.03)
64. Sleep	0.24 (0.08)	0.19 (0.12)	0.23 (0.07)	0.32 (0.19)	0.23 (0.13)	0.14 (0.07)
65. Groom	0.00 (0.00)	0.00 (0.00)	0.03 (0.03)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
66. Swear**	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.02 (0.02)

Note. **Bold values:** ten most fluctuating LIWC categories for each letter type.
* Significant differences ($p \leq 0.05$) between means of different letter types
** Only occurred in one letter type so means could not be compared

6.3.2. Visualizations

The figures below contain the streamgraphs for each letter type. The mean proportion of each LIWC category (over all the letters of the concerning letter types) is plotted per segment (s_1, s_2, ..., s_5) on the x-axis. The panel in Figure 6.3 contains six streamgraphs, one for each letter type. These visualizations show differences in occurrence proportions of LIWC categories throughout each letter type. All 66 LIWC categories are included in these graphs. The darkest shades of every colour show the main (overarching) categories, followed by the corresponding sub categories. The categories are plotted in the same order for each graph. These graphs can be used to find central themes within the letters and overall differences *between* the letters. In the legend, the asterisk (*) behind LIWC categories indicates that there are significant differences between the mean occurrences of the letter types for these categories. The visualizations in Figures 6.4, 6.5 and 6.6 show the ten most fluctuating LIWC categories for each letter type. These graphs can be used to find specific patterns and shifts in the occurrence proportions of LIWC categories *within* the letters.

The streamgraphs in Figure 6.3 show some clear similarities and differences between the six letter types. An interesting finding is that the visualizations for types 1-3 do not seem to differ as much as was expected based on the previous findings of Sools et al. (2015). Overall, the imagination letters (type 1-3) seem to have a calmer flow than the general letters (type 4-6), which show bigger differences in the proportions of the LIWC categories over the five segments. The differences between and within the streamgraphs will now be described in more detail and compared pairwise for the retrospective letters (types 1 and 4), prospective letters (types 2 and 5) and present-oriented letters (types 3 and 6).

Retrospective letters

Sools et al. (2015) found that imagination and general retrospective letters generally have the same structure. This is also reflected by the streamgraphs in Figure

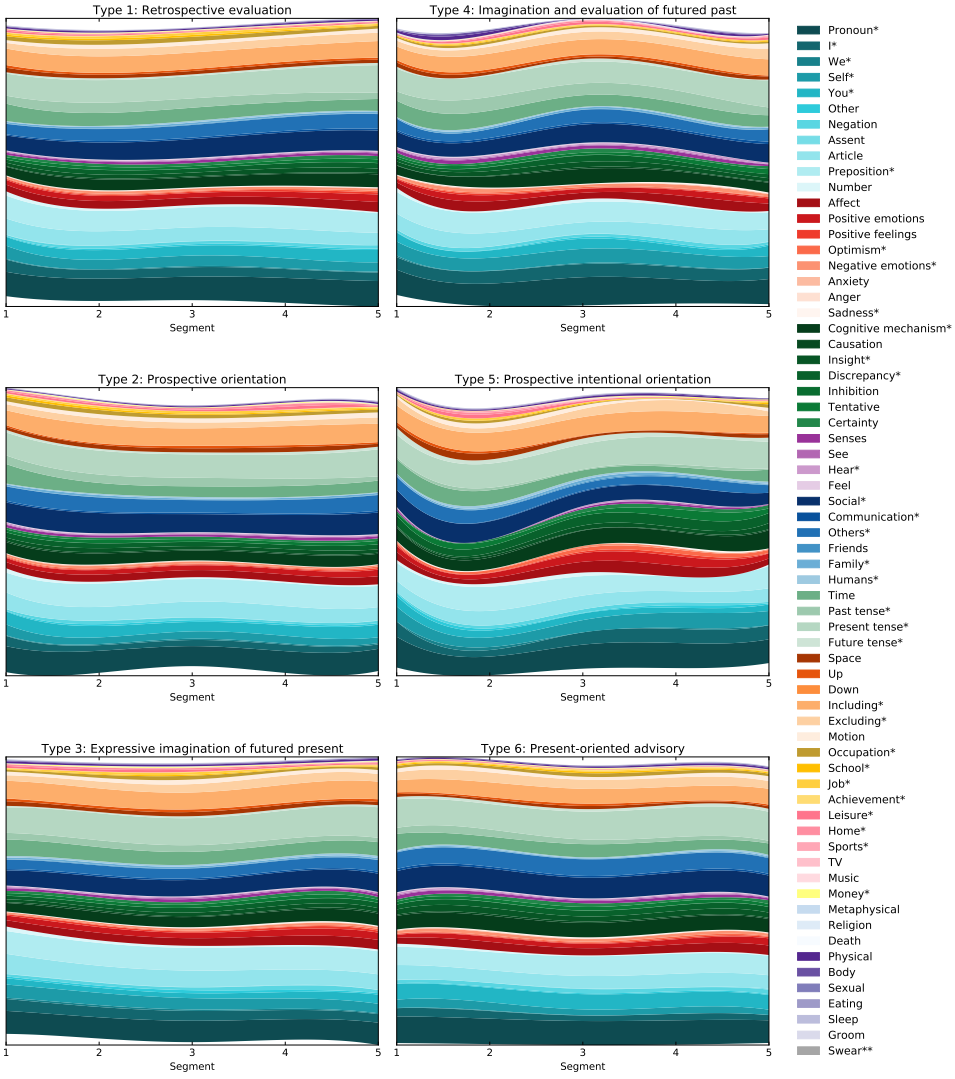


Figure 6.3: Overview LIWC categories per letter type

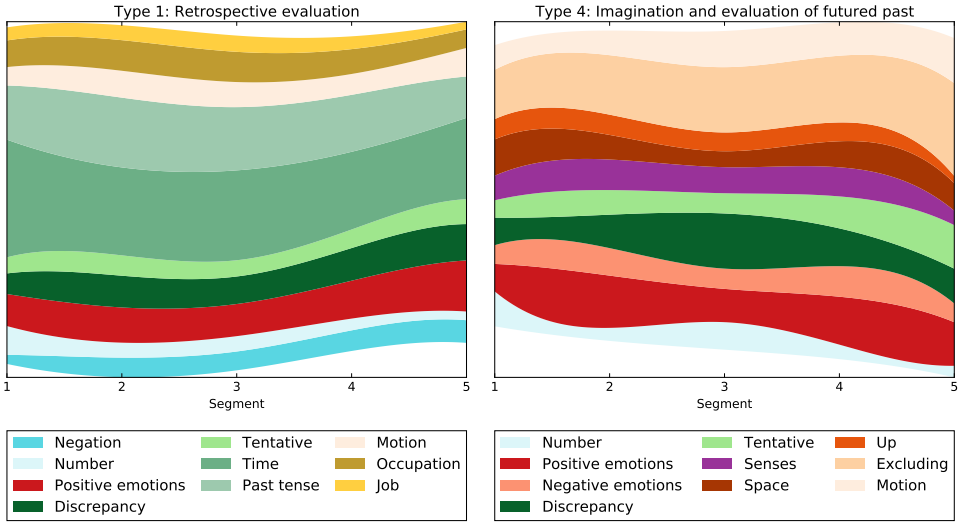


Figure 6.4: Ten most fluctuating categories retrospective letters

6.3, which show that for the majority of the LIWC categories the distribution of the category proportions over the segments is quite similar for both retrospective letters.

According to Sools et al. (2015), the main difference between both retrospective letters would be the verb tenses and the sequence in which these are used. The type 1 letters start with an imaginative future situation in the present tense followed by reminiscence of the future past in the past tense, whereas in the type 4 letters the recounted period actually lies in the past instead of the futured past and is described as a present concern. Based on these findings, one would expect to observe differences in the proportions of used verb tenses both between (Figure 6.3) and within (Figure 6.4) the letters. However, Figure 6.3 shows no observable differences in the proportions of the LIWC categories that regard verb tenses (past, present and future) between letters 1 and 4. Figure 6.4 does show "past tense" as one of the ten most fluctuating categories for letter type 1: the use of past tense slightly increases towards the middle of the letter and then decreases towards the end. The use of present tense does not seem to differ much throughout the type 1 letter as it is not amongst the ten most fluctuating categories included in the graph. None of the used tenses fluctuates much throughout the type 4 letters, as they are not amongst the ten categories included in the graph in Figure 6.4.

Overall it can be observed from both Figure 6.3 and Figure 6.4 that the imagination letters (type 1) contain more words regarding occupation and job, combined with motion words and positive emotions. The words related to occupation and job could be linked to the narrative element "orientation", the first narrative element distinguished by Labov and Waletzky (1967). The motion and positive emotion words could be used to describe the (path towards) the desired future situation

or a period of personal growth ("complicated action"; Labov & Waletzky, 1967). The graphs further show an increasing use of discrepancy words (e.g., should, could, would) from the middle to the end. This supports the findings of Sools et al. (2015), who state that towards the end of the letters conclusions or insights are drawn (pointing towards the narrative elements "evaluation" and "resolution"; Labov & Waletzky, 1967), followed by statements of worldly wisdom self-praising remarks (which could be defined as the "coda"; Labov & Waletzky, 1967).

For type 4 letters, Figure 6.3 and Figure 6.4 show an increase in the use of words from the categories "physical" and "body" combined with both positive and negative emotion words at the beginning and end of the letter. This could indicate that the element "orientation" from the framework of Labov and Waletzky (1967), contains mainly physical characteristics in letter type 4, as opposed to the professional characteristics used in letter type 1. Cognitive mechanisms are used more from the middle (insight and discrepancy) to the end (tentative) of the letters. This could be because these writers are still in the process of reminiscing and evaluating past events (pointing towards the elements "evaluation" and "resolution" of Labov and Waletzky (1967)). It could be that these letters start with a description of physical or emotional complaints or by a recollection of a happier past, which is then processed and evaluated, followed by moral advice or a tentative promise for a better future (distinguished by Sools et al. (2015) as the "coda"). Finally the type 4 letters also contain more words related to senses and leisure. Overall, the general letters seem to be more sensitive, expressive and detailed than the type 1 letters.

