

Deep & Dutch: Exploring language markers in psychiatric stories

Stephanie Verkleij

s.j.a.verkleij@students.uu.nl
Student number: 5638380



Universiteit Utrecht



UMC Utrecht

Utrecht University
Department of Information and Computing Sciences

A thesis submitted for the degree of Master of Business Informatics

First supervisor: Prof. Dr. Marco Spruit
Second supervisor: Prof. Dr. Floortje Scheepers
Daily supervisor: Dr. Kees de Schepper

Abstract

Diagnosing mental disorders is complex due to the genetic, environmental and psychological contributors and the individual risk factors. Language markers for mental disorders can help diagnose a person. The differences in the usage of language between groups are called language markers. Research thus far on language markers and the associated mental disorders has been done mainly with the Linguistic Inquiry and Word Count program (LIWC). In order to improve on this research, we compare the following models LIWC and spaCy and the deep learning models fastText and RobBERT to Dutch psychiatric interview transcriptions. These are analysed with the goal to find out if a person has a mental disorder and if so which one. Furthermore, the second goal of this research is to find out which language markers are associated for a person with a mental disorder and which markers are associated for a person without. LIWC in combination with the classification algorithm random forest performed best in predicting if a person has a mental disorder (accuracy: 0.952; Cohen's kappa: 0.889) and spaCy in combination with random forest predicted best which mental disorder a person has (accuracy: 0.429; Cohen's kappa: 0.304). For the qualitative comparison, stop words were removed in order to investigate the influence on the accuracy and the Mann-Whitney U test and LIME are applied to gain further insight into the language markers. The language markers found for people with a mental disorder were 1SG PRONOUN and FOCUSPAST using LIWC. The spaCy variables were based on lemmas and their dependencies. The spaCy language markers found for people with a mental disorder were IK_DOEN_NSUBJ, IK_GAAN_NSUBJ, IK_HEBBEN_NSUBJ and IK_KOMEN_NSUBJ. The results show that the models which focus on the stylistic difference between words, scored best. The deep learning models were in this research poorly able to find the right cues to label an interview transcription correctly.

Keywords: Language marker · NLP · LIWC · spaCy · RobBERT · fastText · LIME

Acknowledgements

At the start of the Corona quarantine (March 2020), I started with my thesis. I researched language markers and what value it could add to diagnose a person. Furthermore, I looked into different kinds of NLP-techniques to find language markers. I would like to thank the following people: First, I would like to thank Marco Spruit, Kees de Schepper and Floortje Scheepers, for their guidance throughout these difficult times these past months. You all helped me with your enthusiasm, helpful ideas and insights. Second, I would like to thank the Verhalenbank and the people who were brave enough to share their stories about their mental health, without them I would not have been able to do my research. Thirdly, I would like to thank Pablo Romero for his help with the high performing computing cluster and his deep learning tips. And last but certainly not least, I would like to thank my parents, brother and sister for their listening ear, feedback and moral support.

Contents

1		
1	Research plan	9
1.1	Case study: Verhalenbank Psychiatrics . . .	10
1.2	Research questions	11
1.3	Relevance of research	12
1.3.1	Scientific relevance	12
1.3.2	Societal relevance	12
2	Research method	13
2.1	Design Science Framework	13
2.2	CRISP-DM	14
2.3	Literature research protocol	16
3	Theoretical Background	18
3.1	Findings from language markers research . .	18
3.1.1	ADHD	18
3.1.2	Autism	18
3.1.3	Bipolar disorder	18
3.1.4	Borderline Personality Disorder . .	19
3.1.5	Eating disorder	19
3.1.6	GAD	20
3.1.7	MDD	20
3.1.8	OCD	20
3.1.9	PTSD	20
3.1.10	Schizophrenia	21
3.1.11	Summary of disorders	22
3.2	Language Inquiry and Word Count	23
3.3	Dependency parsers	23
3.4	Neural approaches	24
3.4.1	Neural networks	24
3.4.2	Word2Vec	26

	3.4.3	fastText	26
	3.4.4	ELMo	27
	3.4.5	ULMFiT	28
	3.4.6	GPT	29
	3.4.7	BERT	30
	3.4.8	XLnet	32
	3.4.9	T5	33
	3.4.10	Neural network approach conclusion	33
4		Data analysis	35
	4.1	Data description	35
	4.2	Data preparation	36
	4.3	Descriptive statistics	36
	4.4	Quantitative analysis	37
		4.4.1 Traditional results	38
		4.4.2 Deep learning results	39
		4.4.3 Summary of quantitative results . .	40
	4.5	Feature importance	42
		4.5.1 Traditional results	43
		4.5.2 Deep learning results	44
		4.5.3 Summary feature importance results	49
	4.6	Results discussed by domain experts	49
5		Discussion	51
	5.1	Interpretation and implications of results . .	51
	5.2	Limitations	51
	5.3	Future work	52
6		Conclusion	53
	6.1	Sub-questions	53
	6.2	Main research question	54
Appendix A		LIWC	65
Appendix B		spaCy	71
Appendix C		LIWC categories	76
Appendix D		Draft Paper	79

List of Figures

2.1	Engineering Cycle (Wieringa, 2014)	13
2.2	CRISP-DM process model for data mining (Wirth & Hipp, 2000)	15
3.1	Trade-off speed accuracy of dependency parsers (Choi et al., 2015)	25
3.2	CBOW and skip-gram model (Mikolov, Chen et al., 2013)	27
3.3	BiLSTM-based ELMo (Hagiwara, 2018)	28
3.4	BERT pretraining and fine-tuning procedures (Devlin et al., 2018)	31
3.5	An example of pretraining and finetuning the T5	33
4.1	Number of people per mental disorder	37
4.2	Amount of words per mental disorder	38
4.3	LIWC decision tree	43
4.4	LIME explanation quote 1	46
4.5	LIME explanation quote 2	47
A.1	LIWC output part 1	65
A.2	LIWC output part 2	66
A.3	LIWC output part 3	67
A.4	LIWC feature importance top 10 for the binary classification	68
A.5	LIWC summary top 10 features	69
A.6	LIWC Mann-Whitney U-test top 10 features	70
B.1	Frequency of the spaCy variables	72
B.2	spaCy feature importance top 10 for the binary classification	73
B.3	spaCy summary top 10 features	74
B.4	spaCy Mann-Whitney U-test top 10 features	75

List of Tables

1	Language markers per disorder	22
2	Neural networks under consideration	35
3	Predictions	42
4	LIME output for fastText and RobBERT	48
5	Language markers for LIWC and spaCy	49
6	Examples of spaCy variables	71

List of acronyms and abbreviations

ADHD	Attention Deficit Hyperactivity Disorder
AR	Auto-Regressive
BERT	Bidirectional Encoder Representations from Transformers
BPD	Borderline Personality Disorder
CBOW	Continuous Bag Of Words
DSM-5	Diagnostic and Statistical Manual of Mental Disorders
ELMo	Embeddings from Language Models
GAD	Generalised Anxiety Disorder
GPT	Generative Pre-Training
LIME	Local Interpretable Model-agnostic Explanation
LSA	Latent Semantic Analysis
LSTM	Long Short Term Memory
MDD	Major Depressive Disorder
MLM	Masked Language Model
NN	Neural Networks
NLP	Natural Language Processing
OOV	Out-Of-Vocabulary
RNN	Recurrent Neural Network
RoBERTa	Robustly optimized BERT approach
SVM	Support Vector Machine
ULMFit	Universal Language Model Fine-tuning

1 Research plan

The contribution of mental disorders are a major part of the global burden of disease (Whiteford et al., 2013) and in 2017 accounted for 10.7% of the global population (Ritchie & Roser, 2020). This contribution is not keeping an even position, but is rising mainly in developing countries (Whiteford et al., 2013). Furthermore, mental disorders have a substantial long term impact on individuals, caregivers and society (McIntosh et al., 2016). The challenge of diagnosing a mental disorder is the complexity of the multiple genetic, environmental and psychological contributors and the individual risk factors (Russ et al., 2019).

Data science can offer opportunities to diagnose patients and to predict outcomes, patients preferences and treatment reactions (McIntosh et al., 2016). Combining data science with human language is called Natural Language Processing (NLP) and it is increasingly being used on electronic health records (EHRs) to study mental health (Perera et al., 2016). It can for example be used to identify symptoms, treatments and health trajectories (Russ et al., 2019). Furthermore, research has also shown that people with mental health difficulties use distinctive linguistic patterns (Lyons et al., 2018). An increased use of the first singular pronoun (I) is a sign of being more self-focused and is a language marker for mental distress (Lyons et al., 2018).

The research thus far on language markers of mental disorders has focused mainly on the Language Inquiry and Word Count (LIWC) software program (Calvo et al., 2017). This technique calculates the number of words of certain categories that are used in a text based on a dictionary (Pennebaker et al., 2001). LIWC is a traditional technique that analyses at word level. Traditional means that it analyses symbolically at word level and does not use neural networks. The goal of this research is to compare LIWC with other NLP techniques to provide more useful insights into psychiatric stories.

One technique is dependency parsing. This technique can provide insights because it will show the grammatical structure of the sentences and it will provide information about the relationships between words (Davcheva, 2018). For example, Alpino is a Dutch dependency parser used for linguistic exploration, training and evaluating pur-

poses (Van der Beek et al., 2002). By using this technique, the different uses of grammar between mental illnesses are uncovered. This will give further insight into the relationships between sentences of psychiatric stories.

Another technique is deep learning and this technique is chosen because of the complexity of mental disorders. Deep learning is defined by Deng, Yu et al. (2014) as: “a class of machine learning techniques that exploit many layers of non-linear information processing for supervised or unsupervised feature extraction and transformation, and for pattern analysis and classification”. This is why deep learning techniques seem suitable for these complex disorders.

By comparing the results and predictions of the traditional and neural approaches on word and sentence level, between the different mental disorders, differences in language will come to light. These differences in language use, also called language markers are a collective name to distinguish individual difference variables (Addawood et al., 2019) (Groom & Pennebaker, 2002). The objective of this research is to find out to what extent a mental disorder can be diagnosed.

1.1 Case study: Verhalenbank Psychiatrics

The psychiatric department of University Medical Centre Utrecht (UMCU) collects stories of people who have, had or were in contact with psychiatric complaints. Interviews about mental illness are conducted with (ex-)patients, caregivers and medical employees to gain new leads which could benefit the recovery of patients. The interviews are then rewritten into anonymous stories and put on the website of the Verhalenbank.

