

Patient sc-AI-nce

methods for studying patient experiences
with text mining analysis

Research group: Nada Akrouh, Peter van Huisstede, Lea Jabbarian, Jiwon Jung-IO,
Jasper op de Coul & Rik Wehrens

Advisory Board: Hester van de Bovenkamp, Janne Papma, Judith Rietjens, Iris
Wallenburg & Erica Witkamp

23rd December 2021



convergence

Introduction

In the last decades, chronic illnesses have become an ever more pertinent area of attention for researchers and policy makers alike (Thorne, 2006). As more people are living longer with a variety of chronic illnesses, understanding the everyday impact of such illnesses on patients' lives becomes much more important. Understanding the experiences of living with chronic illness calls for an in-depth understanding of the knowledge of patients (Donaldson, 2003; Greenhalgh 2009). This knowledge concerns different aspects of living with a certain condition; sickness (the objective complaints according to medical science), illness (the social aspects of having a condition with attention to the role of the patient and the influence of the environment) and disease (the experiential aspect of having a condition (Frank 2013). Explicating this knowledge can provide peer support to other patients and help (future) professionals and policy makers to improve their services and policies.

Increasingly patients share their detailed experiences and knowledge about their conditions through ego-documents like books and blogs. In the Netherlands, various collections of patient stories have been initiated to bring these stories together. A unique example is the patient stories collection of the EUR which contains over 6000 books written by patients and informal caregivers (www.patientervaringsverhalen.nl). Although these stories contain rich, narrative data, these sources have only sporadically been used for research (Van de Bovenkamp et al., 2020). One explanation for this is that these knowledge sources do not fit easily into traditional and institutionalized knowledge practices in medicine, which often focus on more aggregated forms of knowledge (e.g., standard patient outcome measures (PROMs), clinical guidelines). Such aggregated forms of knowledge, however, do not always capture the themes and experiences that matter to patients in their everyday lives and care trajectories.

A crucial question therefore emerges: how can researchers, healthcare professionals and policymakers find ways to utilize the deep and rich contextual insights of patient experiences captured in patient stories while also building towards more generalizable and comprehensive themes? There are different ways to do so, for example through thematic and narrative analysis (see for examples www.patientervaringsverhalen.nl). In this pilot project we have explored this question by an innovative combination of methods to extend our knowledge about their potential use for peer support, quality improvement of care services, education and policy-making.

While traditionally (patient) stories are analyzed via qualitative methods such as thematic or narrative analysis, novel computational methods (such as text mining and machine learning) offer promising new opportunities to unlock relevant knowledge within the large amount of patient stories. The use of such computational methods in social science is encouraged by many (Antons et al., 2020; Wang, 2017; Németh &

Koltai, 2021) as they promise to identify patterns in the idiosyncratic experiences of patients.

Despite the great promise of the use of machine learning and text mining for retrieving patient knowledge and rendering it part of established medical and policy knowledge practices, the actual use of these methods is limited. In this pilot project, we combine the methodological strengths of novel computational methods (text mining and machine learning) and qualitative analysis. In this study, we have experimented with this combined approach to analyze the stories of persons diagnosed with a psychotic disorder with a specific focus on how they experience social integration and stigma. This requires expertise from various domains, including the social sciences, the medical domain, psychology and the engineering disciplines.

In this document we reflect on our experiences in this Open Mind project and report on the process and results of this pilot. We detail the iterative and interactive approach we have developed in this pilot and outline various scenarios on how computational methods and qualitative research can be productively combined. We discuss the main lessons we have learned and provide several recommendations for other researchers interested in analyzing patient narratives through computational methods. We also summarize the key steps in a methodological guideline to facilitate the use of text mining to retrieve patient knowledge (from p.14 onwards).

The patient sc-AI-nce project: process and outcomes

In September 2021, a group of researchers from different backgrounds embarked on a four-month pilot project called patient sc-AI-nce. These different researchers brought in domain expertise regarding psychiatric care and patient experiences, in-depth methodological expertise in qualitative methods (e.g., thematic analysis, narrative analysis, focus groups) and computational methods (e.g., machine learning, text mining). The small-scale project served as a first pilot that aimed to investigate how computational methods and qualitative research methods could be productively combined in the context of analysing patient stories. The project idea consisted of three work packages: 1) narrative analysis; 2) text mining; 3) evaluation. The prominent research goal was to work towards a methodological guideline that can facilitate other researchers in the challenging task of converging qualitative research methods with computational methods.

In this report, we will elaborate on both the process and outcomes of the Patient Sc-AI-nce project. Figure 1 presents a summary of our work. This includes a triad of processual elements (focused on transdisciplinary collaboration), content-related activities (to extract meaningful insights from the books), and results of the different steps. The guideline is structured according to these three key elements. In the end, we summarize the key lessons and recommendations and based on our exploratory work we discuss a preliminary research agenda.