Prospective letters

Sools et al. (2015) found a clear structure for the imaginative letters (type 2), but not for the general letters (type 5). The imaginative letters were expected to start with a statement about one's present position in life (in present or past tense). This is reflected in the high occurrence of words in the categories "I" and "Self" and words related to "Time" (e.g., end, until) and numbers at the beginning of the letters (see Figure 6.5), which could be used to describe one's present position in life (narrative element "orientation"; Labov & Waletzky, 1967). The increase in the use of words regarding space (e.g., nearby, places, directions), in the middle can reflect concrete imaginary goals and purposes. The path towards the futured situation (possibly the "complicated action"; Labov & Waletzky, 1967) could be indicated by the increasing use of motion words and positive emotions.

With regard to the used tense, Figure 6.3 further shows that, in addition to the present tense, more past tense is used in the type 2 letters, whereas more future tense is used in the type 5 letters. This is in line with the findings of Sools et al. (2015). Overall the general letters contain more affect and emotions, and more cognitive mechanisms towards the end, which could point towards encouraging oneself to realize their goals, as described by Sools et al. (2015).

The intentional element, the major characteristic of the type 5 letters, is clearly reflected in Figure 6.3 by the use of future tense and the high occurrence of tentative words (like "hope", "believe", "try", "possible") and discrepancy words ("must",

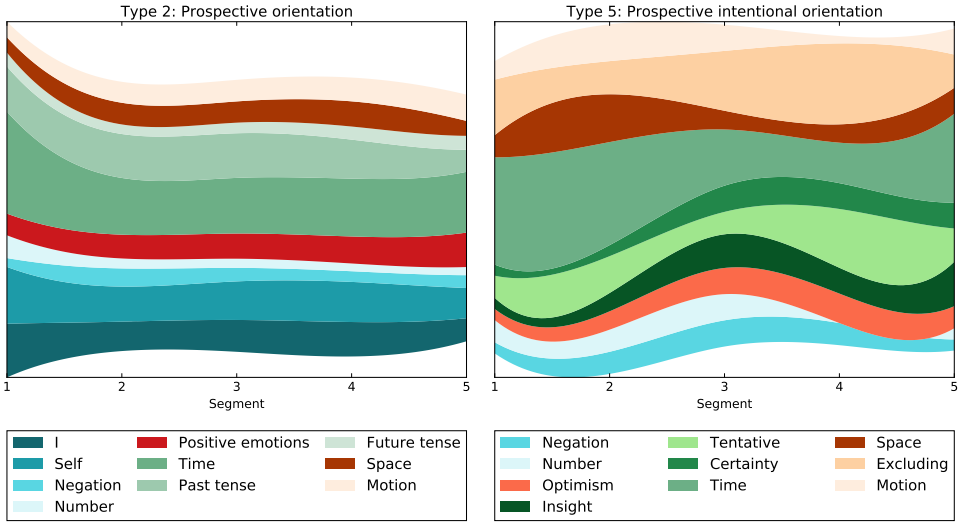


Figure 6.5: Ten most fluctuating categories prospective letters

"wish", "want"). Figure 6.5 further shows that the type 5 letters start and end more tentative, alternated with insight in the middle and end (pointing towards "evaluation"; Labov & Waletzky, 1967). This letter is increasingly optimistic and certain, combined with an increasing use of excluding words. This might point towards an increasing insight in desired versus non-desired situations or future aspects, which may lead to more a more positive and concrete vision for the future in the "coda" (Labov & Waletzky, 1967). However, the increasing use of excluding words combined with the high use of tentative and hesitative words could also reflect the doubt and uncertainty regarding the future related to prospective intentional orientation.

Present-oriented letters

Sools et al. (2015) found no specific sequential order in narrative processes for the present-oriented letters. Figure 6.3 shows that the present-oriented letters are quite similar for both categories. However, the imagination letters (type 3) do contain more words regarding family, leisure, more superlatives (category "up") and slightly more positive emotions and feelings. This is in line with the findings of Sools et al. (2015), who found that type 3 letters are positive, content, and joyful letters.

The letters generally end with hopes and wishes (shown by the increase in discrepancy words) and contain a lot of self-praising remarks (shown by the high increase in the use of "you" in the middle and end). This could point to the narrative elements "resolution" and "coda" (Labov & Waletzky, 1967). The low use of cognitive mechanism and insight words supports the findings of Sools et al. (2015) that the letter contains almost no orientation or evaluation, two of the five narrative elements distinguished by Labov and Waletzky (1967). The high use of excluding words could point towards a breach with the past, without describing the current situation or the path from past to future (no "complicated action"; Labov & Walet-

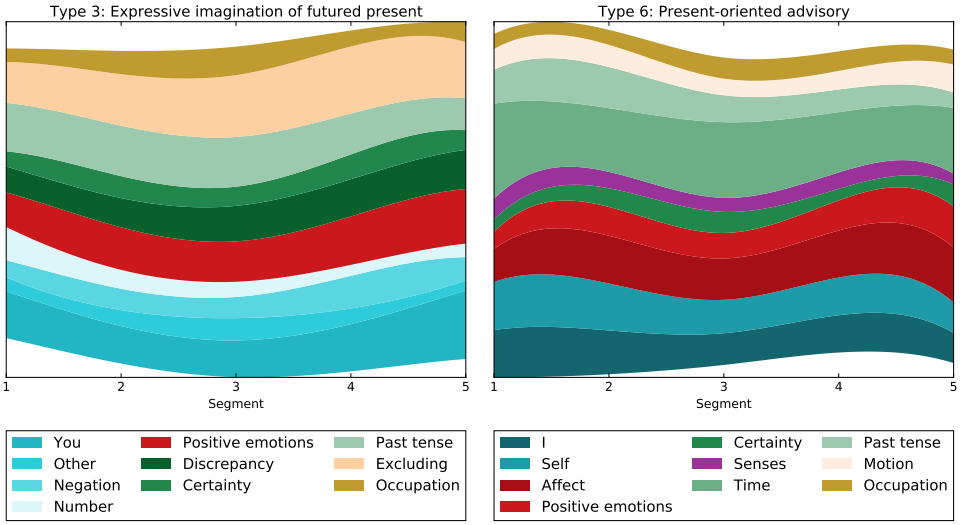


Figure 6.6: Ten most fluctuating categories present-oriented letters

zky, 1967). The additional increase in the use of certainty towards the end indicates that the letters become more stimulating and convincing at the end (indicating “result/resolution” or “coda”; Labov & Waletzky, 1967). It seems that the confidence of the writer increases by imagining the future situation. Finally, regarding the used tense, the type 3 letters are written mainly in the present tense, although Figure 6.6 shows that in both letters the past tense is used more in the beginning than in the middle and end of the letters.

In the general letters (type 6), more insight and discrepancy words are used. These letters also contain more negative emotions and feelings and slightly more sensory words. This supports the findings of Sools et al. (2015), who state that the function of these letters is mainly to provide insight in and guidance for current problems or concerns, followed by statements of worldly wisdom. The finding of Sools et al. (2015) that these letters do not contain a clear path or clarification of how and where certain knowledge or insights have been gained is supported by the fact that these letters contain almost no causation words. The high use of certainty words in the middle of the letter may be explained by the statements of wisdom and moral advice, combined with the fact that these letters do not contain evaluative aspects, which introduce more uncertainty. Apart from the elements “resolution” and “coda” it is difficult to link the letter characteristics from the visualizations to the narrative elements of Labov and Waletzky (1967).

6.4. Discussion

In this chapter, a combination of natural language processing, quantitative analysis and visualization techniques was used to explore differences in letter content, specifically the distribution (sequential order) and proportion of narrative processes

and grammatical elements, both within and between the different types of "Letters from the Future". The visualizations could be used for two purposes; to confirm findings of previous studies on the content of the letters and to explore the letters in a broader sense to come to new insights or theories. Two essential topics in the development of text visualizations – capturing the underlying mathematical narrative structure and choosing a suitable format to visualize changes in letter content throughout the letter – were addressed. In general, the use of text visualizations proved to be a good method to globally explore and compare the underlying structures and differences in contents within and between the letter types. Thanks to the shape of the streamgraphs and the use of sequential colour palettes, the hierarchical time series plots of the letters were easily interpretable and comparable. By combining the visualizations with quantitative analysis of variance and the coefficient of variation, more specific insights in the distribution and proportion of narrative processes and grammatical elements throughout the letters was gained.

All in all, the visualizations were found to be very usable to at least partially confirm the previous findings of Sools et al. (2015). Finding strong additional characteristics or differences between and within the letters turned out to be more challenging. An interesting finding is that the proportional distributions of the LIWC categories, especially those of letter types one, two and three do not differ as much as expected based on the previous findings of Sools et al. (2015). The visualizations for those types look very similar, as opposed to the visualizations for letter types four, five and six. An explanation for this may be that the LIWC categories used as underlying structure are too global or do not directly apply to the current data set. A more specific categorization system developed especially for the "Letters from the Future" data set might perform better. A possibility is to develop a new LIWC dictionary based on the previous findings of Sools et al. (2015) and the visualizations generated in this study, and apply this to a new data set. Potential features to include in this dictionary could be the most informative features that discriminate between the six letter types. These most informative features have been extracted from the current data set for a different study by the authors in which supervised text classification algorithms are used to automatically categorize the letters to their corresponding classes. It would be interesting to visualize the occurrence of these features within the letters.

It could also be that the way the letters are split into five segments influences the proportional distributions. For example, when a certain narrative process starts at the end of the first segment and finishes at the beginning of the second segment, the characteristics for this process are evened out between the first to segments. This may cause a blur in the resulting visualization. It would be interesting to see if splitting the letters manually into five segments, based either on the narrative elements of Labov and Waletzky (1967) or the five narrative processes distinguished by Sools et al. (2015) would lead to more distinctive variations both between the letter segments and the letters as a whole.