This research will help to get a better understanding of what insights a computer can gain from a text which a human is less able to do. The data is in Dutch, which makes it slightly more difficult since most models are trained on English data ¹. There are however NLP techniques especially adapted for Dutch texts.

¹ <https://www.groundai.com/project/robbert-a-dutch-roberta-based-language-model/1>

1.2 Research questions

The main research question in this research is defined as: *To what extent can a diagnosis of mental illness be determined by language markers?* This question is split up in five sub-research questions shown below:

SQ1 Which traditional sentence level approach is suitable to detect language markers?

This question is answered by giving an overview of parsers and a comparison between them.

SQ2 What neural approach is suitable to detect language markers at word level?

To answer this question, first an overview of the different approaches is provided. Next, the approaches are compared and an appropriate approach is chosen.

SQ3 What neural approach is suitable to detect language markers at sentence level?

This question is answered by giving an overview of neural approaches and a comparison between the approaches. Next, the model which performs best on classification tasks, is chosen.

SQ4 How well do the techniques perform on Dutch narratives?

To answer this question, the predictions between the different NLP models and classification algorithms will be compared.

SQ5 To what extent can we identify meaningful language markers for having a mental disorder

The output of LIWC and the dependency parser will be analysed to find significant difference between people with and without a mental disorder. LIME (Local Interpretable Model-agnostic Explanation) will be applied for the deep learning models. LIME is an explanation technique that learns a model locally around the

predictions and explains any classifier's predictions in an interpretable way (Ribeiro et al., 2016).

1.3 Relevance of research

The relevance of this research will be explained in the next section. It will be discussed from the scientific and societal perspective.

1.3.1 Scientific relevance

Comparing a lexical semantic approach, a compositional approach and a stochastic/neural approach will benefit the research so far on language markers in the mental health domain. This is because the research was mainly limited to the LIWC technique (Calvo et al., 2017). By uncovering subtle differences in the use of language with different techniques, identification of psychiatric problems will become easier.

1.3.2 Societal relevance

As said before, mental disorders accounted for 10,7% of the global population in 2017 (Ritchie & Roser, 2020). In 2017 mental health-care costs amounted for 22 billion euro in the Netherlands ². Furthermore, it adds more economic costs than chronic somatic diseases, such as cancer (Trautmann et al., 2016). Economic costs are not only the costs of treatment, hospitalisation and diagnostics, but also the invisible costs, such as early retirement and work absence (Trautmann et al., 2016). The mental health care expenditure will have increased fivefold in the Netherlands in 2060³. This research can help discover early on mental problems based on language markers and prevent more negative health outcomes.

² <https://www.nrc.nl/nieuws/2017/04/12/studie-psychische-klachten-kosten-22-mld-euro-8148765-a1554297>

³ <https://www.rivm.nl/nieuws/zorguitgaven-blijven-tot-2060-stijgen-gemiddeld-met-28-procent-per-jaar>

2 Research method

This section explains which combination of research methods is going to be used. The first two parts will include an explanation of the frameworks used: Design Science Research framework and CRISP-DM. In the third section, the literature research protocol will be explained. Lastly, the key milestones of this research will be laid out.

2.1 Design Science Framework

The Design Science Research (DSR) framework is defined by Wieringa (2014) as: "the design and investigation of artefacts in context". This framework distinguishes between knowledge questions and design goals (Wieringa, 2014). Design goals require an analysis of the stakeholder's goals and it calls for a real world change. Furthermore, there can be many solutions and there is not one single best. Knowledge questions ask for the world as is and do not call for change. These questions have one answer each and are evaluated by truth (Wieringa, 2014).

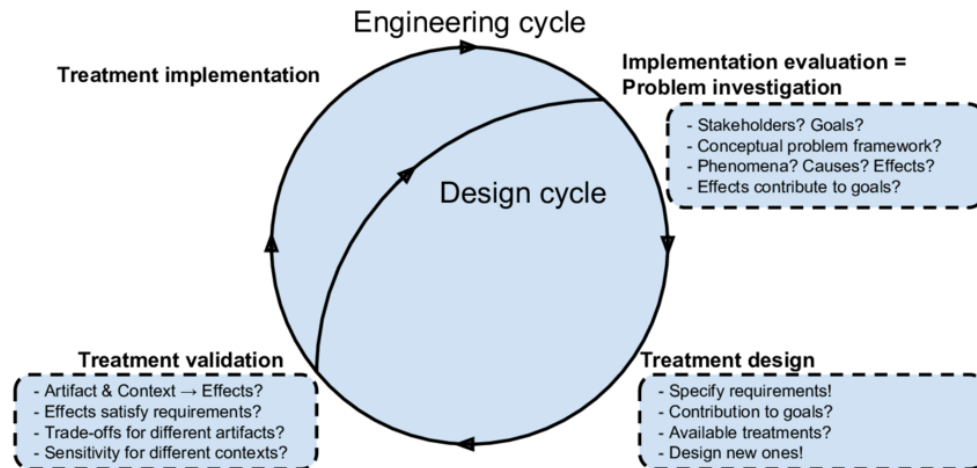


Figure 2.1: Engineering Cycle (Wieringa, 2014)

The research activities are broken down into a set of tasks to achieve the research goal. These tasks are according to a subset of the engineering cycle, which is called the design cycle. The engineering and design cycle are shown in 2.1.

The first task of this framework is the problem investigation. Herein, stakeholders, goals, problems and effects are investigated. In this research, the problem investigation is performed by executing an literature review to get an understanding of the problem. This activity is performed to find the traditional and neural approaches per word and sentence level for analysing the interviews. Furthermore, the LIWC-tool and the different mental disorders are explained. The research protocol is shown in section 2.3.

Next, the treatment is designed. Herein, different approaches are compared to find the best suitable techniques and answer sub questions one, two and three. The aim of this task is to design a model to detect the language markers of a text and give a mental health state. Another activity in this task, is the data analysis. This is performed by following the CRISP-DM cycle (Chapman et al., 1999). This framework will be explained in the next section. The findings from the analysis will result in the model that detects language markers and predicts based on the markers, the mental disorder of the author of the text input. The treatment validation task is to evaluate and validate the model (Wieringa, 2014). In this research, the model is validated with the test dataset and the accuracy is calculated this to answer sub question four. Furthermore, the model is reviewed by experts with knowledge of mental illnesses and sub question five is answered.

2.2 CRISP-DM

The data mining part of this research will be conducted by using the cross-industry standard process for data mining (CRISP-DM) method, which was developed in 1996 (Bosnjak et al., 2009) and belongs to the Knowledge Discovery Process framework (KDP) (Spruit & Lytras, 2018). The main goal of this model is to provide a more structured way in executing data mining processes and to acquire results more efficiently and accurately (Shearer, 2000). One import-

ant characteristic of this method is that it can operate independently from a specific industry or selected tools (Wirth & Hipp, 2000).

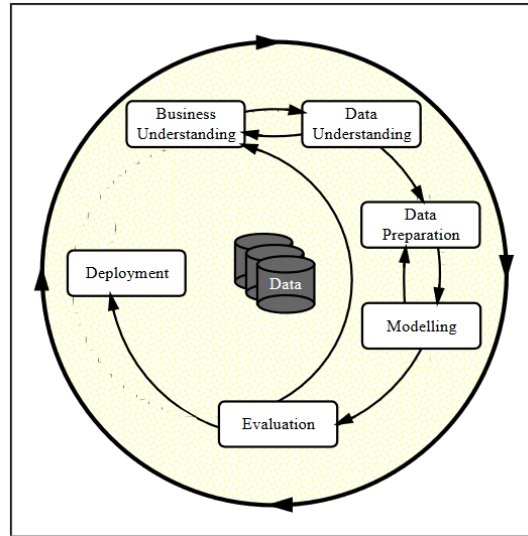


Figure 2.2: CRISP-DM process model for data mining (Wirth & Hipp, 2000)

The method consists of six phases as can be seen in figure 2.2. In the next paragraphs, all the phases will be briefly explained, based on the CRISP-DM 1.0 step-by-step data mining guide (Chapman et al., 1999). The first phase, the business understanding, consists of understanding the projects objectives and requirements from the chosen business. In the same phase these objectives and requirements are transformed into a data mining problem paired with a pre-liminary action plan. The data understanding phase consists of the initial data collection, description, exploration and quality verification. In the third phase, the data is preprocessed. This means that the interview transcriptions are cleaned and constructed to prepare the final data, so that it can be used in the next phases. Next, in the modelling phase, the different NLP modelling techniques were selected and applied to the data. Usually, it will go back and forth between the data preparation and the modelling phase, because some modelling techniques need the data in a specific format to process is correctly.

The fifth phase, the evaluation phase consists of assessing whether the results of the modelling phase are correct and correspond to the previously determined problem. Not only the results, but also the entire process must be evaluated to ensure that all business objectives have been achieved. If not, then one should return to earlier phases to retrieve missing elements. And in the last phase, the deployment phase, the models are deployed. However, this phase is out of the scope of this research.

2.3 Literature research protocol

A literature study is performed to come to an understanding of the status of the research area and to answer the first two research questions. The snowballing method was used to find the relevant literature. Backward snowballing is collecting relevant literature based on the reference list of a paper and forward snowballing is identifying new papers based on papers citing the paper being examined (Webster & Watson, 2002). Recent papers about language markers in mental healthcare were used as starting points. After that, one or two levels deep were snowballed back and forth. The amount of levels snowballed will depend on if new relevant literature is found, if a dead end is reached the snowballing will be stopped. A systematic literature study was not executed, because the results do not significantly differ from the snowballing approach, there were time constraints and the topic is innovative (Wohlin, 2014).