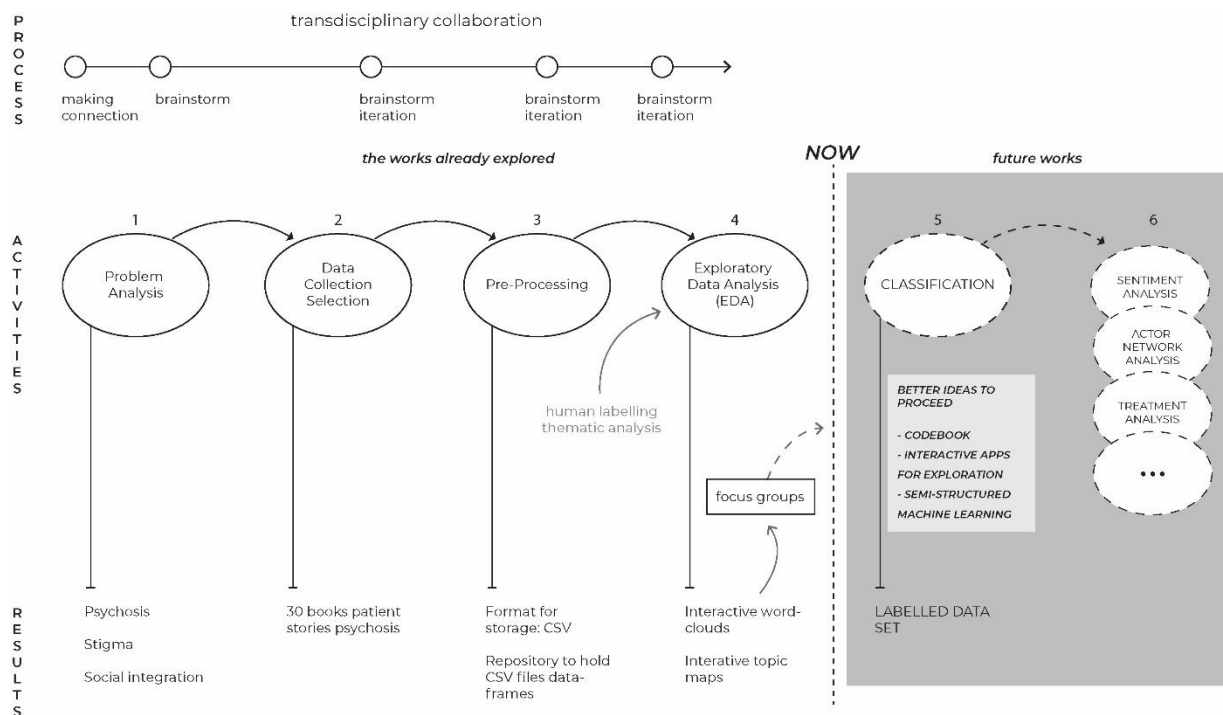


Fig.1 Patient sc-AI-nce project

*Imaged initially created by Peter van Huisstede
Digitally illustrated by Jiwon Jung*

I Process: Transdisciplinary collaboration

The research group consisted of social scientists, data scientists, an industrial engineer and a clinical psychologist. They met on a regular basis during the research project. The first month the collaboration was focused on gaining better understanding of each other's disciplinary backgrounds and methods, and becoming more familiar with the data by reading several patient stories.

For this purpose, we have organized a four-day workshop about text mining at Erasmus School of Health Policy & Management (see box 1). The workshop was attended by the research team and open to other researchers at the institute. The tailored program allowed us to explore each other's practices, language and methods. This facilitated mutual understanding and helped us to get a better grasp on the problem analysis and the methodological tools available to address the research questions.

We also realized that collaboration requires regular meetings around specific topics. Therefore, we organized various meetings where researchers from different disciplinary backgrounds show and explain their working procedures, such as the qualitative coding process of textual data, the writing of a Python script and the text mark-up.

Although we tried to map out the steps needed at the beginning we also realized that we as researchers had to be flexible more than usual. We were aware of the several iterations that would be needed, but how these would look like was hard to predict. That is also why we, as the project proceeded, organized several prolonged brainstorm meetings with the core research team to discuss our ideas. Such regular meetings proved highly useful in integrating knowledge from different disciplines (cf. Stevens et al., 2020). The added value of this is enhanced mutual understanding. On the one hand, it allowed data scientists to better assess the relevance of some outcomes for social scientists. On the other hand, social scientists became progressively better attuned towards interpreting computational results and asking more specific questions, such as asking for the creation of a chart which shows the distribution of certain specific terms (e.g., stigma) across the dataset.

Box 1 Workshop series text mining

In September 2021 Jasper op de Coul and Peter van Huisstede organized a workshop series about text mining at Erasmus School of Health Policy & Management (ESHPM). Four workshops were organized with the aim of giving participants a basic understanding of the processes and possibilities of text mining. The book 'Blueprints for Text Analytics Using Python' by Albrecht, Ramachandran & Winkler (2020) was used as a guideline. In the first workshop the participants were introduced to the field of text mining. They learned about the most important concepts, Exploratory Data Analysis (EDA), data preparation and about the programming language Python. Examples of important concepts are supervised learning, unsupervised learning, TF-IDF (Term Frequency — Inverse Document Frequency) tokenization and vectorization. In the second workshop the focus was on unsupervised learning. Participants became familiarized with topic

modeling, and this was done by a demonstration of this method on a dataset of all scientific articles written by ESHPM. Also, nonnegative matrix factorization (NMF) and word2Vec were demonstrated. In the third workshop supervised learning was addressed. The same dataset of scientific articles used in workshops 1 and 2 were used for performing supervised learning. This dataset was split into a training and testing set, and this exercise showed the participants how the computer predicted the division of the academic articles across the seven departments of ESHPM. The results were interpreted jointly by the participants and the facilitators of the workshop. In the last workshop we learned about sentiment analysis, and the participants reflected upon the lessons learned from the previous workshops.