Splitting the narratives into the structural elements distinguished by Labov and Waletzky (1967) also opens up to a new avenue for future research, namely to investigate variations in the narratives that depend on the characteristics of the

writer. The framework of Labov and Waletzky (1967) has already been used to investigate differences in narrative content between classes (Horvath, 1987; Labov, 1997), gender (Cheshire, 2000; Johnstone, 1990), age (Peterson & McCabe, 1983), (Toolan, 1988), and geography (Johnstone, 1990). Visualizing the narrative structures for groups with different characteristics may lead to new insights or hypotheses for further research on these topics.

As a final note, although the current focus is on visualizing the content of "Letters from the Future", the resulting method can in fact be used to explore any available digital text document or corpus. The methods and results described in this study can be seen as a first step in an ongoing study by the authors and the Storylab to study therapy-related textual features in e-mental health interventions. By using methods like NLP and text visualization to analyze patterns in therapy-related textual features, extracted for example from written narratives or the linguistic interaction between counsellor and client, more insight can be gained in what happens within therapy, when progress is made, or for which persons a certain type of therapy is more effective. This could greatly improve e-mental health interventions and advance therapy change process research. Future research will therefore include expanding the time series to include more letters written by the same person, studying changes between subsequent narratives and analysing counsellor-client interaction.

6.5. References

Abbe, A., Grouin, C., Zweigenbaum, P., & Falissard, B. (2015). Text mining applications in psychiatry: a systematic literature review. *International Journal of Methods in Psychiatric Research*. https://doi.org/10.1002/mpr.1481

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'reilly Media, Inc.

Bloch, M., Byron, L., Carter, S., & Cox, A. (2008). *The ebb and flow of movies: Box office receipts 1986-2007* [Accessed: 2021-12-18]. http://archive.nytimes.com/www.nytimes.com/interactive/2008/02/23/movies/20080223_REVENUE_GRAPHIC.html

Bohlmeijer, E. (2007). *De Verhalen die we leven. Narratieve psychologie als methode.* Boom.

Bradley, J., & Rockwell, G. (1994). What scientific visualization teaches us about text analysis. *ALLC/ACH Conference*.

Brown, C. E. (2012). Coefficient of Variation. In *Applied multivariate statistics in geohydrology and related sciences* (pp. 155–157). Springer Science & Business Media.

Byron, L., & Wattenberg, M. (2008). Stacked graphs - Geometry & aesthetics. *IEEE Transactions on Visualization and Computer Graphics*, *14*(6), 1245–1252. https://doi.org/10.1109/TVCG.2008.166

Cheshire, J. (2000). The telling or the tale? Narratives and gender in adolescent friendship networks. *Journal of Sociolinguistics*, *4*(2), 234–262.

Chowdhury, G. G. (2005). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89. https://doi.org/10.1002/aris.1440370103

Clark, J. (2008). *Tom Sawyer Character StreamGraph* [Accessed: 2021-12-18]. http://www.neoformix.com/2008/TomSawyer.html

Cox, A., & Byron, L. (2008). *The Ebb and Flow of Movies: Box Office Receipts 1986-2008*. The New York Times.

de Carvalho, O., Guimarães, R. F., Gomes, R. A. T., & Silva, N. C. d. (2007). Time series interpolation. *Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International*, 1959–1961.

Havre, S., Hetzler, B., & Nowell, L. (2002). Themerivertm: In search of trends, patterns, and relationships. *IEEE Transactions on Visualization and Computer Graphics*, *8*(1), 9–20.

Hogan, P. (2006). Continuity and change in narrative study. Observations on componential andfunctional analysis. *Narrative Inquiry*, *16*(1), 66–74.

Horvath, B. (1987). Text on conversation: Variability in storytelling texts. In K. Denning, S. Inkelas, F. McNair-Knox, & R. J. (Eds.), *Variation in language*. Department of Linguistics, Stanford.

Johnstone, B. (1990). Variation in discourse: Midwestern narrative style. *American Speech*, *65*(3), 195–214.

Labov, W. (1997). Some further steps in narrative analysis. *Journal of Narrative and Life History*, *7*, 395–415.

Labov, W., & Waletzky, J. (1967). Narrative analysis: oral versions of personal experience. In J. Helm (Ed.), *Essays on the verbal and visual arts* (pp. 12–44). Washington University Press.

Liddy, E. (2001). Natural language processing. In M. Drake (Ed.), *Encyclopedia of library and information science* (2nd). Marcel Decker, Inc.

Lovie, P. (2005). Coefficient of Variation. In *Encyclopedia of statistics in behavioral science*. John Wiley & Sons, Ltd.

Mani, I. (2014). Computational Narratology. In P. Hühn, C. Meister, Jan, J. Pier, & W. Schmid (Eds.), *Handbook of narratology* (pp. 84–92). De Gruyter.

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2003). *Designing Experiments and Analyzing Data*. Taylor & Francis Group.

McGreggor, D. M. (2015). *Mastering Matplotlib*. Packt Publishing.

Meister, J. C., & Matthews, A. (2003). *Computing Action*. De Gruyter.

Pennebaker, J., Chung, C., Ireland, M., Gonzales, A., & Booth, R. (2007). *The development and psychometric properties of LIWC2007*. LIWC.net.

Pennebaker, J., Francis, M., & Booth, R. (2001). *Linguistic inquiry and word count: LIWC 2001 [Software]*. Lawrence Erlbaum Associates.

Peterson, C., & McCabe, A. (1983). *Developmental psycholinguistics: Three ways of looking at a child's narrative*. Plenum Press.

Shumway, R. H., & Stoffer, D. S. (2006). *Time series analysis and its applications. With R examples*. Springer Science & Business Media. https://doi.org/10.1016/j.peva.2007.06.006

Sools, A. M., & Mooren, J. H. (2012). Towards narrative futuring in psychology: Becoming resilient by imagining the future. *Graduate Journal of Social Science*, *9*(2), 203–226. http://www.gjss.org

Sools, A. M., Tromp, T., & Mooren, J. H. (2015). Mapping letters from the future: Exploring narrative processes of imagining the future. *Journal of Health Psychology*, *20*(3), 350–364. https://doi.org/10.1177/1359105314566607

Tomarken, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, *99*(1), 90–99. https://doi.org/10.1037/0033-2909.99.1.90

Toolan, M. (1988). *Narrative: A critical linguistic introduction*. Routledge.

Valéry, P., Rockwell, G., & Bradley, J. (1999). Empreintes dans le sable: Visualisation scientifique et analyse de texte [Printing in Sand; Scientific Visualization and the Analysis of Texts]. In Vuillemin & LeNoble (Eds.), *Litterature, informatique, lecture* (pp. 130–160). Pulmin.

Wattenberg, M. (2005). Baby names, visualization, and social data analysis. *IEEE Symposium on Information Visualization. INFOVIS 2005.*, 1–7.

Wattenberg, M., & Kriss, J. (2006). Designing for social data analysis. *Visualization and Computer Graphics, IEEE Transactions on*, *12*(4), 549–557.

Weber, W. (2007). Text visualization-what colors tell about a text. *Information Visualization, 2007. IV'07. 11th International Conference*, 354–362.

Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., & Crow, V. (1995). Visualizing the non-visual. Spatial analysis and interaction with

information from text documents. *Proceedings of the IEEE Information Visualization Symposium '95*, 51–58.

Zijlstra, H., Van Meerveld, T., Van Middendorp, H., Pennebaker, J., & Geenen, R. (2004). De Nederlandse versie van de 'Linguistic and Word Count' (LIWC). Een gecomputeriseerd tekstanalyseprogramma. *Gedrag & Gezondheid*, *32*(4), 271–281.

7

General discussion

Due to the enormous amount of routine and research data collected in (mental) health care, data reuse is a rapidly growing area with high potential for clinical practice and research. Data reuse is further encouraged by the increasing availability and usability of new technologies such as artificial intelligence (AI) and machine learning (ML) for researchers and care professionals without a technical background. These technologies can help structure and process the large amount of data becoming available in an efficient and reproducible manner. Especially in mental health care, where a large part of the collected data consists of text or audio recordings, AI applications such as text mining (TM) or audio signal processing (ASR) can have a huge advantage over manual annotation and processing. Despite the benefits, data reuse comes with its challenges, as data are being used for purposes other than originally intended. For structured data challenges can be the data formats and classification systems used to store information, while for unstructured text and audio data problems can arise with the quality of the recordings and patient privacy.

The overall aim of this PhD thesis was to investigate how new technologies such as AI can contribute to the successful reuse of clinical data towards improving (mental) health care practice and research. To this end we performed five studies, demonstrating the practice of clinical data reuse based on a range of different available routinely collected or research data sets from (mental) health care, illustrating the challenges met and solutions used to overcome them. This general discussion summarizes the main findings and ends with a discussion of the limitations and perspectives for the future.