Google Scholar with a proxy from Utrecht University was used to execute the search queries. The search queries used in this research:

- "Language marker" "mental health" "LIWC"
- "Language marker" "mental health" "language use"
- "Mental health" "deep learning"
- "Dutch" "parser" "NLP"
- "BERT" "mental health" "classification"
- "Alpino" "dependency parser"
- "spaCy" "lemma" "dependency parser"
- "Language" in conjunction with the words below:
 - ADHD
 - Autism
 - Bipolar Disorder

- Borderline personality disorder
- Eating disorder
- Generalised anxiety disorder
- Major depressive disorder
- OCD
- PTSD
- Schizophrenia

The following papers were starting points of the literature study:

- Calvo, R. A., Milne, D. N., Hussain, M. S. & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23 (5), 649-685.
- Coppersmith, G., Dredze, M., Harman, C. & Hollingshead, K. (2015). From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses, In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*.
- Lyons, M., Aksayli, N. D. & Brewer, G. (2018). Mental distress and language use: Linguistic analysis of discussion forum posts. *Computers in Human Behavior*, 87, 207-211.
- Tausczik, Y. R. & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29 (1), 24-54.
- Young, T., Hazarika, D., Poria, S. & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3), 55-75.

3 Theoretical Background

This chapter describes the theoretical background of this research. First, ten mental disorders and their known language markers are explained in chapter 3.1. Next is a section about LIWC. In 3.3 are the different dependency parsers explained. At last in 3.4, the different neural approaches at sentence and word level are explained.

3.1 Findings from language markers research

This research is not the first looking into language markers of people with a mental disorder. For example Coppersmith et al. (2015) and Lyons et al. (2018) compared mental disorders with the LIWC-tool. The next sections will describe ten different mental disorders, what they are and the highlights of their use of language. This includes mainly their pronoun use, semantic coherence (SC) and word use. After these sections, a summary in the form of a table will be shown.

3.1.1 ADHD

Attention Deficit Hyperactivity Disorder (ADHD) is characterised by hyperactivity, in-attention and impulsivity (Polanczyk et al., 2007). They use more the third person plural (3pl) pronoun and less words of relativity than the control group in the research of Coppersmith et al. (2015). Furthermore, they use more sentences, but less clauses per sentence (Kim et al., 2015).

3.1.2 Autism

Autism is characterised by deficits in social interaction, abnormal language use and ritualistic and compulsive behaviour (Prizant, 1983). People with autism are more self-focused and they use more 1sg pronouns (Nguyen et al., 2013). They also use more words related to motion, home, religion and death in comparison to the control group of the research of Nguyen et al. (2013).

3.1.3 Bipolar disorder

In bipolar disorder, an individual is suffering from manic or hypomanic and depressive episodes (Forgeard, 2008). Manic symptoms

consists of decreased need for sleep, abnormal elevated moods and possible psychotic episodes (Forgeard, 2008). The hypomanic symptoms are milder occurrences of the manic symptoms and are of shorter duration (Forgeard, 2008). People with bipolar disorder use more words related to death and use more first person singular pronouns (Forgeard, 2008).

3.1.4 Borderline Personality Disorder

Borderline personality disorder (BPD) is marked by instability in impulse control, relationships, self-image and affect control (Lieb et al., 2004). Furthermore, clinical signs are impulsive aggression, repeated self-harm, emotional dysregulation and recurrent suicide tendencies (Lieb et al., 2004). People with BPD have a higher use in the pronouns 'he', 'she', 'hers' and 'his', which is the third person singular (3sg) pronoun (Lyons et al., 2018). This could be because they are more sensitive to rejection than patients with other mental disorders and they think that, in their present life, early experiences are still relevant (Lyons et al., 2018). They also curse more, use more past tense words and use less cognitive process words, such as should, maybe and know (Carter & Grenyer, 2012). Furthermore, people with BPD use less complex sentences, but do not score different on semantic complexity (Carter & Grenyer, 2012).

3.1.5 Eating disorder

Eating disorders, consisting of bulimia, anorexia and eating disorders not otherwise specified, are characterised by an obsession with food, body weight or body shape (Fairburn et al., 1993). This results in different symptoms per disorder, such as overeating, vomiting, not eating, limiting calories or setting dietary rules (Fairburn et al., 1993). Language wise, people with eating disorders are self-focused, talk in the present, use a lot of negative emotion words and use a lot of words for cognitive processes in comparison to the control group of Coppersmith et al. (2015). Interesting, but foreseeable word use, are words related to the body (Coppersmith et al., 2015).

3.1.6 GAD

GAD or generalised anxiety disorder consists of intense and extreme worry over a longer period of time (Evans et al., 2008). Furthermore, it is a chronic disorder and often combined with other chronic disorders (Evans et al., 2008). People with GAD use a lot of impersonal pronouns, such as *it* and *those* (Coppersmith et al., 2015). They also use a lot of negative emotion words, especially anxiety words and they use a lot of tentative words, such as *maybe* and *perhaps*. Furthermore, words related to health and death are also used a lot in comparison to the control group of Coppersmith et al. (2015). According to the research of Remmers and Zander (2018), people with anxiety experience impaired use of semantic coherence.

3.1.7 MDD

Next, Major Depressive Disorder (MDD) is a disorder that is characterised by abnormalities in mood, affect, cognition, agitation, retardation, sleep and appetite (Fava & Kendler, 2000). MDD results in people using the first person (1sg) pronoun more and being more self-focused (Trifu et al., 2017). They also use more past tense, have lexical and semantic repetitions and use short, detached and arid sentences (Trifu et al., 2017). Furthermore, the absence of positive thinking and negative biased thinking of people is also a characteristic of people with MDD (Trifu et al., 2017).

3.1.8 OCD

Obsessive compulsive disorder (OCD) is characterised by obsessive thoughts and repetitive purposeful behaviour performed according to a set of rules (March & Mullen, 1998). People with OCD are self-focused and use a lot of words used for cognitive processes, such as *think*, *should*, *never* or *else* (Coppersmith et al., 2015). Furthermore, they use more anxiety words according to Lyons et al. (2018).

3.1.9 PTSD

PTSD, short for post-traumatic stress disorder, follows overwhelming stressful events (Cameron & Gusman, 2003). This could be for example natural disasters, sexual assault or exposure to war (Cameron

& Gusman, 2003). It results in nightmares, avoidance of lookalike situations, feeling numb, being on guard and being watchful and being easily scared (Cameron & Gusman, 2003). The research of Coppersmith et al. (2015) concluded that people with PTSD do not use language extremely different than the control group. However, Papini et al. (2015) concluded that they did use more words related to death, more singular pronouns, less plural pronouns and lower use of cognitive words.

3.1.10 Schizophrenia

Schizophrenia is a disorder with unknown aetiology and it is observed by signs of psychosis (Insel, 2010). Commonly, schizophrenia comes with auditory hallucinations and paranoid delusions (Insel, 2010). It presents itself in early adulthood or adolescence. Research shows that schizophrenia results in language impairment at the semantic/discourse cohesion or coherence (Corcoran & Cecchi, 2020) (Zimmerer et al., 2017). This means that the flow of consistency of references and meaning across sentences are impaired (Corcoran & Cecchi, 2020). Schizophrenia also results in disruption and lack of grammatical structure (Zimmerer et al., 2017). Furthermore, schizophrenia has, compared to other mental disorders, the highest use of 'they' or 'them' also called third person plural (3pl) pronouns (Lyons et al., 2018) (Fineberg et al., 2015). This could be a result of the paranoia (Lyons et al., 2018). Another interesting finding is that people with schizophrenia use more words related to religion and less words related to body and ingestion (Fineberg et al., 2015).

3.1.11 Summary of disorders

Table 1 below shows per mental disorder, the known language markers.

Disorder	Pronoun	SC	Word use	Other
ADHD	3pl	-	-	Relativity, more sentences, less clauses
Autism	1sg	-	Motion, home, religion and death	-
Bipolar	1sg	-	Death	-
BPD	3sg	Normal	Death	Swearing, less cognitive emotion words
Eating	1sg	-	Body	Negative emotion words
GAD	imprs	Impaired	Death and health	Tentative words
MDD	1sg	Impaired	-	Inverse word-order and repetitions
OCD	1sg	-	Anxiety	More cognitive words
PTSD	sg	-	Death	Less cognitive words
Schizophrenia	3pl	Impaired	Religion and Death	Hearing voices and sounds

Table 1: Language markers per disorder

3.2 Language Inquiry and Word Count

As said before, research so far on exploring language markers in mental health has been done mainly with LIWC (Calvo et al., 2017). LIWC is a computerised text-analysis tool and has two central features: the processing component and the dictionaries (Tausczik & Pennebaker, 2010). The processing feature is the program which analyses text files and goes through them word by word. Each word is compared with the dictionaries and then put in the right categories (Tausczik & Pennebaker, 2010). For example, the word "had" can be put in the categories verbs, auxiliary verbs and past tense verbs. Next, the program calculates the percentage of each category in the texts, for example 17% of the words were verbs. A disadvantage of the LIWC program, is that it ignores context, idioms, sarcasm and irony (Tausczik & Pennebaker, 2010). Furthermore, the 89 different categories are based on language research. However, this does not mean that these categories represent reality, because categories could be missing.

3.3 Dependency parsers

The syntactic processing of texts is called dependency parsing (Choi et al., 2015). This processing is valuable because it forms transparent lexicalized representations and it is robust (Choi et al., 2015). Furthermore, it also gives insight into the compositional semantics. Compositional semantics is about the meaning of linguistic sentences and it is composed by the meaning of the individual words or phrases it contains (Hermann, 2014). Small changes in the syntactic structure of a sentence can change the whole meaning of the sentence. For example, John hit Mary or Mary hit John contain the same words, but have different meaning. It is said that compositionality is linked to our ability to interpret and produce new remarks, because once one has mastered the syntax of the language, lexical meanings and modes of composition, one can interpret new combinations of words (Liang & Potts, 2015).

Compositionality is the semantic relationship combined with a syntactic structure (Guevara, 2010). Compositional semantics is driven by syntactic dependencies and each dependency forms, from the con-

textualised sense of the two related lemmas, two new compositional vectors (Gamallo, 2017). So, the technique required for extracting the compositional semantics, needs to contain a dependency parser and a lemmatizer.

Choi et al., (2015) compared the ten leading dependency parsers based on the speed/accuracy trade-off. As can be seen in figure 3.1, is that Mate (Bohnet, 2010), RBG (Lei et al., 2014) and ClearNLP (Choi & McCallum, 2013) score the best of the ten on the unlabeled attachment score (UAS). The only problem is, they all do not include a Dutch dictionary needed for this research. spaCy does include a Dutch dictionary and according to their website, they improved their model and it reached a higher accuracy than the other parsers on the same dataset.

Other Dutch dependency parsers are Frog (Bosch et al., 2007) and Alpino (Van der Beek et al., 2002). Both Frog⁴ and spaCy⁵ includes the Dutch dictionary corpus of Alpino.