II Activities & Results

Problem analysis

Our main research goal for this project was to explore ways of integrating qualitative methods with text mining for the analysis of patient stories. We focused on psychotic disorders, and more specifically how persons diagnosed with a psychotic disorder experience social integration and stigma. This choice was partially pragmatic (several books pertaining to this category were already digitized for another research project at ESHPM, making it easier to apply text mining methods), but also content-driven. Persons diagnosed with a psychotic disorder often perceive relations with care professionals to be complex and even adversarial. Moreover, experiential knowledge in this domain is increasingly recognized as important, as can be seen in the enhanced role for experiential experts (*ervaringsdeskundigen*) in this sector.

The initial idea was to develop a supervised machine learning model to analyze and classify themes within patient stories based on a sample of the annotated stories. However, this sample of hand coded data turned out to be too small for training the data. Because of time constraints it was not possible to generate a considerable amount of manual coded data, which is crucial for the development of a training set (Grimmer & Stewart, 2013). As we got more insights into the different methods as well as the possibilities, we decided to let go of the supervised machine learning idea. The aim was adjusted to the development of a semi-supervised learning model.

Data collection and selection

In line with the problem analysis this project focuses specifically on stories of people diagnosed with a disorder in the psychotic spectrum and is linked to the NWO research project stories about social (re)integration¹. The dataset of digitalized books in this research project was used for the selection of the books for our research project. All these books written in Dutch and derived from patientervaringsverhalen.nl. The dataset

¹ <https://www.eur.nl/eshpm/onderzoek/verhalen-over-maatschappelijke-integratie>

consisted of 38 books. From this dataset we excluded duplicates, comic books and books that were for a big part theoretical (i.e., books that are not completely autobiographical). This resulted in a dataset of 30 books (see Appendix I).

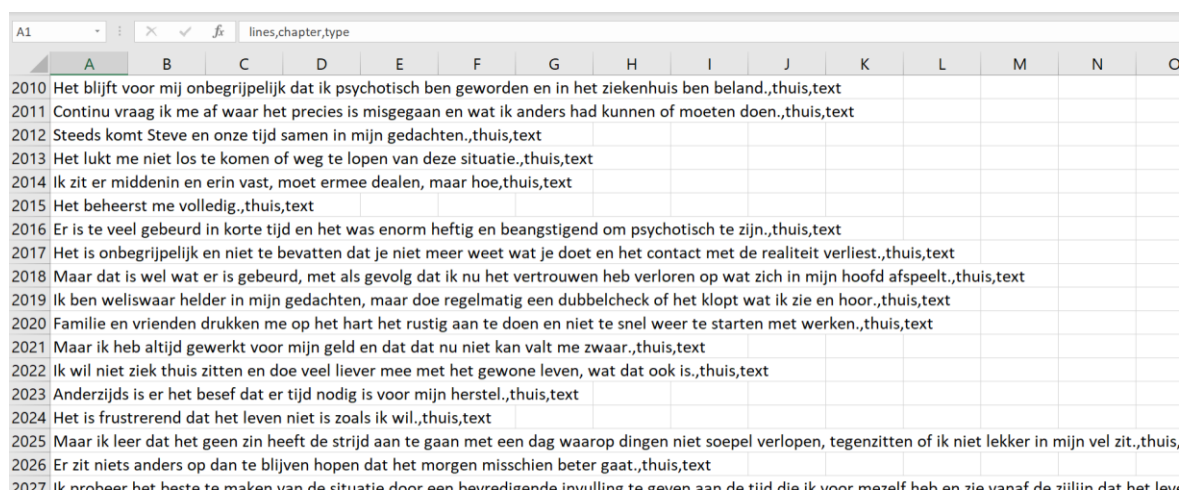
Pre-processing

All the 30 books were available in pdf format. The digital texts were explored with command line interface (CLI) tools, terminal and unix, like: sed, grep, ack, and awk. The digitalized texts consisted of multiple Optical Character Recognition (OCR) errors. These OCR errors were corrected. An example of an OCR error is the merge of the letters *n and i* into *m*. In this example this is solved by replacing the *m* by the *n and i*. The task of cleaning the data is a time-consuming but necessary step in the process.

Then, we moved on to the mark-up of the texts. In this step metadata is added to the text files. We decided to go for a more general approach where each book consists of lines each holding a sentence of a book. Front- and back matter were discarded. If a book contains poems, then these poems were collected in a separate file. This resulted in a format to store patient stories as "standardized" plain text files. These files can be easily read into a Pandas dataframe for analysis. A Pandas dataframe is a row/column format that is organized as follows: rows contain observations on variables that are represented by columns. The first column, "lines", holds all sentences of a book as rows. Columns can be easily added either by NLP software (i.e. add a column containing all nouns from a sentence).

We chose to write the dataframe to CSV for storage and/or further processing. CSV files are suitable for further use within collaborative research. This results in a repository containing the patient stories in CSV format, both the standard ones (only containing the sentences of the books) and the enriched ones (which also entail metadata).