7.1. Main findings

7.1.1. Fitness for purpose of routinely recorded health data

The first study presented in this dissertation (Chapter 2) examined the usability of routinely recorded primary and secondary care data for the identification and validation of patients with complex diseases such as primary Sjögren's syndrome (pSS). pSS is an underdiagnosed, long-term autoimmune disease that affects particularly salivary and lachrymal glands. pSS patients were identified in primary care by translating the formal inclusion and exclusion criteria for pSS that are used in secondary care into a patient selection algorithm using data from Nivel Primary Care Database (PCD), covering 10% of the Dutch population between 2006-2017. The pSS patients found by the algorithm were compared to Diagnosis Related Groups (DRG) recorded in the national hospital insurance claims database (DIS). International Classification of Primary Care (ICPC) coded general practitioner contacts and disease episodes, combined with the mention of "Sjögren" in the disease episode titles, were found to best convert the formal classification criteria to a selection algorithm for pSS. 1,462 possible pSS patients were identified in primary care (mean prevalence 0.7‰, against 0.61‰ reported globally). The DIS contained 208,545 patients with a Sjögren related DRG or ICD-10 (International Classification of Diseases-10) code (prevalence 2017: 2.73‰). 2,577,577 patients from Nivel PCD were linked to the DIS database, among which 1,296 of the 1,462

pSS patients. 716 of the linked pSS patients (55.3%) were confirmed based on the DIS. We found that identifying complex disease patients in primary care largely depends on the availability of structural and granular information. Although prevalence rates in primary care were in line with those reported in literature, rates in secondary care seemed highly inflated. Formal diagnostic information remains required to determine whether routine electronic health record (EHR) data are fit for the identification and study of pSS patients. This study mainly reused structured data that were encoded using several general classification systems for diseases, diagnoses, prescriptions, and diagnostic tests. In addition, a simple keyword search was used to identify pSS cases from the disease episode titles. Although simple text searches are easy to implement, the use of supervised text classification may lead to a more accurate identification of cases from text. The development and use of text classification models to automatically process large volumes of text was described in Chapter 3.

7.1.2. Automated supervised text cassification tool

The second study in this dissertation (Chapter 3) provided a thorough description of supervised text classification, a popular TM application in which textual objects are assigned to a set of predefined class labels using a classification model. Supervised text classification is increasingly used in (psychological) research, as it enables researchers to process, organize, or analyze unstructured text data more efficiently and it improves research consistency and reproducibility. To make this method available for researchers with little to no experience in computer science, statistical modeling, or programming, this chapter provided step-by-step instructions on the development of new binary and multiclass classification models. The study addressed the complete text classification pipeline, including model selection and evaluation using nested cross-validated parameter grid search. The elements of the pipeline (preprocessing, feature extraction, feature selection, and machine learning using support vector machines) were described and the main parameters were reviewed. In addition, an Automated Supervised Text Classification Tool (ASTeCT) was provided, which enables researchers to apply the complete procedure directly to their own text data set to generate their own classification models. The chapter ends with an example in which the tool was applied to a Dutch data set from psychological research practice. ASTeCT was also tested on a public English test data set for classification research, which showed that the procedure and tool can be applied to text data from different (psychological) contexts and in different languages.

7.1.3. Improving treatment intake for mental health disorders

In the third study presented in this dissertation (Chapter 4), the developed tool was applied to textual responses on a patient intake questionnaire to automatically screen for multiple mental health (including substance use) disorders. TM and ML can potentially save a lot of time and effort in the diagnosis and monitoring of patients. Previous studies showed that mental disorders can be detected based on text, but those focused on screening for one predefined disorder instead of multiple

disorders simultaneously. This study developed a Dutch multiclass text classification model to screen for a range of mental disorders, in order to refer new patients to the most suitable treatment. Based on patients' (N = 5,863) textual responses to a questionnaire currently used for intake and referral, a seven-class classifier was developed to distinguish between anxiety, panic, posttraumatic stress, mood, eating, substance use, and somatic symptom disorders. A linear support vector machine (SVM) was fitted using nested cross-validation grid search. The highest classification rate was found for eating disorders (82%). Scores for panic (55%), posttraumatic stress (52%), mood (50%), somatic symptom (50%), anxiety (35%), and substance use disorders (33%) were lower, likely due to overlapping symptoms. The overall classification accuracy (49%) was reasonable for a seven-class classifier. Though this study enabled simultaneous screening for multiple disorders, the performance for disorders other than eating disorders needs to be improved before implementation in mental health practice.

7.1.4. Recognizing hotspots in Brief Eclectic Psychotherapy

In addition to text data, audio data also contain a lot of information that can be extracted automatically using new technologies. The fourth study in this dissertation (Chapter 5) illustrated the development of a multimodal (text and audio) supervised classification model to automatically recognize hotspots from recorded therapy sessions. Identifying and addressing hotspots is a key element of imaginal exposure in Brief Eclectic Psychotherapy for PTSD (BEPP). Research shows that treatment effectiveness is associated with focusing on these hotspots and that hotspot frequency and characteristics may serve as indicators for treatment success. This study aimed to develop a model to automatically recognize hotspots based on text and speech features, which might be an efficient way to track patient progress and predict treatment efficacy. A multimodal supervised classification model was developed based on analog tape recordings and transcripts of imaginal exposure sessions of ten successful and ten non-successful treatment completers. Data mining and machine learning techniques were used to extract and select text (e.g., words and word combinations) and speech (e.g., speech rate, pauses between words) features that distinguished between "hotspot" (N = 37) and "non-hotspot" (N = 45) phases during exposure sessions. The developed model resulted in a high training performance (mean F_1-score of 0.76) but a low testing performance (mean F_1-score = 0.52). This shows that the selected text and speech features could clearly distinguish between hotspots and non-hotspots in the current data set, but would probably not recognize hotspots from new input data very well. In order to improve the recognition of new hotspots, the described methodology should be applied to a larger, higher quality (digitally recorded) data set. As such this study should be seen mainly as a proof of concept, demonstrating the possible application and contribution of automatic text and audio analysis to therapy process research in posttraumatic stress disorder (PTSD) and mental health research in general.

7.1.5. Exploring "Letters from the Future" by visualization

The last study in this dissertation (Chapter 6) showed the use of data visualization to explore differences in narrative styles. As stated before, the growing supply of online mental health tools, platforms, and treatments results in an enormous quantity of digital narrative data to be structured, analyzed and interpreted. Natural language processing (NLP) is very suitable to automatically extract textual and structural features from narratives. Visualizing these features can help to explore patterns and shifts in text content and structure. In this study, streamgraphs were developed for different types of "Letters from the Future", an online mental health promotion instrument. The visualizations showed differences between as well as within the different letter types, providing directions for future research in both the visualization of narrative structure and in the field of narrative psychology. The method presented here was not limited to "Letters from the Future", the object of study, but can in fact be used to explore any digital or digitalized textual source, like books, speech transcripts, or email conversations.

7.2. Interpretation of findings

As stated in the Introduction (Chapter 1), data quality, completeness, and privacy are known critical elements for succesful data reuse. This dissertation illustrates what issues can arise with regard to these elements, how these issues influence the research process and outcomes, and to what extent AI applications can be used to deal with these issues.

Data quality was found to be a major challenge in reusing audio recordings of real-life therapy sessions (Chapter 5). As session recordings contain real, authentic emotions embedded in a broader context, such data are highly valuable for the study of speech sounds and emotion recognition, which are often based on emotions portrayed by actors. To analyze speech content, automatic speech recognition was expected to be a valuable and efficient alternative to manual transcription of each recording, enabling the processing and analysis of large numbers of recordings. However, due to the use of basic recording equipment and the transitory nature of analog recordings, recording quality drastically reduced over the years. Moreover, the heavy emotional outbreaks and the different cultural backgrounds of the patients made any automated processing impossible and as such manual transcription was required. Consequently, our data set was a lot smaller than intended and our results were not generalizable to future data. Luckily, thanks to today's digital recording equipment, audio data quality currently is much higher and data no longer needs to be digitized for further analysis. This eliminates a part of the data quality challenges encountered in our study, making it easier to process and analyze larger amounts of audio data.

The main challenge in reusing text data is patient privacy, especially when dealing with narrative data on mental health problems, which may contain sensitive and personal information (Chapter 4). Our solution for this was to "blindly" process the text and develop a text screening model using the text classification tool developed in Chapter 3. This tool enabled us to work on the sensitive information locally and without any insight in the textual content, which reduced the risk of privacy issues

but also of possible confirmation bias due to prior knowledge. However, by using a tool we were limited by the choice of models and parameters made beforehand, during the development of the tool. Adding or changing the tool's settings based on new insights is quite laborious, as this requires developing, validating, updating, and installing a new version. Therefore, we adopted a common and proven classifier, pipeline, and text features, which may have led to a less successful classification model, at least for disorders other than the eating disorder which was identified very well.

Finally, when reusing routinely collected EHR data for the identification of patient groups (Chapter 2), the lack of sufficiently complete, detailed information or outcome scores made it difficult to apply formal patient inclusion and exclusion criteria and to validate possible patient selection algorithms, as well as the resulting patient set. Even when data was enriched on the patient level with information from a second source, the available EHR data still lacked reliable diagnostic information required to draw formal conclusions. In order to successfully reuse routine EHR data for patient selection, it is important to check beforehand what formal diagnostic information is available, what classification and coding systems are used, and what linkable external data sources exist.

Our findings show that, even when rich or large amounts of data and data processing techniques are available, data reuse does not automatically lead to faster, more generalizable, or better results. In some cases data reuse can become a long journey leading to unsatisfactory results. To prevent or at least be aware of this, a set of guidelines or minimal requirements for successful data reuse might be valuable. The FAIR data principles (Findability, Accessibility, Interoperability, and Reuse of digital objects; European Commission Expert Group on FAIR Data, 2018) offer many useful pointers, especially regarding interoperability (e.g., data should be encoded using community agreed schemas and vocabularies, be processed using open data formats and software, and be easily linkable to other data sets). However, it is difficult to foresee what the effect will be if a data set does not meet all conditions. As this is generally learned during the process, we hope the studies presented in this dissertation have provided more insight in the practice of data reuse.

7.3. Limitations

A first limitation is that this thesis shows how clinical data reuse for scientific purposes works out in practice based on only four data sets. We are aware that no hard conclusions can be drawn from such a limited number of examples. Moreover, all data sets originate from Dutch (mental) health care practice and research, which means our findings may mainly apply to the practice of secondary data use in the Netherlands. Data quality, completeness, and privacy are internationally reported issues when it comes to data reuse (Sherman et al., 2016). However, the solutions described, such as linking data from different sources to validate a developed patient selection algorithm (Chapter 2) may not be possible in every country, as data sources may not be available or may not cover the same populations as in the Netherlands.