So, there are three options to parse a text: Alpino, Frog and spaCy. All three use Alpino in some way, but because of equipment constraints, spaCy is chosen.

3.4 Neural approaches

This section starts with an introduction of neural networks, how it all started and what the current developments are. After that, an exhaustive list of neural networks is discussed. An explanation will be given per model and in the end the most suitable model is chosen for this research.

3.4.1 Neural networks

The use of neural networks in NLP started around the early 2000s and were really making a difference in the 2010s, because of the growing power of computer systems (Otter et al., 2020). Features made for traditional NLP systems, were frequently handcrafted, time consuming and incomplete (Otter et al., 2020). Neural networks on the other hand can automatically, based on dense vector representations,

⁴ <https://github.com/LanguageMachines/frog/releases/>

⁵ <https://spacy.io/models/nl>

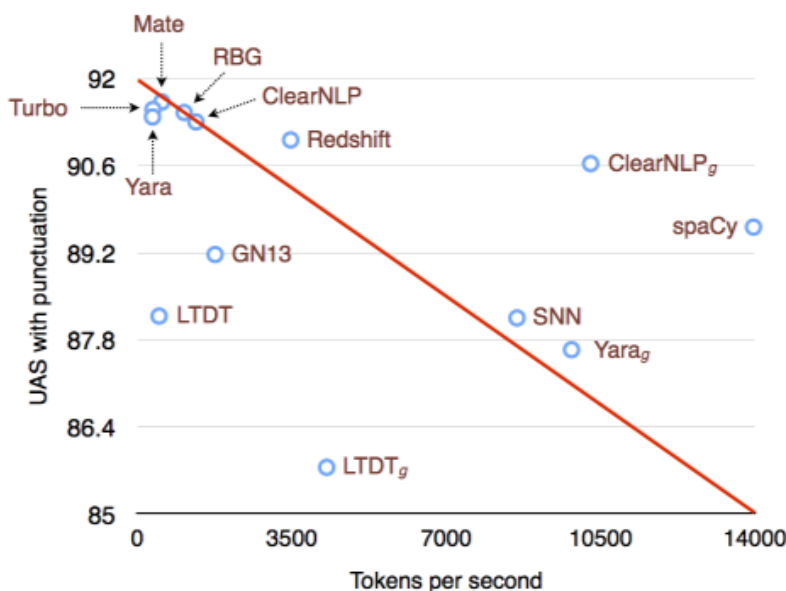


Figure 3.1: Trade-off speed accuracy of dependency parsers (Choi et al., 2015)

learn multilevel features and give better results (Young et al., 2018). The trend of neural networks has been caused by the success in deep learning and word embeddings (Young et al., 2018).

Recurrent neural networks (RNN) capture the context of a network and are tailor-made for modelling context dependencies (Young et al., 2018). With context is meant how one word is related to other surrounding words in a sentence. A RNN also allows the looping of hidden layers back to themselves which results in the RNN being able to handle variable length sequences input (Lopez & Kalita, 2017). Furthermore, RNNs are able to process data where time and order are of value, because it can loop over the input and see what has come previously (Lopez & Kalita, 2017). An improved type of RNN is the LSTM (Long Short Term Memory). The LSTM captures the long-distance context and goes from left to right (Lopez & Kalita, 2017). The improved version of LSTM is called BiLSTM (bidirectional LSTM) and this RNN looks in forward and backward

directions at word sequences (Honnibal, 2016). This can be seen in figure 3.3 where ELMo is trained based on BiLSTM.

Another development in neural networks is transfer learning. It is defined as transferring knowledge from a related already learned task to a new task (Torrey & Shavlik, 2010). This results in a better performance of the target task and it is also less expensive to (re-) collect data (Pan & Yang, 2009).

In 2017, the transformer architecture was introduced (Vaswani et al., 2017). It allows parallel inputs and it improves the pretraining of large text corpora. This leads to major gains for several tasks, such as text classification and language understanding (Wolf et al., 2019). The transformer-XL is an improved version of the transformer and was published in 2019 (Dai et al., 2019). This transformer can process longer sequences of inputs simultaneously without disrupting temporal coherence (Dai et al., 2019).

3.4.2 Word2Vec

Word embeddings, such as the skip-gram model and the continuous bag-of-words (CBOW) model (Mikolov, Sutskever et al., 2013), distribute high quality vector representations and are often used in deep learning models as the first data processing layer (Young et al., 2018). The difference between the two word embedding techniques is that the skip-gram model predicts the surrounding context words given the target word and the CBOW model calculates the conditional probability of the target word, given the surrounding context words (Young et al., 2018). This can be seen in figure 3.2. The Word2vec algorithm uses neural networks to learn these vector representations (Mikolov, Chen et al., 2013). It can use the skip-gram model or the CBOW model and it works for both small and large datasets (Mikolov, Chen et al., 2013).

3.4.3 fastText

Out-of-vocabulary (OOV) words, also known as unknown words, are a common issue for languages with large vocabularies (Young et al., 2018). fastText can overcome this problem by handling each word as a bag-of-character n-gram (Bojanowski et al., 2017). This is done by using the skip-gram model from Word2Vec as an extension.

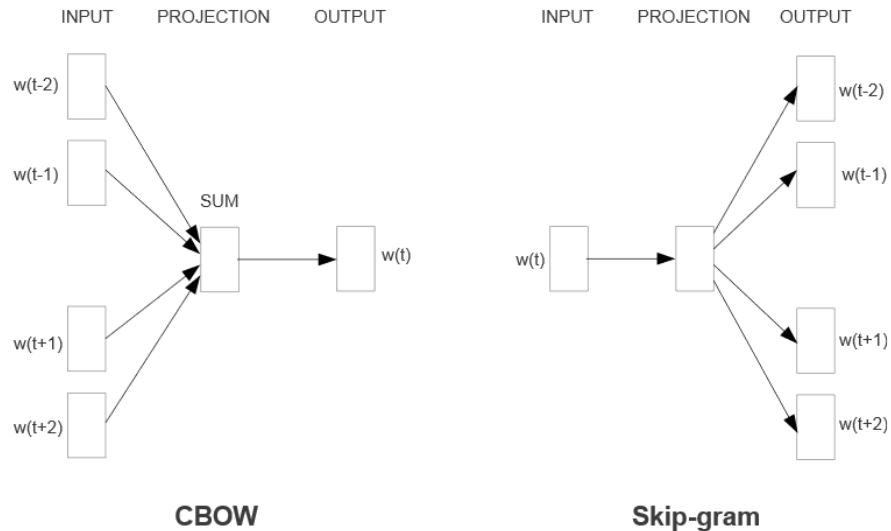


Figure 3.2: CBOW and skip-gram model (Mikolov, Chen et al., 2013)

These n-grams are used to represent the sum of the n-gram vectors (Bojanowski et al., 2017). Other techniques represent each word as a vector, but ignore the internal structure of a word. This limits morphologically rich languages, such as Finnish or Turkish, because these languages contain many word forms (Bojanowski et al., 2017). If a training corpus does not have certain word forms, it will be difficult to learn good word representations (Bojanowski et al., 2017). So by including character level information, the vector representations for morphologically rich languages will be improved (Bojanowski et al., 2017).

3.4.4 ELMo

Embeddings from Language Models (ELMo) is a word representation technique that models complex characteristics of word use (Peters et al., 2018). It reads the whole input sentence before assigning a contextualized word embedding to it (Peters et al., 2018). An example of ELMo can be seen in figure 3.3, where ELMo is trained based on the BiLSTM (Hagiwara, 2018). The figure shows a two-layer ELMo architecture (Hagiwara, 2018). The more layers, the more context is learned from the input. The lower levels catch basic syntax and

grammar and the higher levels catch contextual semantics. The advantage of ELMo trained with BiLSTM is that it reads bidirectional, so from start to finish and vice versa (Hagiwara, 2018). This makes it able to catch the whole context of the words in a sentence (Hagiwara, 2018).

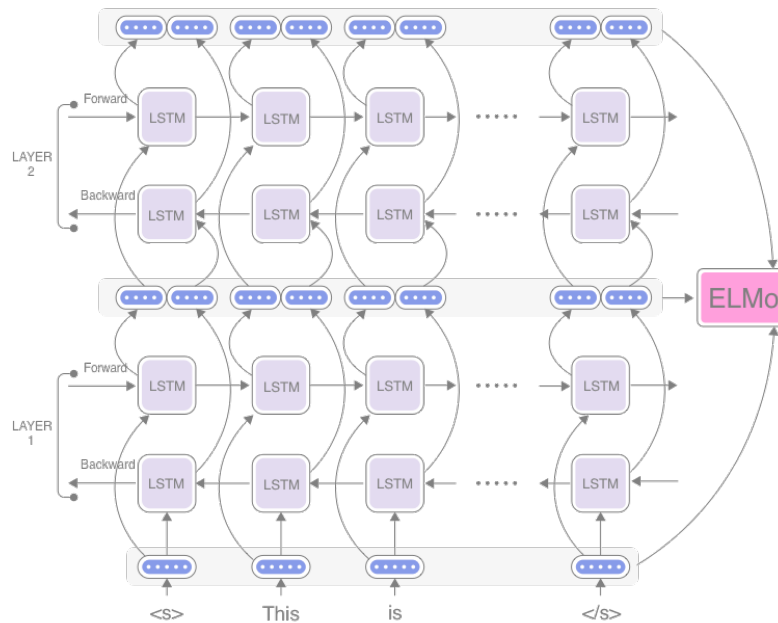


Figure 3.3: BiLSTM-based ELMo (Hagiwara, 2018)

3.4.5 ULMFiT

Universal Language Model Fine-tuning (ULMFiT) is a transfer learning method that can be applied to any task in NLP (Howard & Ruder, 2018). It uses the LSTM-architecture without additional hyperparameters (Howard & Ruder, 2018). Furthermore, ULMFiT also prevents overfitting by leveraging general domain pretraining and novel fine-tuning techniques (Howard & Ruder, 2018). The method consists of three stages: pretraining, fine-tuning and classifier fine-tuning (Howard & Ruder, 2018). The pretraining consists of the language model training on the general domain-corpus to catch the

general features and next this model is fine-tuned on the target task data to learn task-specific features. To fine-tune the classifier on the target task, the fine-tuned language model is used to get output distribution probabilities (Howard & Ruder, 2018).