Box II Prepared text in CSV format



lines	chapter	type
2010		Het blijft voor mij onbegrijpelijk dat ik psychotisch ben geworden en in het ziekenhuis ben beland.,thuis,text
2011		Continu vraag ik me af waar het precies is misgegaan en wat ik anders had kunnen of moeten doen.,thuis,text
2012		Steeds komt Steve en onze tijd samen in mijn gedachten.,thuis,text
2013		Het lukt me niet los te komen of weg te lopen van deze situatie.,thuis,text
2014		Ik zit er middenin en erin vast, moet ermee dealen, maar hoe.,thuis,text
2015		Het beheerst me volledig.,thuis,text
2016		Er is te veel gebeurd in korte tijd en het was enorm heftig en beangstigend om psychotisch te zijn.,thuis,text
2017		Het is onbegrijpelijk en niet te bevatten dat je niet meer weet wat je doet en het contact met de realiteit verliest.,thuis,text
2018		Maar dat is wel wat er is gebeurd, met als gevolg dat ik nu het vertrouwen heb verloren op wat zich in mijn hoofd afspeelt.,thuis,text
2019		Ik ben weliswaar helder in mijn gedachten, maar doe regelmatig een dubbelcheck of het klopt wat ik zie en hoor.,thuis,text
2020		Familie en vrienden drukken me op het hart het rustig aan te doen en niet te snel weer te starten met werken.,thuis,text
2021		Maar ik heb altijd gewerkt voor mijn geld en dat dat nu niet kan valt me zwaar.,thuis,text
2022		Ik wil niet ziek thuis zitten en doe veel liever mee met het gewone leven, wat dat ook is.,thuis,text
2023		Anderzijds is er het besef dat er tijd nodig is voor mijn herstel.,thuis,text
2024		Het is frustrerend dat het leven niet is zoals ik wil.,thuis,text
2025		Maar ik leer dat het geen zin heeft de strijd aan te gaan met een dag waarop dingen niet soepel verlopen, tegenzitten of ik niet lekker in mijn vel zit.,thuis,
2026		Er zit niets anders op dan te blijven hopen dat het morgen misschien beter gaat.,thuis,text
2027		Ik probeer het beste te maken van de situatie door een bevredigende invulling te geven aan de tijd die ik voor mezelf heb en zie vanaf de riilijn dat het lev

Above we see the prepared text of the book 'Wendingen' (in English 'Turns') written

by Anna de Witt. This book has been 'cleaned' by removing noise and correcting all OCR-mistakes. In this CSV file the book is organized by dividing it into lines, chapter where the sentence falls under and type of data. All sentences in the book are represented in the lines in this file. In the screenshot of the CSV file above we see that these sentences fall under the chapter 'thuis' (in English 'home'). Here the author describes how she experiences being home after being discharged from the hospital. This way of structuring the text allows for a more targeted application of analysis. For instance, when focusing on the concept of social integration we can select only the text components that capture this very concept.

Thematic analysis

A selection of the books were manually coded by the social scientists. Atlas.ti Web was used for coding the books. When coding the books, the focus was on the concepts of stigma and social integration. Seven books were read, and the coding resulted in a total of 439 codes. An annotation guideline in the form of a hierarchical codelist (see Appendix II) was developed. This list was continuously being sharpened as we took an abductive approach (Tavory & Timmermans, 2014). Based on theories of stigma and social integration a first list was made, which was completed with terms that emerged from the data. Although we did not use the human labelling as training data, we did use it for input during the Exploratory data analysis (EDA).

Exploratory data analysis (EDA)

Exploratory data analysis is usually a standard step taken in text mining analysis. The goal is to learn about the dataset by visualizing it in different ways. In the heuristic stage we analyzed simple word frequencies, used TF-IDF to compare document word frequencies weighted against the corpus of books. We used bigrams (two-word sequence of words) and trigrams (three-word sequence of words) to get a better grasp of the concepts used in the books. This step was interesting since it gave us, for the first time, quantitative insights into (elements of) the books. The descriptive statistics and TF-IDF showed us that the books vary in terms of the number of words per book. This led to questions about distribution and representativeness of the patient experiences and the implications thereof for the analysis. These are relevant questions for follow-up research as these matters also relate to data preparation. Thereupon we decided to move on to other methods of EDA.

A Jupyter notebook was created that reads in the texts of the books selected for this project. After preprocessing of the texts, we further explored the contents of the books against the two research goals: social integration and stigma. Various kinds of exploratory data analyses were explored, such as word clouds and topic modeling. Generating word clouds based on word frequency and importance is a well-known way of generating preliminary insights from text via computational methods. They can be

used to introduce common themes and give a description of a certain setting using 'grand tour' questions (Albris et al., 2021). Topic modeling is a machine learning method that discovers themes in a large textual dataset using algorithms (Blei, 2012). Topic modeling therefore is an unsupervised learning method. With the topic models in hand, it is the task for researchers to interpret the clusters created. Both LDA and NMF topic modeling are performed on the dataset of books, and the relevance of the outcomes were discussed by the research team. The research group came together to deliberate about the development of a tool where qualitative methods could be combined with these unsupervised learning methods. After several brainstorm sessions the group eventually aimed at the development of an interactive word cloud tool and interactive topic maps using Python scripts:

1. An interactive word cloud tool that allows users to zoom in on certain concepts using parameters like window-size over the text to be worked with (an application of the KWIC, keyword in context, approach) and possibilities, in the form of sliders, for adjusting the weight of TF-IDF, bigrams, nouns, and verbs (see figure 2).
2. An interactive topic modelling application that visualized the topics of the books and the distribution of the terms considered for this against all terms of the corpus.