A second limitation is that the solutions used in this thesis specifically apply to the data quality, completeness, or privacy related issues present in our cases. Each secondary data set comes with its own challenges, and although AI applications such as NLP, TM, and ASR can be very useful in processing large quantities of data, this dissertation also shows that such techniques may not be applicable to all data sets. For example, when working with low quality audio recordings (Chapter 5), ASR leads to poor results and manual data processing is still required. Similarly, in our studies we limited the use of AI to supervised machine learning. To make described solutions available for other researchers or care professionals, this dissertation comes with an easy, readily available tool for people to develop their own supervised text classification models that can be directly applied in their own research and care setting. However, not all data sets allow for supervised learning, as this requires annotated labels of sufficient quality.

Finally, when working with unstructured patient data, most information is in the details. For a TM algorithm to reach a sensitivity comparable to a trained expert such as a therapist, large training data sets are needed. In practice, when collecting or reusing patient data in the Netherlands, most data sets are too small to develop strong classification models. This limitation was seen not only in the cases presented in this dissertation; also other studies set in the Netherlands (e.g., Smink, 2021) showed that even after years of data collection, the resulting Dutch mental health intervention data set was still too small for successful supervised learning. This underlines a recurring theme throughout the chapters of this dissertation, namely that the available data sets were too small for the complex models that were fitted, and that text preprocessing tools and dictionaries were primarily developed (and sometimes only available) for the English language. Although this may have affected our results, we still think it is important to show what is possible when larger, or higher quality, data sets would be available in the future.

7.4. Future perspectives

Data reuse is a highly relevant topic, which is shown by the emergence of the FAIR data principles and data requirements of scientific journals and funders. However, not all data sets are suitable for sharing and secondary use. This is especially the case for (mental) health data, which are often rich in content but therefore also highly sensitive and possibly personally identifiable. At the moment, most mental health studies, such as those aimed at traumatic stress, have not focused on data preservation, sharing, or reuse (Kassam-Adams & Olff, 2020). However, active efforts are currently made to collect data sets that other researchers may reuse (e.g., Global Collaboration on Traumatic Stress (GC-TS), https://www.global-psychotrauma.net/data-sets). Moreover, initiatives making traumatic stress data more FAIR (e.g., GC-TS, https://www.global-psychotrauma.net/fair-data) and enabling the analysis of highly privacy sensitive research data on location (e.g., Personal Health Trains; Deist et al., 2020) have been launched. Future studies should focus on the practical application and integration of such initiatives in health care research and practice.

In addition, more effort should be put in promoting the benefits and possibili-

ties of data sharing. If more researchers would share their recently or previously collected data, this could speed up research, facilitate efficient collaborations, and ultimately benefit patients (Olff, 2020). Moreover, data sharing can also benefit the researchers who originally collected the data set by drawing more attention and increasing the visibility of their research. However, researchers often see barriers when it comes to sharing their data, e.g., regarding privacy and ownership, or just do not know how to share their data (Kassam-Adams & Olff, 2020). Attention thus should also be paid to educating researchers on different data sharing possibilities, emphasizing the difference between FAIR and open data. FAIR data does not necessarily mean open data for instance; data can also be only available upon reasonable request (Kassam-Adams & Olff, 2020). The GC-TS currently examines traumatic stress researchers' views and experiences regarding data sharing and reuse. All these efforts should help to increase and possibly improve the secondary use of data.

7.5. Conclusion

The central question in this dissertation was how new technologies such as AI can contribute to the successful reuse of clinical data towards improving (mental) health care practice and research. Data reuse is widely encouraged and has the potential to improve health care quality, reduce costs, and lead to more effective clinical research. AI certainly makes it more interesting and worthwhile to reuse existing data sets, as it enables a renewed, more profound analysis of rich and ecologically valid material that may be scarce, difficult to collect, or too extensive for manual processing. However, researchers and health care professionals may encounter several difficulties when reusing existing data sets for purposes other than originally intended.

This thesis demonstrates the practice of clinical data reuse based on four different available data sets, collected during different phases in the care process and from different care settings in the Netherlands. Each study provided insight in possible challenges one can meet when reusing routine and research data, and how AI and other techniques can help deal with these challenges. Despite the growing availability of data, the possibility to link and enrich these data, and the use of AI techniques such as supervised text classification, automated speech processing, and data visualization, reusing data was found to be quite a complicated and lengthy process.

Although AI was regarded a useful tool in (secondary) data processing, for example when analyzing large amounts of text data, extracting speech characteristics from therapy session recordings, and dealing with privacy sensitive data by executing a blind analysis, AI is not the solution to any given problem. Successful data reuse depends more on the data quality (e.g., the quality of an audio recording), label quality (in terms of annotation and classification systems), the size and distribution of the data over different classes (class balance), and the scope of the data set. If those elements are insufficient or do not fit the research question, applying AI cannot be expected to lead to more efficient research or more successful data reuse. It is therefore of great importance to take these elements into consideration

before reusing data.

7.6. References

Deist, T. M., Dankers, F. J., Ojha, P., Scott Marshall, M., Janssen, T., Faivre-Finn, C., Masciocchi, C., Valentini, V., Wang, J., Chen, J., Zhang, Z., Spezi, E., Button, M., Jan Nuyttens, J., Vernhout, R., van Soest, J., Jochems, A., Monshouwer, R., Bussink, J., … Dekker, A. (2020). Distributed learning on 20 000+ lung cancer patients – The Personal Health Train. *Radiotherapy and Oncology*, *144*, 189–200. https://doi.org/10.1016/j.radonc.2019.11.019

European Commission Expert Group on FAIR Data. (2018). *Turning FAIR into reality*. https://doi.org/10.2777/1524

Kassam-Adams, N., & Olff, M. (2020). Embracing data preservation, sharing, and reuse in traumatic stress research. *European Journal of Psychotraumatology*, *11*(1). https://doi.org/10.1080/20008198.2020.1739885

Olff, M. (2020). To share or not to share –10 years of European Journal of Psychotraumatology. *European Journal of Psychotraumatology*, *11*(1). https://doi.org/10.1080/20008198.2020.1844955

Sherman, R. E., Anderson, S. A., Dal Pan, G. J., Gray, G. W., Gross, T., Hunter, N. L., LaVange, L., Marinac-Dabic, D., Marks, P. W., Robb, M. A., Shuren, J., Temple, R., Woodcock, J., Yue, L. Q., & Califf, R. M. (2016). Real-World Evidence — What is it and what can it tell us? *New England Journal of Medicine*, *375*(23), 2293–2297. https://doi.org/10.1056/nejmsb1609216

Smink, W. (2021). *What Works When for Whom? A methodological reflection on Therapeutic Change Process Research*. University of Twente. https://doi.org/10.3990/1.9789036550338

Summary

Given the enormous amount of routine and research data collected in (mental) health care, data reuse is a rapidly growing area with high potential for clinical practice and research. Data reuse is further encouraged by the increasing availability and usability of new technologies such as artificial intelligence (AI) and machine learning (ML) and the implementation of the FAIR data principles. AI and ML technologies can help structure and process the large amount of available data in an efficient and reproducible manner. Especially in mental health care, where a large part of the collected data consists of unstructured text or audio recordings, applications such as text mining or audio signal processing can have a huge advantage over manual annotation and processing. The implementation of the FAIR data principles encourages researchers to make their primary data available for reuse, or to look for reusable existing data sets before collecting new. Despite these developments, data reuse can be challenging, as data sets are being used for purposes other than originally intended. For structured data, the data formats and classification systems used for encoding information can be incompatible, whereas for unstructured text and audio data problems can arise with data quality and patient privacy. The aim of this PhD thesis was to investigate how new technologies such as AI can contribute to the successful reuse of clinical data towards improving (mental) health care practice and research. The five studies presented in the previous chapters contributed to this overarching aim.

The first study examined the usability of routinely recorded primary and secondary care data for the identification of patients with complex diseases. Taking primary Sjögren's syndrome (pSS), an underdiagnosed, long-term autoimmune disease, as an example, a patient selection algorithm was developed and applied to Nivel Primary Care Database (PCD). This study mainly reused structured data that were encoded using several general classification systems for diseases, diagnoses, prescriptions, and diagnostic tests. In addition, a simple keyword search was used to identify pSS cases from recorded disease episode titles. International Classification of Primary Care (ICPC) coded general practitioner (GP) contacts and disease episodes, combined with the mention of "Sjögren" in the disease episode titles, were found to best translate the formal classification criteria to a selection algorithm for pSS. The pSS patients found by the algorithm were compared to Diagnosis Related Groups (DRG) recorded in the national hospital insurance claims database (DIS), by linking routine care data from both sources. Just over half (55.3%) of the pSS patients identified in primary care were confirmed based on data from secondary care. Although prevalence rates in primary care were in line with those reported in literature, rates in secondary care seemed highly inflated. In summary, identifying complex disease patients in primary care was found to largely depend on the availability of structural and granular information. Additional formal diagnostic in-

formation was required to determine whether routine electronic health record data is fit for the identification and study of pSS patients.

Although a keyword search as used in the first study is easy to execute and implement, a more accurate identification of cases from text may be achieved by using supervised text classification. Supervised text classification is a popular text mining application in which textual objects are assigned to a set of predefined class labels using a classification model. This enables one to efficiently process, organize, or analyze large volumes of unstructured text data and improves research consistency and reproducibility. The second study described the development of new binary and multiclass text classification models, addressing the complete text classification pipeline including model selection and evaluation using nested cross-validated parameter grid search. The elements of the pipeline (preprocessing, feature extraction, feature selection, and machine learning using support vector machines) were described and the main parameters were reviewed. In addition, an Automated Supervised Text Classification Tool (ASTeCT) was provided, which enables researchers to apply the complete procedure directly to their own text data set to generate their own classification models. The tool was applied to a Dutch data set originating from an online mental health promotion instrument and tested on a public English test data set for classification research. This showed that the procedure and tool can be applied to text data from different contexts and in different languages.