3.4.6 GPT

Generative Pre-Training (GPT) is an approach which makes use of generative pretraining and discriminative fine-tuning for each specific task (Radford et al., 2018). To achieve effective transfers with minimal adaptations to the architecture of the model, task-aware input transformations are used (Radford et al., 2018). Furthermore, the Transformer (Vaswani et al., 2017) was used for handling long-term dependencies to obtain a more structured memory (Radford et al., 2018). This results in a robust transfer performance (Radford et al., 2018).

GPT-2 The Generative Pre-trained Transformer 2 (GPT-2) is an unsupervised transformer language model with over an order of magnitude more parameters than the GPT (Radford et al., 2019). The model was demonstrated in a zero-shot setting, which means that the model’s parameters or architecture is not altered after pretraining (Radford et al., 2019). This was done because researchers were arguing that language models do not generalise well and are only trained for a specific task (Radford et al., 2019).

GPT-3 The third GPT-model is the Generative Pre-trained Transformer 3 (GPT-3) (Brown et al., 2020). It is an auto-regressive language model, which is not fine-tuned and is without any gradient updates (Brown et al., 2020). The GPT-3 is so advanced that it is able to write samples of news articles, which are not recognised by human evaluators as computer-written texts (Brown et al., 2020). This is also why the paper includes a section about societal impacts of the GPT-3. Furthermore, this model is not available for the public yet, only a few selected people are given access to the private beta⁶.

⁶ <https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/>

3.4.7 BERT

BERT is a language representation model designed to pretrain, from unlabelled text, deep bidirectional representations (Devlin et al., 2018). This is reached by looking at both the left and right context in all the layers of the corpus and jointly conditioning this (Devlin et al., 2018). The pretrained BERT model can be fine-tuned with one extra output layer as a result (Devlin et al., 2018). BERT uses the masked language model (MLM) as pretraining objective. The MLM masks some of the tokens of the input at random and tries to predict the original words under the masks based on the context. This enables the representation to fuse the left and the right context of the masked word, with as result a deep bidirectional Transformer (Devlin et al., 2018). Additional, next sentence prediction is used to train text-pair representations. The framework of BERT consists of two parts: the pretraining and the fine-tuning (Devlin et al., 2018). The model is trained with different pretraining tasks on unlabelled data, during pretraining (Devlin et al., 2018). During fine-tuning, the model is first initialised with the pretrained parameters and after that, by using the labelled data from downstream tasks, the parameters are fine-tuned (Devlin et al., 2018). So, the pretrained BERT model gets for every different task different inputs and outputs to fine-tune the parameters and this can be seen in figure 3.4 (Devlin et al., 2018), where on the left, the model is pretrained with MLM and on the right, the model is fine-tuned based on the different tasks. The pink boxes show the tokens and masked tokens, the orange boxes show the embeddings in the blue neural network and the green boxes are the embeddings processed by the transformer.

RoBERTa An improved version of BERT is RoBERTa, which stands for Robustly optimized BERT approach (Liu et al., 2019). The main difference is the training of the model. The main changes are that RoBERTa trains longer, on more data, with bigger batches and on longer sequences (Liu et al., 2019). Furthermore, the next sentence prediction objective is removed from the training procedure and the masked pattern applied to the training data is changed dynamically (Liu et al., 2019).

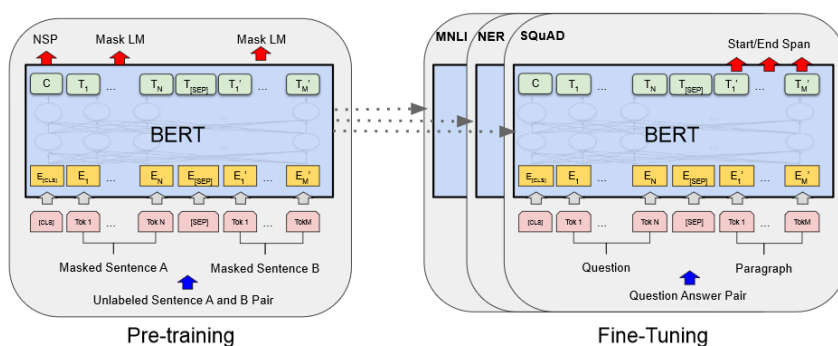


Figure 3.4: BERT pretraining and fine-tuning procedures (Devlin et al., 2018)

ClinicalBERT ClinicalBERT is another BERT version and analyses clinical notes and uncovers high-level relationships between medical concepts (Huang et al., 2019). These clinical notes consist of unstructured data and are high-dimensional and sparse (Huang et al., 2019). It processes the notes of patients and calculates the readmission risk score of a patient or whether he/she is back in 30 days (Huang et al., 2019). This helps with early intervention and better informed decisions. ClinicalBERT can also do more tasks, such as diagnose, predict risk of mortality and calculate the duration of stay in the hospital (Huang et al., 2019).

StructBERT Another extension of BERT is StructBERT (Wang et al., 2019). This language model incorporates language structures in the pretraining (Wang et al., 2019). It contains two additional tasks in the pretraining: leveraging sentence level ordering and leveraging word level ordering (Wang et al., 2019). This results in StructBERT being able to predict words and sentences in the right order (Wang et al., 2019).

ALBERT The lite version of BERT is called ALBERT (Lan et al., 2019). ALBERT was developed because of the increasing model sizes (Lan et al., 2019). The CPU/GPU limitations and longer training times were becoming a problem (Lan et al., 2019). Therefore, two new techniques were introduced to reduce the memory consump-

tion and improve training speed (Lan et al., 2019). One technique is the cross-layer parameter sharing and this prevents the parameter from growing into the network’s depth (Lan et al., 2019). The second technique is factorised embedding parameterization and consists of separating the vocabulary embeddings from the size of the hidden layers. This is established by decomposing the matrix of the vocabulary-embeddings into two small matrices (Lan et al., 2019). Furthermore, BERT’s limitations regarding inner-sentence coherence, is improved by introducing self-supervised loss for sentence-order prediction (Lan et al., 2019). Despite of ALBERT having less parameters than BERT-large, it reaches state-of-the-art results (Lan et al., 2019).

3.4.8 XLnet

XLnet is a generalised auto-regressive (AR) pretraining method (Yang et al., 2019). According to the researchers from Carnegie Mellon University, BERT is neglecting dependencies between masked tokens and it is also suffering from a discrepancy in the pretraining and finetuning of the model (Yang et al., 2019). XLnet overcomes this problem by using its AR language modelling, learning contexts bidirectional and integrating some ideas from the Transformer-XL into its architecture (Yang et al., 2019). AR language modelling estimates the probability distribution of a corpus with an AR model. The research of Sarhan and Spruit (2020) shows that XLnet outperforms BERT and XLM-RoBERTa when training and testing on the dataset. The AR language modelling is the main solution to the problem, because it does not hide tokens and is therefore able to model the joint probability using the product rule (Yang et al., 2019). Furthermore, the Transformer-XL ideas that are used are the segment recurrence mechanism and relative encoding scheme, which improves the performance, because it allows longer text sequences (Yang et al., 2019). So, because XLnet uses AR and bidirectional context, it maximizes the expected likelihood with respect to all permutations of the factorisation order (Yang et al., 2019). Furthermore, it also does not rely on corrupt data, because it does not use masks (Yang et al., 2019).

3.4.9 T5

Text-to-Text Transfer Transformer (T5) is a unified approach to transfer learning (Raffel et al., 2019). It provides a framework that transforms NLP-problems into text-to-text formats (Raffel et al., 2019). With text-to-text format is meant using text as input and produce as output text (Raffel et al., 2019). An example ⁷ can be seen in figure 3.5, where the T5 was pretrained with "knowledge" by learning to fill in dropped-out text. Next, the T5 is finetuned to answer questions without extra information. This results in the T5 answering questions by using information from pretraining. The T5 allows for any NLP task, to use the same model, hyperparameters and loss functions. For example, question answering, document summarization and classification (Raffel et al., 2019).

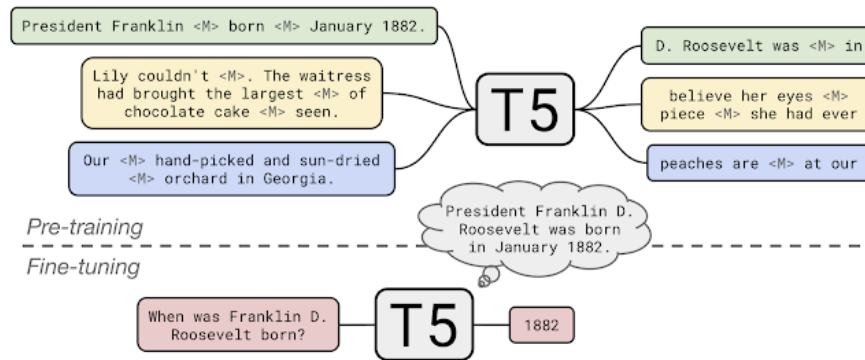


Figure 3.5: An example of pretraining and finetuning the T5

3.4.10 Neural network approach conclusion

In figure 2 an overview of the different neural networks can be seen. The choice of best fit is limited, because of the small and Dutch dataset. Two neural networks are chosen for this research, one based on words and one based on sentences. Furthermore, the neural networks need to have a Dutch model. So the choice is between Word2Vec and fastText at word-level and BERT, mBERT and RoBERTa at

⁷ <https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html>

sentence level. Other models, such as ClinicalBERT could also be used in combination with a transfer learning model such as XLM (Cross-lingual Language Model) to tackle the Dutch data. However, these models have not been used extensively in the medical domain (Khattak et al., 2019). This could be because the interpretability and performance of a model are equally important. Even though deep learning models can perform superior in comparison to the more traditional models, they are hard to explain or understand (Miotto et al., 2018). Hence, this approach is not used for this research. Furthermore, fastText has proven that it results in better performance in comparison to Word2vec (Le et al., 2019) and it is able to handle OOV words, because of the n-grams. The Dutch version of BERT is called BERTje (de Vries et al., 2019), the Dutch version of RoBERTa is called RobBERT (Delobelle et al., 2020) and mBERT is the multilingual BERT with more than 100 languages (Devlin et al., 2018). A choice between the three BERT’s is made by looking at their performance on the classification task, because that will be used in this research. The research of Delobelle et al. (2020) shows that RobBERT (ACC = 95.1%) performs best on the classification tasks compared to mBERT (ACC = 84.0%) and BERTje (ACC = 93.0%) with the full dataset. So, the neural networks chosen for this research are fastText and RobBERT.