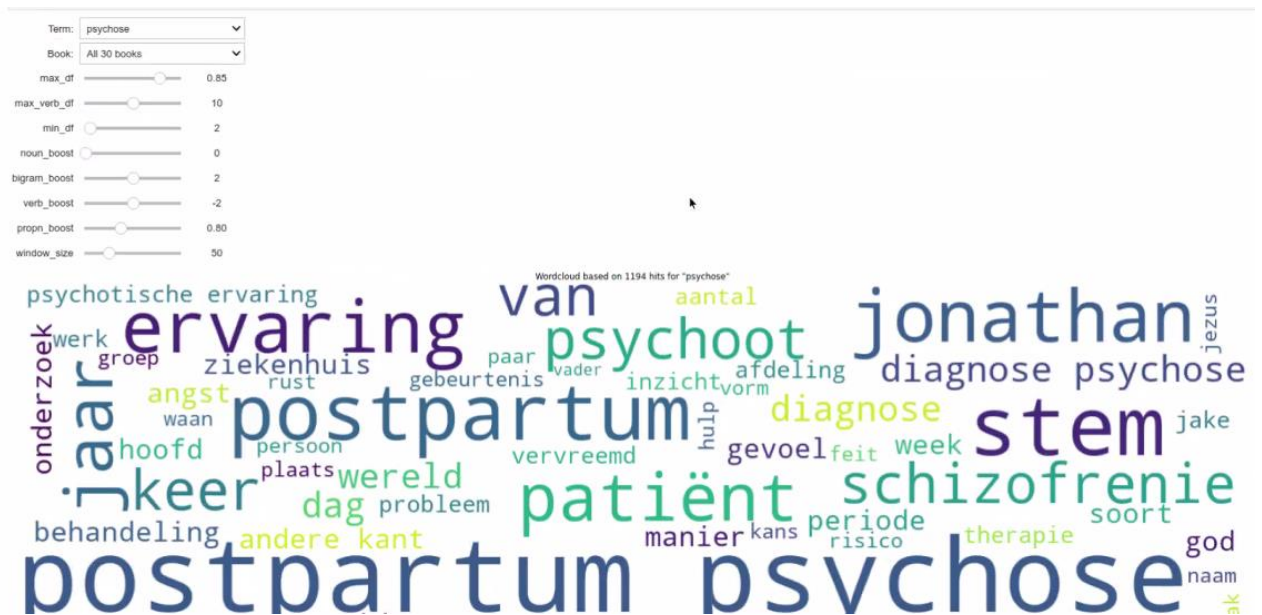


Fig.2 Interactive word cloud tool

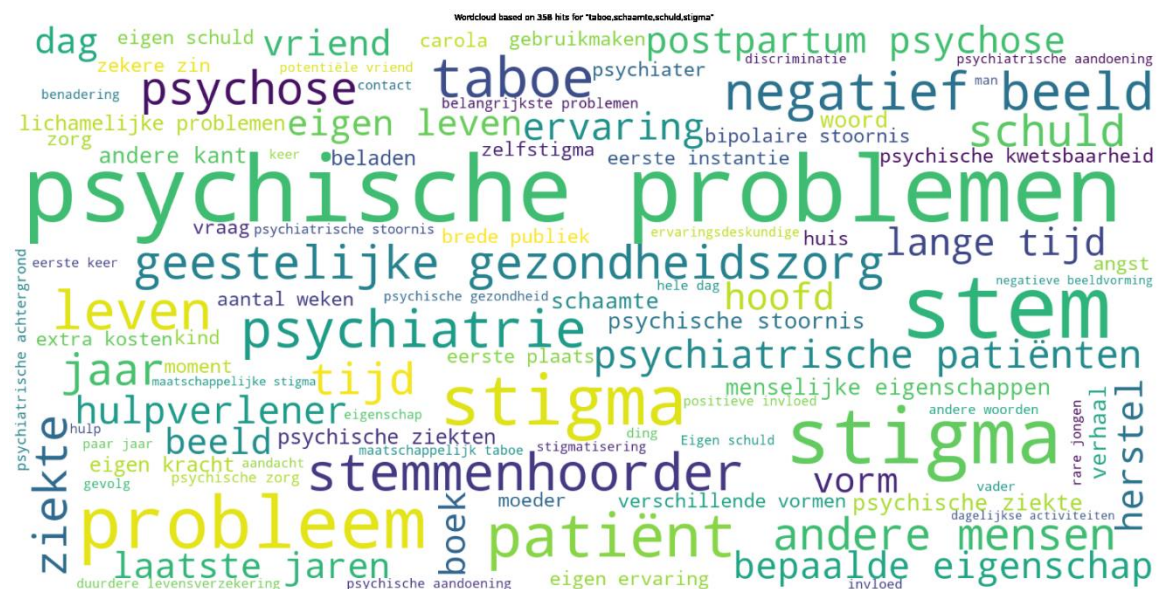
As it was important for us to conduct focus groups with various people, we chose to only include the word cloud tool during these focus groups. We found out that topic modelling is somewhat hard to grasp if one is not familiar with the method. This is not the case for word clouds, these are quite accessible because the visuals are easier to understand and prior knowledge into this method is not required for understanding them. As such, word clouds offered the most suitable way for us to present insights of text mining to different groups because they aligned most with our aim to use the outcomes as a *tool for facilitating in-depth discussion*. We expected that this way of

presenting the preliminary findings would allow respondents to reflect more easily on these results and enable a richer and more meaningful conversation.

When the word cloud tool was generated, the next task was to generate different word clouds which would be used during the focus groups. The first step is to come up with input terms. We decided to base these input terms on the thematic analysis we conducted on a selection of the books. Then we organized a meeting with the research team for the selection of the word clouds.