In the third study, the developed tool (ASTeCT) was applied to textual responses on a patient intake questionnaire to automatically screen for multiple mental health (including substance use) disorders. Previous studies showed that mental disorders can be detected based on text, but those focused on screening for one predefined disorder instead of multiple disorders simultaneously. This study developed a Dutch multiclass text classification model to screen for a range of mental disorders, in order to refer new patients to the most suitable treatment. Based on patients' textual responses to a questionnaire currently used for intake and referral, a seven-class classifier was developed to distinguish between anxiety, panic, posttraumatic stress, mood, eating, substance use, and somatic symptom disorders. A linear support vector machine was fitted using nested cross-validation grid search. The developed model was found to perform particularly well in identify eating disorders. Although this study enabled simultaneous screening for multiple disorders at once, the performance for disorders other than eating disorders needs to be improved before implementation in mental health practice.

In addition to text data, audio data also contain a lot of information that can be extracted automatically using new technologies. The fourth study illustrates the development of a multimodal (text and audio) supervised classification model for the automatic recognition of hotspots (key elements of imaginal exposure during Brief Eclectic Psychotherapy for posttraumatic stress disorder) from recorded therapy sessions. A supervised classification model was developed based on analog tape recordings and transcripts of imaginal exposure sessions of ten successful and ten non-successful treatment completers. Data mining and machine learning techniques were used to extract and select text (e.g., words and word combinations) and speech (e.g., speech rate, pauses between words) features that distinguish

between "hotspot" and "non-hotspot" phases during exposure sessions. The developed model resulted in a high training performance but a low testing performance. This showed that the selected text and speech features could clearly distinguish between hotspots and non-hotspots in the used data set but will probably not recognize hotspots from new input data very well. To improve the automatic recognition of new hotspots, the described methodology should be applied to a larger, higher quality (digitally recorded) data set.

The fifth study showed the use of data visualization to explore differences in narrative styles. Whereas natural language processing (NLP) is suitable to automatically extract textual and structural features from narratives, visualizing these features can help to explore patterns and shifts in text content and structure. In this study, streamgraphs were developed for different types of "Letters from the Future", which were written in the context of an online mental health promotion instrument. The visualizations showed differences between as well as within the different letter types, providing directions for future research in both the visualization of narrative structure and in the field of narrative psychology.

The five studies presented here all made use of secondary data, demonstrating the practice of data reuse based on a range of available routinely collected or research data sets from different health care settings in the Netherlands. It was found that, although data reuse is widely encouraged and has the potential to improve health care quality, reduce costs, and enable more effective clinical research, it can be challenging in practice. Despite the growing availability of data, the possibility to link and enrich these data, and the emergence of new (AI) techniques such as supervised text classification, automated speech processing, and data visualization, reusing data was found to be quite a complicated and lengthy process. AI can certainly further successful data reuse, for example by reproducibly processing large amounts of unstructured data or by "blindly" analyzing privacy sensitive data using a text classification tool such as ASTeCT. However, critical elements such as data quality (e.g., quality of an audio recording), label quality (in terms of annotation and encoding), the sample size and distribution of the data over different classes (class balance), and the scope and granularity of the data set also play an important role. It is of great importance to judge these elements beforehand, because if they are insufficient or do not match with the secondary research purpose, applying AI will most likely not lead to more efficient or successful research.

Samenvatting

De enorme hoeveelheid routine- en onderzoeksdata die in de (geestelijke) gezondheidszorg worden verzameld, maakt data hergebruik een snelgroeiend domein met veel potentie voor zowel de klinische praktijk als het onderzoeksveld. Data hergebruik wordt verder gestimuleerd door de toenemende beschikbaarheid en bruikbaarheid van nieuwe technologieën zoals kunstmatige intelligentie (artificial intelligence; AI) en machine learning (ML) en de implementatie van de FAIR-dataprincipes. Met behulp van AI en ML kunnen grote hoeveelheden data efficiënt en reproduceerbaar worden verwerkt en gestructureerd. Vooral in de geestelijke gezondheidszorg, waar een groot deel van de verzamelde data bestaat uit ongestructureerde tekst- of audio-opnames, kunnen toepassingen als tekstmining of audiosignaalverwerking een enorm voordeel bieden ten opzichte van handmatige annotatie en transcriptie. De implementatie van de FAIR-dataprincipes stimuleert onderzoekers om hun primaire data beschikbaar te stellen voor hergebruik, of om op zoek te gaan naar bestaande, herbruikbare datasets voordat zij zelf nieuwe data verzamelen. Ondanks deze ontwikkelingen kan data hergebruik een uitdaging vormen, omdat datasets worden gebruikt voor andere doeleinden dan oorspronkelijk bedoeld. In het geval van gestructureerde data kunnen bijvoorbeeld de gewenste en de gebruikte gegevensformaten of classificatiesystemen waarmee informatie is vastgelegd onverenigbaar zijn, terwijl voor ongestructureerde tekst- en audiodata de gegevenskwaliteit en de privacy van de patiënt een probleem kunnen vormen. Het doel van dit proefschrift was om te onderzoeken hoe nieuwe technologieën zoals AI kunnen bijdragen aan het succesvol hergebruik van data ter verbetering van zowel de klinische als de onderzoekspraktijk. De vijf onderzoeken die in de voorgaande hoofdstukken zijn gepresenteerd, hebben bijgedragen aan dit overkoepelende doel.

De eerste studie onderzocht de bruikbaarheid van routinematig vastgelegde eerste- en tweedelijnszorg gegevens voor de identificatie van patiënten met complexe aandoeningen. Met als voorbeeld het primaire syndroom van Sjögren (pSS), een ondergediagnosticeerde, langdurige auto-immuunziekte, werd een patiëntenselectie algoritme ontwikkeld en toegepast op de Nivel Zorgregistraties Eerste Lijn (NZR). In dit onderzoek werden voornamelijk gestructureerde gegevens hergebruikt die waren gecodeerd met behulp van verschillende algemene classificatiesystemen voor ziekten, diagnoses, medicatievoorschriften en diagnostische tests. Daarnaast werd een eenvoudige zoekopdracht op trefwoord gebruikt om pSS-gevallen te identificeren op basis van geregistreerde ziekte-episode titels. Internationale Classificatie van Eerstelijnszorg (ICPC) gecodeerde huisartscontacten en ziekte-episodes, gecombineerd met de vermelding van "Sjögren" in de ziekte-episode titels, bleken de formele classificatiecriteria het beste te vertalen naar een selectie-algoritme voor pSS. De pSS-patiënten die door het algoritme werden gevonden, werden vervolgens

vergeleken met Diagnose Behandelcombinaties (DBC) codes die zijn geregistreerd in het DBC-Informatiesysteem (DIS) van de Nederlandse Zorgautoriteit, door routinematige zorggegevens uit beide bronnen te koppelen. Iets meer dan de helft (55,3%) van de pSS-patiënten die in de eerste lijn werden geïdentificeerd, werd bevestigd op basis van gegevens uit de tweede lijn. Hoewel de prevalentiecijfers in de eerste lijn overeenkwamen met de prevalentie gerapporteerd in de literatuur, leek de prevalentie in de tweede lijn sterk verhoogd. Samenvattend bleek het identificeren van patiënten met een complexe ziekte in de eerste lijn grotendeels af te hangen van de beschikbaarheid van structurele en gedetailleerde informatie. Aanvullende formele diagnostische informatie bleek nodig om te bepalen of routinematig vastgelegde gegevens uit het elektronisch patiëntendossier geschikt zijn voor de identificatie en studie van pSS-patiënten.

Hoewel het zoeken op trefwoorden zoals in de eerste studie eenvoudig is uit te voeren en te implementeren, kan de identificatie van patiënten op basis van tekst nauwkeuriger, bijvoorbeeld door middel van gesuperviseerde tekstclassificatie. Dit is een populaire tekstmining toepassing, waarbij tekstuele objecten worden toegewezen aan een set vooraf gedefinieerde categorieën (klasselabels) met behulp van een classificatiemodel. Dit maakt het mogelijk om grote hoeveelheden ongestructureerde tekstgegevens efficiënt te verwerken, organiseren, of analyseren, en verbetert bovendien de consistentie en reproduceerbaarheid van het onderzoek. De tweede studie beschreef de ontwikkeling van nieuwe binaire en multiklasse tekstclassificatiemodellen en ging in op de volledige tekstclassificatie pijplijn, inclusief veelgebruikte strategieën voor modelvalidatie, -selectie, en -evaluatie. De elementen van de pijplijn (voorbewerking, feature extractie, feature selectie, en machine learning met behulp van support vector machines) werden beschreven en de belangrijkste parameters werden uitgelegd. Daarnaast is een Automated Supervised Text Classification Tool (ASTeCT) geleverd, waarmee onderzoekers de volledige procedure rechtstreeks op hun eigen tekst dataset kunnen toepassen om hun eigen classificatiemodellen te genereren. De tool werd toegepast op een Nederlandse dataset afkomstig van een online instrument voor geestelijke gezondheidsbevordering en werd getest op een openbare Engelse test dataset voor classificatieonderzoek. Hieruit bleek dat de procedure en de tool toepasbaar zijn op tekstgegevens uit verschillende contexten en in verschillende talen.

In de derde studie werd de ontwikkelde tool (ASTeCT) toegepast op een vragenlijst voor de intake van patiënten, met als doel automatisch te screenen op meerdere psychische stoornissen. Eerdere studies hebben aangetoond dat psychische stoornissen kunnen worden opgespoord op basis van tekst, maar die waren gericht op het screenen op één vooraf gedefinieerde stoornis in plaats van meerdere stoornissen tegelijk. Onze studie ontwikkelde een Nederlands multiklasse tekstclassificatiemodel om te screenen op een reeks psychische stoornissen, om nieuwe patiënten door te verwijzen naar de voor hen meest geschikte behandeling. Op basis van de tekstuele antwoorden van patiënten op een vragenlijst die momenteel wordt gebruikt voor intake en doorverwijzing, werd een classificatiemodel met zeven klassen ontwikkeld om onderscheid te maken tussen angst-, paniek-, posttraumatische stress-, stemmings-, eet-, middelengebruik- en somatische symptoom-stoornissen.