Model	Dutch?	Architecture	Input level	Selected?
Word2Vec	Yes	CBOW & Skip-gram	Word	No
fastText	Yes	RNN	Word	Yes
ELMo	Yes	(Bi)LSTM	Sentence	No
ULMFit	Yes	Transformer	Sentence	No
GPT	No	Transformer	Sentence	No
GPT-2	No	Transformer	Sentence	No
GPT-3	No	Transformer	Sentence	No
BERT	Yes	Transformer	Sentence	No
RoBERTa	Yes	Transformer	Sentence	Yes
ClinicalBERT	No	Transformer	Sentence	No
XLnet	No	Transformer-XL	Sentence	No
StructBERT	No	Transformer	Sentence	No

ALBERT	No	Transformer	Sentence	No
T5	No	Transformer	Sentence	No

Table 2: Neural networks under consideration

4 Data analysis

The findings from the literature review form the answers for the first three subquestions. Subquestions four and five are answered by performing a data analysis according to the CRISP-DM method. It starts by the description and basic understanding of the data. Next, the data is prepared and after that it is modelled. The procedure of the data analysis and the results are shown in the next sections. Furthermore, the source code of the data analysis is available at: <https://github.com/StephanieVx/ExploringLinguisticMarkers>.

4.1 Data description

The data used for this research were interviews with (ex-)psychiatric patients, caretakers and medical employees. The interviews were held by a psychologist and executed with an unstructured approach. Each patient was asked to tell a story about themselves. For example, a story about their youth, about their family or about their current work. The available interviews with caregivers and medical employees were used to compare the people with mental health disorders to people without it. Furthermore, the interviews were recorded and transcribed by Transcriptie Online ⁸. There are in total 72 interviews from people with and 36 from people without a mental disorder taken from 2016 to 2020.

Important to know is that these interviews were not held in an acute phase. This means that the (ex-)psychiatric patients were not in a psychiatric crisis and were not a possible danger to themselves or others. This does not mean that they were cured at that moment. They were in their remission or recovery phase. Furthermore, people

⁸ <https://www.transcriptieonline.nl/>

with a mental disorder can successfully manage their disorder, but will almost never be "cured" (Wakefield, 1992). So, whichever phase the (ex-)psychiatric patients were when interviewed, they were influenced by their mental disorder.

Furthermore, the data is privacy sensitive. Therefore, names, places and health care institutions were replaced by generic labels, such as [name] or [place], from the interviews beforehand by Transcriptie Online.

4.2 Data preparation

Most of the data was already collected before the start of this research. The rest was collected during the research. After the collecting, the data needed to be prepared before analysing it. This was done in R (version 3.6.0) by removing the questions of the psychologists and removing the headers of the documents. So, the corpus only included the text the interviewee had said during the interview.

Furthermore, the data needed to be labelled. Every interview was read and a label was attached based on the diagnosis the patient told the psychologist. Two patients were not clear about their diagnosis, so another psychologist took a look at those two interviews and labelled them.

4.3 Descriptive statistics

The number of people per mental disorder in our dataset is shown in figure 4.1. As can be seen, the group with dissociation (a disconnection between a person's memories, feelings, perception, or sense of self ⁹) contains the least number of people in this dataset and the group with psychosis is the largest. Furthermore, there are two labels with personality. The difference between personality and personality+ is that it shows in different ways. Personality includes obsessive-compulsive personality disorder, avoidant personality disorder, dependent personality disorder and unspecified personality disorders. Personality+ contains in this research only BPD because

⁹ <https://www.psychiatry.org/patients-families/dissociative-disorders/what-are-dissociative-disorders>

of the data, but in the DSM-5¹⁰ (Diagnostic and Statistical Manual of Mental Disorders) it also contains anti-social personality disorder, histrionic personality disorder and narcissistic personality disorder. Figure 4.2 shows a boxplot of the amount of words per mental disorder. It can be seen that the people with eating disorders used less words than the other groups.

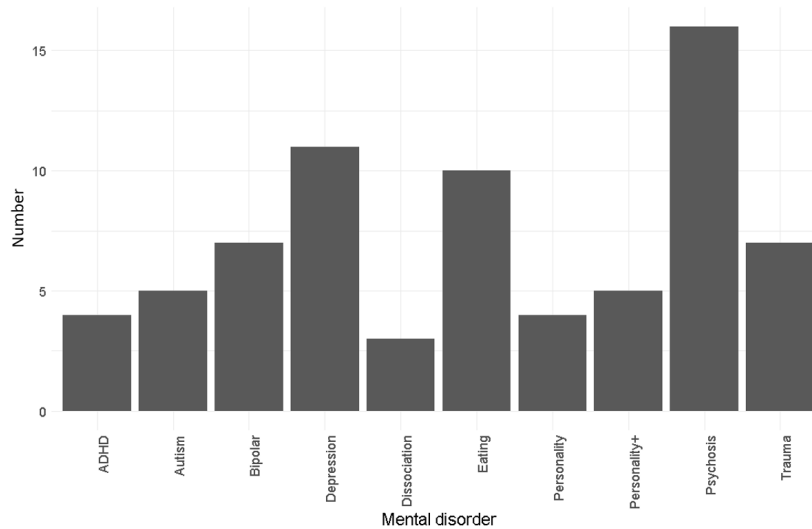


Figure 4.1: Number of people per mental disorder

4.4 Quantitative analysis

The next section consists of a comparison between mental disorder and no mental disorder and a comparison per mental disorder. The different models are applied to the data and with the results predictions are made. For the predictions, the data was randomly split in 80% training set and 20% test set. Every model scored on the testset and the scores are expressed by accuracy and Cohen's kappa. Accuracy instead of F1-score is used because accuracy is easier to interpret. Furthermore, it looks at correctly classified observations.

¹⁰ <https://www.ggzstandaarden.nl/zorgstandaarden/persoonlijkheidsstoornissen/specifieke-omschrijving-persoonlijkheidsstoornissen>

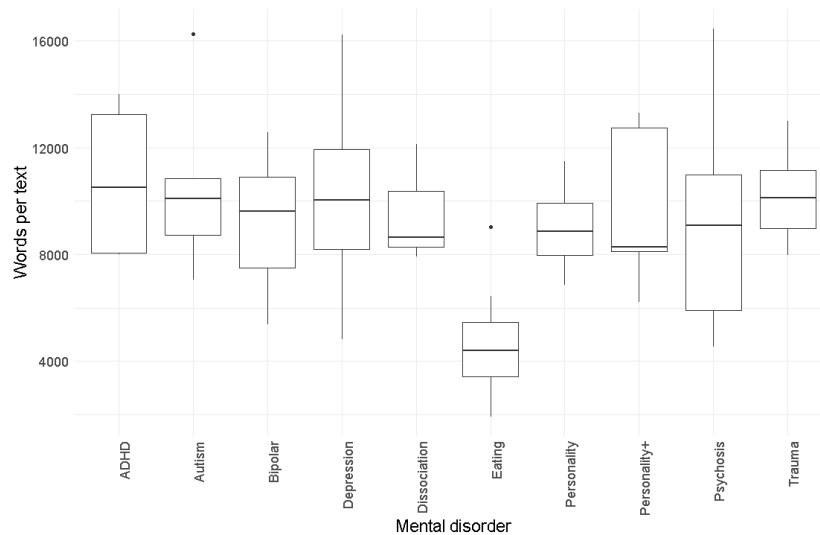


Figure 4.2: Amount of words per mental disorder

A disadvantage of accuracy is that it does not take an imbalanced class distribution into account. The distribution of the first comparison is 67% MD - 33% noMD. The accuracy threshold is set at 0.67 because from that point the model scores better than the odds. For the second comparison, the distribution is as shown in figure 4.1 and this results in a threshold of 0.16.

Cohen's kappa was also applied to every model because it calculates the improvement in accuracy over the sample and takes imbalance into consideration. This is done by assessing the inter-classifier agreement (Cohen, 1960). It takes the probability that the mental disorder vs no mental disorder or the 10 different kinds of mental disorders in the comparisons, agree by chance into consideration when quantifying how much they agree. So, it measures the reliability of the different models classifying the same transcription and is corrected by how often they classify the same by chance. The results are shown in table 3.

4.4.1 Traditional results

Every interview transcription went through LIWC and the output was used to predict whether a person has a mental disorder or not,

and if so which one. The compositional semantics of the interviews were analysed with spaCy. The lemmas and dependencies of every interview were collected with spaCy and bound together in syntactic n-grams. Examples of the different n-grams are shown in table 6 in appendix B. To reduce noise impact on the data, a feature selection of n-grams based on their frequency was used. The frequency of the spaCy variables are shown in B.1 in appendix B. As can be seen, a few thousand variables were collected. If an n-gram was used less than 5 times, it was not used in the analysis. Nevertheless, almost 1000 different n-grams were collected. For the prediction of the different labels, the algorithms used were decision tree, random forest and SVM (Support Vector Machine). The decision tree algorithm builds a classification model and this is represented in a tree. It is able to handle missing values, which could be an advantage for the data. The random forest algorithm builds a classification model by making multiple trees and by aggregating them together, preventing a negative influence of outliers on the outcome. This algorithm was used because it tends to not overfit on the data (Friedman et al., 2008). The third algorithm used was the SVM and this algorithm separates the classes based on the optimal hyperplane (Moraes et al., 2013). According to research, this algorithm performs with more accuracy than most other algorithms for text classification (Moraes et al., 2013).

4.4.2 Deep learning results

The deep learning models first needed to be fine-tuned, three epochs for RobBERT and 50 epochs for fastText. The amount of epochs for RobBERT was based on the associated scientific paper (Delobelle et al., 2020). The standard range of epochs for fastText is between 5 and 50 and 50 epochs was chosen because this was also done in the Huggingface transformers framework. The Simple Transformers library, which is used in this research, is based on this framework. Simple transformers is used, because a model can be fine-tuned, evaluated and tested with few lines of code¹¹. The deep learning model used for analysis on sentence-level was the pdelobelle/robbert-v2-

¹¹ <https://github.com/ThilinaRajapakse/simpletransformers>

dutch-base¹². After fine-tuning, the models were validated on the test-set.