Besides the fact that word clouds are more accessible for the general public, the choice to focus on the interactive word cloud tool and interactive topic maps fits the experimental character of this research.

Box III word cloud theme stigma



In the figure above we see a word cloud that is generated with our word cloud tool. The input terms were 'taboo', 'shame', 'guilt' and 'stigma'. This word cloud consists of all kinds of words: verbs, nouns, adjectives and personal pronouns. It also consists of both unigrams (e.g. 'taboe') and bigrams (e.g. 'bepaalde eigenschap').

Focus groups

Already at the start of the research we wanted to organize a focus group with the aim of gathering insights from experts in the field. Besides enabling a deeper discussion about the outcomes of the text mining analysis, we also aimed to use the focus groups to gain better insights into how text mining as a method is experienced: what meanings participants attribute to it, as well as what benefits and limitations they identify. In the focus groups we therefore focused on whether the themes that emerged from the data

analysis resonate with the lived experiences and professional knowledge of participants, why and how they did or did not resonate, and whether these emerged themes can indeed facilitate a meaningful dialogue about lived experiences.

At the end of the second month the preparations for the focus groups started. Initially we aimed at organizing one focus group, which would be attended by both mental health care professionals and experts by experience. However, as we sharpened our aim of the focus groups we decided, after deliberation with the research team and advisory committee, to split the groups and plan two focus groups. This meant that one focus group was dedicated to professionals and one to experts by experience.

The two focus groups took place in November 2021. By organizing the focus groups, we aimed at extracting professional knowledge as well as experimental knowledge on the initial analyses that came forth out of EDA (Jalonen et al., 2021). In the first place we aimed at understanding how participants appraise the word clouds. Moreover, we aimed to understand how matters around stigma and social integration were experienced by professionals and experts by experience.

The first focus group was attended by seven mental health care professionals who work at the psychiatry department of a hospital. A selection of five word clouds was shown during this focus group. We divided the focus group in three sections: 1) introduction to the research project and aims; 2) presentation of the word clouds and content-based discussion; 3) joint methodological reflection. The main part consisted of the discussion about the word clouds. We presented three word clouds: one based on the theme of 'psychosis', one based on the theme 'stigma', and one based on the theme 'social integration'. To facilitate the discussion, we asked the participants three questions:

- What surprises (in positive or negative way) you in the word cloud?
- What do you recognize in the word cloud?
- What do you miss in the word cloud?

Results show that the word clouds and our three questions activated substantive discussion about psychosis, stigma and social integration and our used methodology. The professionals' perception of the psychosis experiences of patients as depicted in the word clouds and our used methodology were two central points. Professionals recognized most of the terms in the word clouds about the themes 'psychosis' and 'stigma'. They could relate these terms to examples of their contact with patients. Most surprising was the word cloud about social integration. Professionals found it difficult to connect to the terms in this word cloud, because most terms referred to aspects outside the environment of health care. What professionals missed in the word clouds were the themes such as fatigue, trauma and specific therapies. These points were consequently linked to the methodological points raised about the word cloud tool and the sources. Participants also had questions about the context of the stories, they were

curious about the background of the authors. They were interested in details like the diagnosis, age and the motivation of authors to write a book. They had the idea that by knowing more about the authors they could make better sense of the terms in the word clouds.

The second focus group was attended by six authors who wrote about their experiences with psychosis. Here we followed the same structure as the first focus group, but made several changes based on the experiences during the first focus group. One difference is that we shared the word clouds we would discuss during the focus group beforehand with the participants. This choice was made because we learnt from the first focus group that participants needed a considerable amount of time to read and digest the content of the word clouds. We tried to overcome this during the second focus group by asking participants to prepare by taking a look at the word clouds before the focus group took place. A second change we made was the decision to present several word clouds based on *individual* books. The reason for this was that we felt that presenting these individual word clouds better captured the large diversity of experiences we witnessed across the books. We also expected that this would enable a deeper discussion.

Also, this focus group activated an interesting discussion about both the content of the word clouds as well as the methods we used. A difference in the second focus group was that the discussion about our used methods was central. The participants differed in their perceptions toward the content of the word cloud. While some were positively surprised by the terms and took meaning out of it, others were critical and questioned the significance of the word clouds. The latter point formed the rest of the discussion towards the question: what are valuable ways for using the stories in research using text mining? Participants for instance missed information about the context, such as how authors portray their psychotic episodes. This was seen as the essence of psychosis which could not be identified when analyzing the word clouds. In order to see how the contexts about psychosis change over time, historical analysis of the books was seen as an interesting idea that came out of the discussion.

Future work

Above we described our mixed method approach on the analysis of stories about persons diagnosed with a disorder in the psychotic spectrum. As the project proceeded, we learned from each other's expertise and put these together to explore and develop methods for combining qualitative research methods and computational methods. On the one side our research was theory driven. We used theories of stigma and social integration as foundation for labelling the books. On the other side our research was data driven, as we were open to exploring new aspects. Where supervised methods are most suitable for a theory driven approach, unsupervised methods are most suitable for a data driven approach (Németh & Koltai, 2021). The combination of both methods is a common practice. We tried to combine these during the EDA, where we chose to focus

on the word cloud tool. This word cloud tool allowed us to generate a variety of word clouds based on the themes stigma and social integration. A selection of these word clouds was presented during two focus groups.