Het ontwikkelde model bleek bijzonder goed te presteren bij het identificeren van eetstoornissen. Hoewel deze studie gelijktijdige screening op meerdere stoornissen mogelijk maakte, dient de identificatie van andere stoornissen dan eetstoornissen te worden verbeterd voordat een dergelijk model in de geestelijke gezondheidszorg kan worden geïmplementeerd.

Naast tekstgegevens bevatten ook audiogegevens veel informatie die met nieuwe technologieën automatisch kan worden geëxtraheerd. De vierde studie illustreerde de ontwikkeling van een multimodaal (tekst en audio) gesuperviseerd classificatiemodel voor het automatisch herkennen van hotspots (hoofdelementen in imaginaire exposure tijdens Beknopte Eclectische Psychotherapie voor Posttraumatische stressstoornis; BEPP) uit opnames van therapiesessies. Een gesuperviseerd classificatiemodel werd ontwikkeld op basis van analoge bandopnames en transcripties van imaginaire exposure-sessies van tien succesvol en tien niet-succesvol voltooide behandelingen. Data mining en machine learning technieken werden gebruikt om tekst features (zoals woorden en woordcombinaties) en spraak features (zoals spraaksnelheid en pauzes tussen woorden) te extraheren en te selecteren die onderscheid maken tussen "hotspot" en "niet-hotspot" fases in exposure sessies. Het ontwikkelde model resulteerde in een hoge training score maar een lage test score. Dit betekent dat de geselecteerde tekst en spraak features duidelijk onderscheid konden maken tussen hotspots en niet-hotspots in de gebruikte dataset, maar dat hotspots waarschijnlijk niet goed zullen worden herkend in nieuwe datasets. Om de automatische identificatie van nieuwe hotspots te verbeteren, moet de beschreven methode worden toegepast op een grotere dataset van hogere kwaliteit (digitaal opgenomen).

De vijfde studie toonde het gebruik van datavisualisatie om verschillen in narratieve stijl te onderzoeken. Waar natural language processing (NLP) zeer geschikt is om automatisch tekstuele en structurele kenmerken uit teksten te halen, kan het visualiseren van deze kenmerken helpen om patronen en verschuivingen in tekstinhoud en structuur te ontdekken. In dit onderzoek werden stroomdiagrammen ontwikkeld voor verschillende typen "Brieven vanuit de Toekomst", die zijn geschreven in de context van een online instrument voor geestelijke gezondheidsbevordering. De visualisaties lieten zowel tussen als binnen de verschillende typen brieven variatie zien. Dit kan richting geven aan toekomstig onderzoek in zowel de visualisatie van narratieve structuur als op het gebied van narratieve psychologie.

De vijf onderzoeken in dit proefschrift toonden verschillende kanten van datahergebruik in de praktijk op basis van beschikbare routinematig verzamelde gegevens en onderzoeksdatasets uit verschillende zorgcontexten in Nederland. Er werd bevonden dat, hoewel hergebruik van data op grote schaal wordt aangemoedigd en het in potentie kan bijdragen aan effectiever klinisch onderzoek en het verbeteren van de kwaliteit van de gezondheidszorg, er zich in de praktijk ook vele uitdagingen voordoen. Ondanks de groeiende beschikbaarheid van data, de mogelijkheid om deze data te koppelen en te verrijken, en de opkomst van nieuwe (AI) technieken zoals gesuperviseerde tekstclassificatie, geautomatiseerde spraakverwerking en datavisualisatie, bleek het hergebruik van data een behoorlijk gecompliceerd en langdurig proces. AI-toepassingen kunnen het succesvol hergebruik van gege-

vens zeker bevorderen, bijvoorbeeld bij het reproduceerbaar verwerken van grote hoeveelheden ongestructureerde gegevens of door het 'blind' analyseren van privacygevoelige gegevens met behulp van een tekstclassificatie tool zoals ASTeCT. Kritische elementen zoals datakwaliteit (bijvoorbeeld de kwaliteit van een audio-opname), labelkwaliteit (in termen van annotatie en codering), de steekproefomvang en verdeling van de gegevens over verschillende klassen (klassenbalans), en de reikwijdte en fijnmazigheid van de dataset spelen echter ook een belangrijke rol. Het is van groot belang deze elementen voorafgaand aan het hergebruik te beoordelen, want als deze onvoldoende zijn of niet passen bij het secundaire onderzoeksdoel, zal het toepassen van AI hoogstwaarschijnlijk niet leiden tot efficiënter of succesvoller hergebruik van data voor onderzoek.

Dankwoord

Eindelijk is het zo ver en mag ik, als allerlaatste onderdeel van dit proefschrift, mijn dankwoord schrijven. Over de jaren heen zijn er heel wat partijen en mensen betrokken geweest bij dit proefschrift en het leven daarbuiten. Zonder deze mensen had dit proefschrift er nu wellicht niet gelegen. Ik wil een aantal van hen hier graag persoonlijk voor bedanken.

Allereerst mijn promotoren, **Bernard Veldkamp** en **Miranda Olff**. Bernard, bedankt voor de mogelijkheid om te kunnen promoveren bij OMD. Toen ik terugkwam van mijn reis was het best lastig om weer aan het werk te gaan en ik ben blij dat ik toen deze kans heb gekregen. Ik heb veel geleerd, ook over mezelf, en wil je bedanken voor het vertrouwen en de vrijheid die ik heb gekregen om dit proefschrift af te ronden op mijn eigen manier en tempo. Miranda, bedankt voor alle positieve aanmoediging en je kritische blik, waardoor dit proefschrift naar een hoger niveau is getild. Daarnaast wil ik je bedanken voor de manier waarop je mij op het AMC hebt verwelkomd in je team van onderzoekers. Ik heb een hele fijne tijd gehad bij jullie en vind het erg leuk dat ik een aantal jaar heb mogen deelnemen aan alle team-uitjes en andere activiteiten.

Daarnaast gaat mijn dank uit naar **Cees Glas** en **Sjoerd van Tongeren** (†), die samen mijn gecombineerde aanstelling als onderzoeker bij OMD en data manager bij het IGS Datalab mogelijk hebben gemaakt.

Ook wil ik graag de voorzitter en leden van de promotiecommissie bedanken voor hun kennis en tijd.

Dit proefschrift had als uitgangspunt het hergebruiken van bestaande data afkomstig uit eerder onderzoek of de zorgpraktijk. Ik wil dan ook verschillende partijen bedanken die mij de mogelijkheid hebben gegeven hun data te gebruiken.

Allereerst **Anneke Sools** van het Storylab. Jouw "Brieven vanuit de toekomst" dataset was de eerste set waarmee ik aan de slag ging en deze vormde een ware inspiratiebron. Aan de hand van deze set heb ik mijn text mining skills ontwikkeld en van jou heb ik geleerd hoeveel informatie je uit tekst kunt halen. Daarnaast bleken de brieven een waardevolle test set bij het ontwikkelen van de ASTeCT tool. Tot slot hebben de Text Analysis café's die jij organiseerde mij verder wegwijs gemaakt in de wereld van de tekst analyse. Hartelijk dank daarvoor!

Van Interapy wil ik graag **Bart Schrieken** en vooral ook **Maurice Hidajat** bedanken. Bart, bedankt voor het beschikbaar stellen van de rijke Interapy data. Deze dataset vormde niet alleen de basis voor een mooi paper, het heeft me ook uitgedaagd om een tool te ontwikkelen waarmee deze dataset blind geanalyseerd kon worden. Maurice, ontzettend bedankt voor al je tijd, geduld en enthousiasme. Ik heb niet bijgehouden hoe vaak je wel niet een nieuwe versie, een nieuwe run, of een nieuwe check voor mij hebt uitgevoerd. Hoewel ik mij soms bezwaard voelde

dit van je te vragen was jij altijd enthousiast en geduldig. Ik heb ons contact hierin erg gewaardeerd.

Wat betreft de hotspots data wil ik in het bijzonder **Mirjam Mink-Nijdam** en **Arjan van Hessen** bedanken. Mirjam, wat vond ik het interessant om in jouw hotspots data te duiken! Je hebt me wegwijs gemaakt in de dataset en in het herkennen van hotspots. Daarnaast heb je veel met me meegekeken naar de codering. Hiermee hebben we de hotspots data weer een stukje verrijkt en zijn we tot een heel mooi paper gekomen. Dankzij jou rustige begeleiding vond ik dit een fijn project om aan te werken. Arjan, wat een geluk dat ik op jouw kennis en enthousiasme omtrent spraakanalyse heb mogen meeliften! Je hebt mij ontzettend geholpen met de audio analyses en alle voorbereiding die daarbij kwam kijken, waar ik als leek totaal geen weet van had. Zonder jou was het denk ik niet gelukt, en daarnaast was je enthousiasme aanstekelijk. Ook wil ik hier **Khiet Truong** (HMI) en **Laurens Satink** en **Michel Boedeltje** van Telecats bedanken, die mij hebben geholpen met de audio analyses en het oplijnen van de tekst en spraak data.

Tot slot ben ik blij dat ik in mijn tijd bij het Nivel gebruik heb kunnen maken van data uit de Nivel Zorgregistraties Eerste Lijn, die perfect binnen mijn proefschrift pastte. Ik wil vooral **Rodrigo Davids** bedanken voor het meedenken over de data specificaties en de aanlevering van de dataset.