4.4.3 Summary of quantitative results

Table 3 shows the accuracy of the two comparisons and Cohen’s Kappa per prediction. LIWC and spaCy score above the threshold in the first comparison and the deep learning algorithms score below. The LIWC program in combination with the random forest algorithm reached the highest accuracy when comparing mental disorder versus no mental disorder. Most models scored above the threshold except for spaCy in combination with a decision tree in the second comparison. spaCy in combination with the random forest algorithm reached the highest accuracy when comparing the different kinds of mental disorders. Furthermore, the table shows that the random-Forest algorithm performs better than the SVM algorithm with stop words in the first comparison and for the n-grams in the second comparison. For every combination of model and prediction algorithm, the influence of stop words was tested. This can be seen in Table 3 in the sixth and seventh column. The accuracy scores were not improved when removing the stop words, except when using LIWC and the SVM algorithm in the mental disorder vs no mental disorder comparison and RobBERT in the multi-class comparison. Furthermore, spaCy performs more negative without stop words with differences of 0.4 in accuracy and 0.7 in kappa. This could be because spaCy needs those words to keep ordering information. For example, removing the stopwords from ”Ben gives his toy to a friend” changes the sentence into ”Ben gives toy friend”. Cohen’s kappa was calculated for each model and prediction algorithm. Some of the models have classifiers below 0.4 which means that there is a slight agreement. This applies to both deep learning algorithms in both comparisons. A kappa of above 0.6 means that the classifiers have a substantial agreement, for example the n-grams input with the SVM model in the MD (mental disorder) vs Control group comparison. When the kappa is between 0.8 and 1 it means that the classifiers have an almost perfect agreement. This applies to the LIWC-output with

¹² <https://huggingface.co/pdelobelle/robbert-v2-dutch-base>

the random forest model in the second comparison with a kappa of 0.889.

	Input	Model	Acc.	Kappa	Acc. no SW	Kappa no SW
Mental Disorder vs no Mental Disorder	LIWC-output	rpart	0.857	0.667	0.857	0.674
	LIWC-output	random-Forest	0.952	0.889	0.952	0.877
	LIWC-output	SVM	0.857	0.64	0.905	0.738
	n-grams	rpart	0.810	0.391	0.444	-0.309
	n-grams	random-Forest	0.762	0.173	0.389	-0.370
	n-grams	SVM	0.714	0.115	0.528	-0.275
	raw data	fastText	0.643	0.172	0.607	0.072
	raw data	RobBERT	0.607	0.000	0.607	0.000
Mental Disorder multi-class	LIWC-output	rpart	0.286	0.157	0.286	0.177
	LIWC-output	random-Forest	0.214	0.120	0.214	0.144
	LIWC-output	SVM	0.286	0.114	0.143	0.0718
	n-grams	rpart	0.143	-0.0120	0.071	-0.052
	n-grams	random-Forest	0.429	0.304	0.214	0.078
	n-grams	SVM	0.357	0.067	0.143	0.091
	raw data	fastText	0.286	0.000	0.200	0.000
	raw data	RobBERT	0.200	0.000	0.267	0.120

Table 3: Predictions

4.5 Feature importance

The analysis of the most importance features will look at the most prominent differences in language use for people with and without a mental disorder.

4.5.1 Traditional results

In appendix A in figure A.1, A.2 and A.3 the output of the LIWC tool can be seen in jitter-plots. Every plot shows the distribution of the data. As can be seen, some variables have more overlap than others, such as NUMBER and QUANT (quantifier). 1SG and SOCIAL have a clear split between people with and without a mental disorder. These categories can also be seen in the decision tree in figure 4.3.

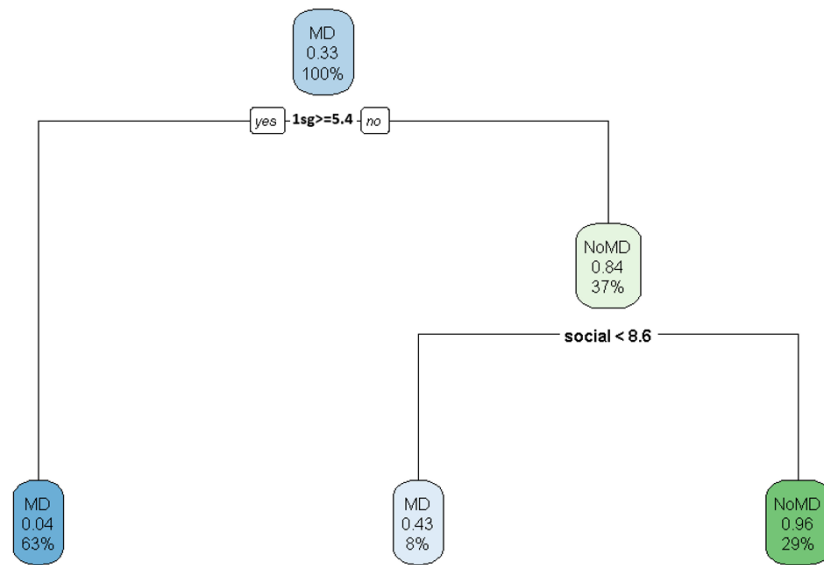


Figure 4.3: LIWC decision tree

A feature selection with random forest is applied to both LIWC and spaCy to determine which variables add most value to the predictions. The figures A.4 and B.2 are in appendix A and B and show the top ten variables per model. Furthermore, a summary per variable per model is shown in appendix A in figure A.5 and in appendix B in figure B.3. The table shows for the top ten features per category and n-gram the mean, standard deviation and standard error.

To assess the influence of people with and without a mental disorder on the use of the 10 variables of LIWC and spaCy, hypothesis pairs are formulated and tested.

- H_0 : The use of variable X by a person with a mental disorder is equal to the use of variable X by a person without a mental disorder.
- H_{1a} : A person with a mental disorder uses variable X significantly more than a person without a mental disorder.
- H_{1b} : A person without a mental disorder uses variable X significantly more than a person with a mental disorder.

For both ten variables, the Mann-Whitney two-tailed U tests were performed to determine whether the means of the two groups per variable were equal to each other. This test was used because both the assumptions, normal distribution and equal variance, were rejected and thus only the Mann-Whitney could be applied for this data. The results of the test are in figure A.6 and B.4 and show whether there are significant differences between people with and without a disorder and if so is it greater or less than the population mean of the other group. A summary of the significant results is shown in table 5. The Mann-Whitney U test revealed no significant difference between people with and without mental disorders for four of the spaCy variables. So the H_0 was rejected for these four.

4.5.2 Deep learning results

LIME was applied to both fastText and RobBERT to gain further insight into the black-box models. For example, quote 1 is from someone who has been diagnosed with schizophrenia and the text is labelled by RobBERT as a mental disorder. The word 'eh' has been highlighted because it explains according to LIME why it was labelled as mental disorder (class = 0). In Figure 4.4, the ten words with the highest probabilities can be seen. These words contribute most to the classification of mental disorder vs no mental disorder. When removing for example the first "Yes" in the graph from the text, the probability of classifying it as a mental disorder goes down with 1,2%. Some words appear multiple times in the figure and this is because it looks locally at a text and words appear in a different context. This also means that sometimes a word will be an explanation for a mental disorder and other times not, based on the context. The second quote is from someone with an eating disorder and analysed by fastText. The word 'Eh' was highlighted because

it explained why the transcription was labelled as a mental disorder (class = `_label_md`). Figure 4.5 shows the ten words with the highest probabilities from that transcription.

Table 4 shows for eight interviews which words resulted in the label mental disorders and which in no mental disorder. The first four interviews were with stop words and as can be seen, most of the words are stop words or 'meaningless' words. They could however be related to insightful words, which also showed in the quotes. This could be supposedly because RobBERT looks both left and right in the context of a word in all of the layers of the transcription and conditions it. As can be seen, some words appear at the mental disorder column and in the no mental disorder column and this is because these words appear in different contexts. So, in one context it contributes to the mental disorder classification and in another context not. The stop words were removed from the last four interviews to look if LIME found more meaningful words. In for example interview 7 with the fastText model, LIME found the words `psychiatric` and `performance` as markers for a mental disorder for that interview and LIME found in interview 8 the words `healing` and `job`. The words found without stop words give a bit more insight than with stop words. However, the words found by LIME are different for almost every interview and thus not applicable for other interviews.

Quote 1: "I ehm, [silence] the most poignant I will you- Yes, the most poignant what I can tell you is that, I have weekend leave on the weekend and then [name][wife] and I lay together in bed. And nothing happens there. Because I don't need that, haha. But I can't even feel that I love her. I know it, that I love her. And I know that my wife is and I, and I. But that's all in here eh, but **I** don't feel it. And that is the biggest measure which you can set.. Yes. And I talked about it with her. "

Quote 2: "Yes it gives kind of a kick or something to go against it and to see that people you really eh yes I don't know. That your that your eating disorder is strong and people find that then. And then you think oh I am good at something . And then yes I don't know. Then you want there that you want

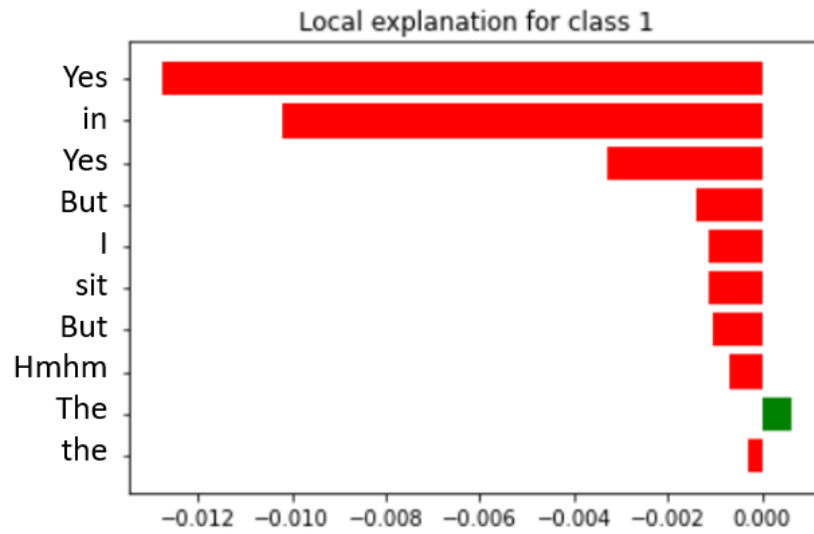


Figure 4.4: LIME explanation quote 1

to be doing something you are good at . . . **eh** I am able to walk again since two months. Before I eh stayed in bed and in a wheelchair around half a year, because I eh could not walk myself. And I was just too weak to do it. and eh yes I still cannot do quite a lot of things. I am really happy that I can walk again by myself.”