This research project provided us with several lessons. Most important is that we learned that collaborative efforts of domain experts, data scientists and qualitative researchers are very valuable and enriching for every discipline. It allowed us to reflect on different steps in the process of text mining and to identify possibilities and limitations.

The exploratory character of our research allowed us to develop several ideas for future research. A first idea is to combine our approach with citizen science. The point of departure in citizen science is that citizens participate in all stages of research. By including citizen scientists in collaboration various possibilities emerge. Their involvement and input could help to further shape the substantive process of the research. An example is by developing tools where citizens can label the dataset.

A second idea is to further develop our tools into interactive apps which can be used in several ways like in focus groups. Based on our experiences of the focus groups we remarked that participants were especially interested in the *generation* of the word clouds. Participants could be given the opportunity to generate word clouds using the word cloud tool. This would be relevant because it allows both professionals and experts-by-experience to identify central words, themes or descriptions that they feel capture key dimensions of the patient experiences. As such, working interactively with word clouds can also provide new insights for researchers and policymakers.

Another idea is to further develop our labelled dataset, and apply various existing data analysis methods, such as sentiment analysis, actor-network analysis and association rule mining. Each technique will lead to unique analysis results to understand the experience of patients, such as the affective status of the text, stakeholders involved, relationship between stakeholders and frequency of co-occurring events.

A getting started guide for combining qualitative analysis and text mining analysis for analyzing patient stories

In this section we offer a guide for starting out with the method of combining qualitative analysis and text mining analysis of patient stories. This guide is based on all the steps taken during our research project. These steps and lessons can offer guidance for one who is starting out with combining qualitative analysis and text mining analysis for analyzing patient stories. With starting out we imply the steps up to and including the exploratory data analysis. Besides the methodological steps we also include the process of facilitating transdisciplinary collaboration. This is important since this collaboration is eminent in the sense that it allows for joining forces and integrating knowledge and expertise of different disciplines.

Step 1 Set up a transdisciplinary collaboration

1. Connect expertise on qualitative research and data science
 - a. Collaboration between qualitative researchers and data scientists
 - b. Qualitative researcher with data mining skills/ data scientist with qualitative research skills

Step 2 Problem analysis

1. Decide about research questions based on field of expertise

Step 3 Data collection and selection

1. Specify disease group
 - a. Broad (e.g., dementia) or specific (e.g., vascular dementia)
2. Specify perspective
 - a. All perspectives
 - b. Patient/client
 - c. Family
3. Specify population group
 - a. Age category (children/adolescents/adults/elderly)
4. Specify time/period

Step 4 Select method

1. Select text mining method
 - a. Supervised learning
 - b. Unsupervised learning
 - c. Semi-supervised learning

2. Select qualitative research methods
 - a. Human labelling
 - i. Development annotation guideline (based on research question)
 - ii. Annotation by researchers
 1. Reading # books and open coding
 2. Intercoder agreement
 3. Development hierarchical code-list
 - iii. Annotation by experts (professionals/ expert by experience)
 - iv. Comparison annotations
 - v. Set-up systematic thesaurus
 - b. Focus groups
 - i. Professionals
 1. Informing participants
 2. Selecting visualizations during focus group:
 3. Preparing questions
 4. Provide participants with hand-outs before focus group
 - ii. Patients/experts by experience (EBE)
 1. Informing participants:
 2. Selecting visualizations during focus group:
 3. Preparing questions
 4. Provide participants with hand-outs before focus group
 - c. Interviews
 - i. Professionals
 1. Informing participants
 2. Selecting visualizations during focus group:
 3. Preparing questions
 4. Provide participants with hand-outs before focus group
 - ii. Patients/experts by experience (EBE)
 1. Informing participants:
 2. Selecting visualizations during focus group:
 3. Preparing questions
 4. Provide participants with hand-outs before focus group

Step 5 Text preparation

1. Digitalization of selected books using OCR software
2. CLI text exploration
3. Solving OCR mistakes
4. Markup text (e.g., start and end of chapter, poems, patient files, images)
 - a. Standard (only sentences)
 - b. Enriched (sentences + metadata)
5. Removing noise

- a. Remove unnecessary or irrelevant parts in text (e.g. front and back matter)
6. Tokenization

Step 6 EDA

1. Descriptive statistics
 - a. Number of words per book
2. TF-IDF
3. Stemming
 - a. Unigrams
 - b. Bigrams
 - c. Trigrams
 - d. Lemmatization
4. Word clouds (book word cloud tool)
 - a. Run on whole corpus
 - b. Run on one specific document or multiple documents
 - c. Run on specific term(s) with sliders (window etc.)
 - i. Choose input term (s)
 1. Based on theory
 2. Based on human coding of data (open coding list)
 3. Abductive
 - ii. Choose documents
 1. All books
 2. One specific or multiple documents
 - d. Interpretation by researchers
 - e. Validation
 - i. Generating word clouds with different researchers
 - ii. Check with all involved researchers
 - f. Representativeness
 - i. Checking distribution of input word across documents
5. Topic modelling
 - a. Topic modelling 1 (NMF)
 - b. Topic modelling 2 (LDA)
 - c. Interpretation by researcher

Bibliography

Albris, K., Otto, E. I., Astrupgaard, S. L., Gregersen, E. M., Jørgensen, L. S., Jørgensen, O., ... & Schønning, S. (2021). A view from anthropology: Should anthropologists fear the data machines?. *Big Data & Society*, 8(2), 205395172111043655.