Daarnaast heb ik al die jaren waarin ik aan mijn proefschrift heb gewerkt heel wat werkomgevingen en collega's mee mogen maken. Allereerst natuurlijk mijn collega's van OMD, waar het allemaal begon. Bedankt voor de gezellige teamuitjes, de lunchwandelingen en de kennisuitwisseling tijdens de colloquia. **Qiwei He**, your text mining research was an inspiration and a great starting point. In het bijzonder wil ik mijn kamergenote **Maaike Heitink** bedanken, met wie ik zoveel heb gedeeld, gekletst en gelachen in mijn UT-tijd en daarna. Ik heb veel van jou geleerd en ik ben blij dat jij na al die jaren mijn paranimf wilt zijn! Ik ben je ook super dankbaar voor de mooie omslag voor mijn proefschrift, heel fijn dat je mij hiermee hebt willen helpen.

Ik wil mijn AMC-collega's bedanken dat jullie me zo fijn hebben opgenomen in jullie team. De gezellige lunches, teamuitjes en het ESTSS congres in Denemarken heb ik heel erg gewaardeerd en ik voelde me echt onderdeel van het team.

Mijn collega's bij het Rijnstate, ook al was ik er maar kort, jullie zijn nog steeds mijn favoriet! Mijn eerste ervaring in een gezamenlijke kantoortuin waarin zoveel werd samengewerkt en gelachen. Ik kijk met ontzettend veel plezier op mijn tijd bij jullie terug, en naast het werk had ik nog genoeg energie over om aan mijn proefschrift te besteden. Ook in moeilijke tijden was het prettig om weer naar kantoor te komen. **Elise van Zandbrink**, ik ben jou voor altijd dankbaar voor de tijd en ruimte die je mij hebt gegund rondom het overlijden van mijn vader, en dat je mij helemaal naar Friesland hebt gebracht zodat ik geen belangrijke momenten heb hoeven missen.

Mijn Nivel collega's **Anouk**, **Lotte** en **Isabelle**; bedankt voor de gezelligheid en jullie interesse in mijn proefschrift. De laatste loodjes wogen toch erg zwaar! **Christine**, bedankt voor de herkenning en wat ben ik blij dat ik via jou mijn nieuwe

baan bij IKNL heb gevonden. Gezellig om je weer regelmatig tegen te komen op de werkvloer. Daarnaast wil ik mijn FluCov collega's bedanken, die mij het laatste jaar bij Nivel hernieuwde energie hebben gegeven om door te pakken met mijn proefschrift. En in het bijzonder **John**, thanks for all your funny anecdotes and your support; I've finally let go of the boat!

Tot slot ben ik blij met mijn huidige IKNL collega's en de organisatie skills die ik daar nu opdoe. Ik vind het leuk dit laatste stukje van mijn promotietraject met jullie te kunnen delen.

Naast het werk heb ik de afgelopen jaren veel energie kunnen opdoen bij volleybalvereniging Boni. Bedankt aan al mijn teamgenootjes en aan de EBLC in het bijzonder. Wat hebben we veel lol gehad in de voorbereiding van het lustrum! **Monica**, **Rianne** en **Marijke**, ik vind het fijn dat we ook naast het volleyballen contact houden en de grote life events met elkaar kunnen delen.

Marloes, wat hebben we een toptijd gehad op de beachvelden en daarnaast! Ik koester veel mooie herinneringen aan onze sportieve activiteiten en uitgaansavonturen. Nu we braaf en burgerlijk zijn vind ik het heerlijk om (alleen of met Joost en Eva) bij jou en Steven langs te gaan, waar het ons nooit aan iets ontbreekt. Ik kijk ernaar uit om met jullie en de lieve Dorus samen nog vele mooie momenten te mogen delen.

Mijn oudste vriendinnen uit de Enschede-tijd, **Karin** en **Bregje**; wat fijn dat we nog steeds bij elkaar in de buurt wonen en elkaar nog regelmatig opzoeken. Ik ben dankbaar dat we al die tijd al lief en leed delen en dat onze kindjes nu zo gezellig samen kunnen spelen. Jullie blijvende interesse en vertrouwen in mijn proefschrift heeft me erg geholpen al die tijd.

Natuurlijk wil ik hier ook mijn familie bedanken. Of ik nu een grote reis maak of aan een nieuwe baan begin, jullie staan altijd voor me klaar en steunen me in mijn keuzes. **Heit**, wat jammer dat je dit allemaal niet meer mee kunt maken. Ik had nog zo graag meer bijzondere momenten met je gedeeld. Gelukkig heb jij al je meiden goed achtergelaten. **Mam**, **Marian** en **Willemien**, ik ben trots dat we het met zijn vieren zo goed doen. Marian, wat fijn dat je mijn paranimf wilt zijn! Mijn schoonfamilie wil ik bedanken voor hun niet aflatende interesse in mijn proefschrift en alle praktische steun in de afgelopen jaren.

En als allerlaatste maar ook allerbelangrijkste, mijn partner **Joost**. Ik waardeer alle tijd en ruimte die ik van jou heb gekregen om aan mijn onderzoek te werken en wil je bedanken voor je onuitputtelijke vertrouwen, steun en positiviteit. Zelfs op momenten dat ik het zelf niet meer zag zitten hield jij vertrouwen en wist je me te motiveren om door te zetten. Ik kan je niet genoeg bedanken voor al je hulp bij zowel de totstandkoming als de afronding van dit proefschrift. Zonder jou was dit proefschrift er niet geweest en had het er niet zo mooi uitgezien. Ik ben trots op ons! En tot slot, onze dochter **Eva**. Wat hebben we het gezellig met jou erbij in ons leven. Je bleek de beste stok achter de deur om een punt achter dit proefschrift te zetten, zodat ik vanaf nu alle tijd en aandacht heb voor ons gezin.

Curriculum Vitæ

Sytske Wiegersma

Sytske Wiegersma was born in Smallingerland, the Netherlands, on 16 August 1988. She graduated from secondary school at the Drachtster Lyceum in Drachten and moved to Enschede in 2006 to study Educational Sciences at the University of Twente. During her studies she developed an interest in research methodology and data analysis, especially when applied in the (mental) health care domain. After graduating in 2011, she worked as a data analyst for the department of Pulmonary Medicine at the Medisch Spectrum Twente (MST) before traveling through South America for 1.5 years. When she returned to the Netherlands in 2013 she started her PhD research on the reuse of clinical data, which she combined with a position as Data Manager at the IGS Datalab. After three years, her position at the University of Twente ended and she started working as Data Scientist at the Rijnstate Hospital in Arnhem on a short project on the prediction and prevention of hospital readmissions. She then worked as a Researcher at Nivel for three years, focusing on the reuse of routinely recorded primary care data and the learning health care system. She now works at the Integraal Kankercentrum Nederland (IKNL) as Adivsor medical guidelines for pallative care. She combined her work with her PhD research, leading to the completion of this thesis in 2022 under supervision of prof.dr.ir. B.P. Veldkamp and prof.dr. M. Olff.

Sytske lives together with her partner Joost van Noije. Together they have a daughter Eva (2020) and a baby on the way.

Contact: sytske@wiegersma.nl

List of Publications

Related publications

1. **Wiegersma, S.**, Hidajat, M., Schrieken, B., Veldkamp, B.P., & Olff, M. (2022). Improving Web-based Treatment Intake for Multiple Mental and Substance Use Disorders by Text Mining and Machine Learning: Algorithm Development and Validation. *JMIR Mental Health, 9*(4):e21111. https://doi.org/10.2196/21111

2. **Wiegersma, S.**, Nijdam, M.J., Van Hessen, A.J., Truong, K.P., Veldkamp, B.P., & Olff, M. (2020). Recognizing hotspots in Brief Eclectic Psychotherapy for PTSD by text and audio mining, European Journal of Psychotraumatology. *European Journal of Psychotraumatology, 11*(1). https://doi.org/10.1080/20008198.2020.1726672

3. **Wiegersma, S.**, Flinterman, L.E., Seghieri, C., Baldini, C., Paget, J., Barrio Cortés, J., & Verheij, R.A. (2020). Fitness for purpose of routinely recorded health data to identify patients with complex diseases: The case of Sjögren's syndrome. *Learning Health Systems, 4*(4), e10242. https://doi.org/10.1002/lrh2.10242

4. Smink, W., Sools, A.M., Van der Zwaan, J.M., **Wiegersma, S.**, Veldkamp, B.P., & Westerhof, G.J. (2019). Towards text mining therapeutic change: A systematic review of text-based methods for Therapeutic Change Process Research. *Plos one, 14*(12), e0225703. https://doi.org/10.1371/journal.pone.0225703

5. **Wiegersma, S.**, Sools, A.M., & Veldkamp, B.P. (2016). Exploring "Letters from the Future" by Visualizing Narrative Structure. *Proceedings of the 7th Workshop on Computational Models of Narrative (CMN 2016), 53*, 5:1-5:18. https://doi.org/10.4230/OASIcs.CMN.2016.5

Related presentations

1. **Wiegersma, S.**, Mink-Nijdam, M.J., Van Hessen, A.J., Olff, M., & Veldkamp, B.P. (2017). *Recognizing hotspots in Brief Eclectic Psychotherapy for PTSD by text and audio mining.* 15th European Society for Traumatic Stress Studies, Odense, Denmark.

2. **Wiegersma, S.**, Van Noije, A.J., Sools, A.M., & Veldkamp, B.P. (2016). *DIY Text Classification – The development of a supervised text classification tool.* RCEC workshop Item Response Theory and Educational Measurement, Enschede, The Netherlands.

3. **Wiegersma, S.**, Sools, A.M., & Veldkamp, B.P. (2016). *Exploring "Letters from the Future" by visualizing narrative structure.* Computational Models of Narrative - Digital Humanities, Cracow, Poland.

4. **Wiegersma, S.**, Sools, A.M., Veldkamp, B.P., & Westerhof, G.J. (2016). *What works when for whom? Advancing therapy change process research by mining for therapy-related textual features in effective e-mental health interventions.* Supporting Health by Technology VII, Groningen, The Netherlands.