Antons, D., Grünwald, E., Cichy, P., & Salge, T. O. (2020). The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. *R&D Management*, 50(3), 329-351.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

van de Bovenkamp, H. M., Platenkamp, C., & Bal, R. (2020). Understanding patient experiences: The powerful source of written patient stories. *Health Expectations: an international journal of public participation in health care and health policy*, 23(3), 717.

Donaldson, L. (2003). Expert patients usher in a new era of opportunity for the NHS. *BMJ: British Medical Journal*, 326(7402), 1279.

Frank, A. W. (2013). *The wounded storyteller: Body, illness, and ethics*. University of Chicago Press.

Greenhalgh, T. (2009). Chronic illness: Beyond the expert patient. *BMJ: British Medical Journal*, 338(7695), 629–631.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3), 267-297.

Jalonen, H., Kokkola, J., Laihonen, H., Kirjavainen, H., Kaartemo, V., & Vähämaa, M. (2021). Reaching hard-to-reach people through digital means—Citizens as initiators of co-creation in public services. *International Journal of Public Sector Management*.

Németh, R., & Koltai, J. (2021). The potential of automated text analytics in social knowledge building. *In Pathways Between Social Science and Computational Social Science* (pp. 49-70). Springer, Cham.

Stevens, M., Wehrens, R., & de Bont, A. (2020). Epistemic virtues and data-driven dreams: On sameness and difference in the epistemic cultures of data science and psychiatry. *Social Science & Medicine*, 258, 113116.

Tavory, I., & Timmermans, S. (2014). *Abductive analysis: Theorizing qualitative research*. University of Chicago Press.

Thorne, S. (2006). Patient-provider communication in chronic illness: A health promotion window of opportunity. *Family & Community Health*, 29(1), 4S–11S.

Wang, Y. (2017). Education policy research in the big data era: Methodological frontiers, misconceptions, and challenges. *Education Policy Analysis Archives*, 25, 94.

Appendix I: Book list

Book title	Author	Publisher	Year of publication
Uitgedokterd	Brenda Froyen	Manteau	2016
Herstel na dertien jaar overleven in de psychiatrie	Mieni Westerink	Elikser	2018
Julia	Hanneke Joosten	Water	2018
Het weerwolf clubje van DEF-huis	Chester Navarro	Lecturium	2019
Trap door gesloten kamers	Vaya Drent	Boekscout	2020
Een pscyhose door therapeutische training	Carlos Monteiro	Brave New Books	2014
Uit de schaduw van de waan	Karen van Dartel	Hogrefe Uitgevers	2012
Het zit allemaal in mijn hoofd	Mick	Free Musketeers	2016
Prettig gestoord is zo gek nog niet	Marijke Groenwoudt	Self-published	2013
De weg kwijt	Lenneke Derks	Boekscout	2011
Wendingen	Anna de Witt	Brave New Books	2015
Aangedane liefde	John Jongejan	Boekscout	2013
PsyCHAOTiSch	Claudia de Kok	Self-published	2011

Oscar	Nicolai van Doorn	Nationale media	2013
Hoezo gek	Phytia	Free Musketeers	2014
Elske de kolerehanger	Elske van Oenen	Memes	2011
Ik heb een gek te temmen	Nicky Samsom	Tobi Vroegh	2016
Vervreemd van mijzelf en mijn omgeving	Anja Stevens, Fleur Schreurs	Tobi Vroegh	2013
Verwarde man	Mark Verhoogt	De Graaff	2016
Psychose? Doe normaal!	Ayla	Boekscout	2015
Ongeluk in mijn hoofd	Yasmin Vermeer	Boekscout	2010
Omkeren	Joke de Jong	Boekscout	2017
Uit de goot	Paul Roozendaal	Lucht	2016
Op zoek naar mijzelf	Sharita Hart	Boekscout	2016
Alleen	Wouter Kusters	Lemniscaat	2007
De strijd van een psychoot	Jake Dutch	Free Musketeers	2016
Het ei van columbus	Ashley Geerlings	Brave New Books	2020
Angst en onrust	Karin den Oudsten	Free Musketeers	2011
Leven met stemmen	Marius Romme, Sandra Escher	Stichting leven met stemmen	2012
Briefwisseling tussen een	Felix Sperans, Axel Ruiters	Uitgeverij Gelderland	2019

schizofreen en een autist			
------------------------------	--	--	--

Appendix II: Hierarchical codelist

Stigma

- Self stigma
 - Self stigma thoughts
 - Self stigma actions
- Public stigma
 - Public stigma friends
 - Public stigma family
 - Public stigma professionals
 - Public stigma at work
 - Public stigma citizens
 - Public stigma media
- Reaction/dealing of patient with stigma
- Label
- Taboo

Social integration

- Social relationships/network
 - Family
 - Partner
 - Children
 - Parents
 - Siblings
 - Other family
 - Friends
 - Acquaintances (e.g. neighbor)
- Work
 - Work reintegration
 - Role of colleagues
- Study
- Living
- Activities
 - Writing
 - Sports
 - Going out
- (Different) behaviour after psychosis
 - New activities after psychosis
 - New thoughts after psychosis
 - New actions after psychosis
 - Reflections on life (old me vs. new me)
 - Acceptation
- Social support
 - Social support from friends and family
 - Lack of social support friends and family
 - Social support at work or school
 - Lack of social support at work or school
- Recovery
 - After care