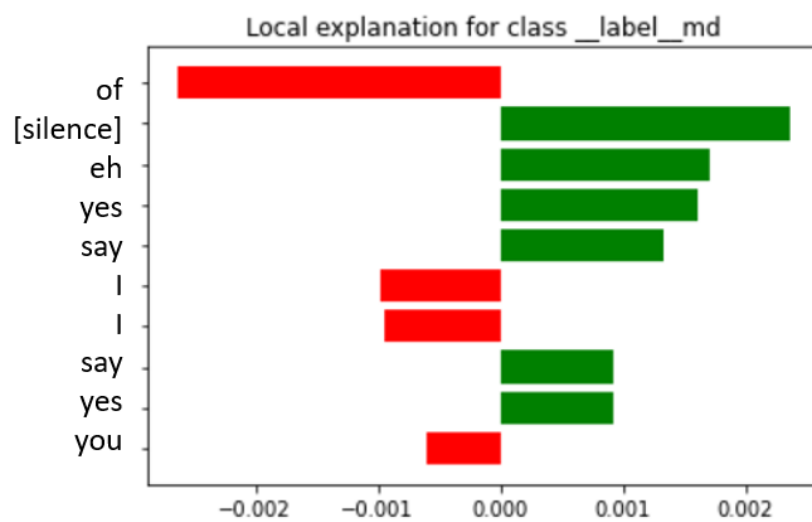


Figure 4.5: LIME explanation quote 2

ID	MD?	SW?	Rob- BERT	fast- Text	Words MD BERT	Words noMD BERT	Words MD fastText	Words noMD fastText
1	Y	Y	0.68	0.77	everyone, too, because, Yes, For example, too, Yes, I, did	-	yes, with, is, ., common, me	from, common, common, eh
2	Y	Y	0.55	0.69	feel, allowed, I, really, eh, angry, they, You	[name], there	together, am, well, well, .	am, I, me, my
3	N	Y	0.39	0.45	happy, the, looking back, Well, belongs, eh, always, no, well, think	-	say, come, yes, and, causing	not, that, [place name], week, say
4	N	Y	0.37	0.23	could, can, And, That, sat, be, chats, and, whole	walked	protected, to, is, do, bad, have, is, physical, am	walks
5	Y	N	0.68	0.77	ehm, one, bill, yes, distraction, recovery	sat, eh, real, goes	yes, well, that, yes, well, rest	if, but, better, care
6	Y	N	0.58	0.65	eh	hospital, And, whole, whole, she, one, also, eh, again	whole, completely, ., further, times	stood, sick, selfish, and, eh
7	N	N	0.41	0.46	eh, nineteen ninety seven, of, notices of objection, say, team	car, ehm, team, through, But	psychiatrics, performance, one, he	that, en route, exciting, we, go, and
8	N	N	0.49	0.43	married, common, a, sit, heaven, times, and, The	ehm, ehm	sewn, healing, and, but, job	huh, hear, term, ready, busy

Table 4: LIME output for fastText and RobBERT

4.5.3 Summary feature importance results

The table 5 shows per tool the uncovered language markers for LIWC and spaCy. The 1SG LIWC pronoun tested as a language marker for a person with a mental disorder and this pronoun is also for spaCy the basis for every language marker for a mental disorder. The LIME results of RobBERT and fastText had no clear patterns, for every interview different words were found that indicated a mental disorder or no mental disorder. So, no language markers were found with the deep learning models.

	Language marker	Mental Disorder?	W; $p < 0.05$
LIWC	1sg	Yes	2487
	focuspast	Yes	1856
	affiliation	No	380
	drives	No	568
	female	No	937
	male	No	767
	3sg	No	454
	social	No	281
	3pl	No	882
	1pl	No	217.5
spaCy	ik_doen_nsubj	Yes	1700.5
	ik_gaan_nsubj	Yes	1726
	ik_hebben_nsubj	Yes	1796.5
	ik_komen_nsubj	Yes	1852.5
	er_zijn_advmod	No	849
	ze_hebben_nsubj	No	768.5

Table 5: Language markers for LIWC and spaCy

4.6 Results discussed by domain experts

The results of the different models were discussed with data scientists, researchers and psychologists of the UMCU.

The data used for this research is static data and by that is meant that somebody tells their story and that is it. There is no new data of that person added after a while. By following a person in their healing process and their language usage over a longer period of time could result in interesting outcomes according to the data scientists. Furthermore, the language markers found by LIWC and spaCy were discussed. The data is from people with a mental disorder who tell their own story and from medical employees and family members who talk about people with a mental disorder. This influences the outcome of this research, because if a person tells their own story, they will probably use more the 1sg pronoun. Also, if a medical employee talks about an experience with a patient, they could use more the 3sg and 3pl pronouns. The people with a mental disorder also tell their story when they are not in an acute phase and they could be talking more about a completed story in their past. So, the language markers are really logical according the experts. The classifications are abandoned in the psychiatry, because they do not really help a person according to the psychologist. However, if the outcome is changed to for example how far someone is in their healing process, you could also find interesting results. The models used in this research could be applied for this new direction. A hypothesis was that people who are further into their healing process, tell a more integrated story about their past than a person who is less far. So, focuspast could be marker for someone further into the healing process. Another idea was that this research could be used to look at symptoms instead of the diagnosis. What kind of treatment will help a person based on what somebody says. Or look at suicidality or aggression, what can a text tell us about that? So, find out what a person is not explicitly telling, by analysing the deeper layers to find possible patterns or symptoms. One domain expert said: "The strength of this research lays not in the exact results, but in the application of the different models and the potential questions which could be answered by these models."

5 Discussion

In this section the results from previous sections, limitations and future work will be discussed.

5.1 Interpretation and implications of results

The results from the predictions of the different models showed that LIWC and spaCy performed best. When looking more into the why, one could conclude that both look at the stylistic differences between language users. This was also shown in the use of the 1sg pronoun, because with both models the people with a mental disorder use the pronoun significantly more. Furthermore, the focuspast could be explained by the fact that the people with their mental disorders talked about what happened to them in their past. The pronouns 3sg, 3pl and 1pl were in the literature review associated with several mental disorders, but were in this research associated with no mental disorder. This could be explained by the fact that the people talked about their relatives or patients in the interviews. Furthermore, stop words do not appear to have a positive influence on the performance of the classifications except when using LIWC and the SVM algorithm in the mental disorder vs no mental disorder comparison and RobBERT in the multi-class comparison.

5.2 Limitations

This research had a number of limitations and all of them are linked to the dataset. First, the people with the mental disorders were labelled based on their diagnosis. However as said in the introduction, diagnosing a mental disorder is complex due to the different factors. So, it could be that some interviews were not labelled correctly. Second, the interviews were in Dutch. This resulted in less options in language models to choose from. Third, the people without a mental disorder talked about their family or patients with a mental disorder. This could influence the performance of the models. Fourthly, the dataset was static. The interviews were a snapshot of somebody's use of language. It could be that a person had a good or bad day with their disorder and this could have influenced their use

of language. Lastly, the dataset was limited. Only 72 interviews of people with a mental disorder were collected and there were around 10 kinds of mental disorders. So, the dataset included not even 10 interviews per disorder and some disorders contained of only 2 or 3 interviews transcriptions. This was also why the deep learning models were used, because they exploit transfer learning to extract features that can easily be reused for classification and thus require less data to achieve high performance (Feng et al., 2021). Furthermore, the balance between people with and without a mental disorder was skewed, which could also influence the results.

5.3 Future work

This research project is a start for new possibilities in the field of language markers in the mental health care domain. First, alternative NLP-models could be applied. For this research, the approach of applying for example XLM combined with ClinicalBERT was not used, because these models have not been used extensively in the medical domain. Research into applying different kinds of models to interview transcriptions could lead to new insights. Moreover, the two deep learning models were chosen because of performance. However, it could be that these two were not the best in picking up mental disorder cues and others can reach a better accuracy.

Second, new language markers could be found by exploring more advanced methods to explain blackbox-models. This research focused on LIME, because SHAP (SHapely Additive exPlanation) does not support a fastText-model. However, when using other models or other explanation techniques, new insights could be gained.

Third, the data and outcome of the models could be changed. This research could be the stepping stone for looking into symptoms or stage in the healing process of a person based on the use of language. The different models in this research show their strengths and weaknesses and will help by choosing the most suitable one.

Fourth, when more data is available for more different kinds of mental disorder new insights and language markers could be gained. Moreover, the DSM-5 is the manual to classify a person and when language markers per mental disorder are mapped, it could be a tool to support medical professionals.

6 Conclusion

The main question of this research was: *"To what extent can a diagnosis of mental illness be determined by language markers?"*. First, the five sub-questions are answered and next the main question is answered.

6.1 Sub-questions

SQ1 Which traditional sentence level approach is suitable to detect language markers?

This question was answered in section 3.3. Several dependency parsers were already compared by another researcher. However these parsers did not have a Dutch dictionary. In the end, the three Dutch dependency parsers Frog, Alpino and spaCy were compared and spaCy was chosen based on equipment constraints.

SQ2 What neural approach is suitable to detect language markers at word level?

In section 3.4 two deep learning models were compared at word-level, Word2Vec and fastText. The model chosen was fastText, because it can handle OOV-words.

SQ3 What neural approach is suitable to detect language markers at sentence level?

Furthermore, the other deep learning models compared in section 3.4 were all at sentence-level. The models without a Dutch dictionary could be combined with a transfer learning model, for example ClinicalBERT. However research suggested that this was not extensively used in the medical domain. The models with a Dutch dictionary were compared based on classification performance. Between BERTje (accuracy = 93%) and RobBERT (accuracy = 95.1%), RobBERT was chosen because it outperformed BERTje.

SQ4 How well do the techniques perform on Dutch narratives?

The models and classification algorithms were used to make predictions for the two comparisons. The binary prediction between mental disorder and no mental disorder reached an accuracy of 95% with LIWC and random forest. The mental disorder multi-class comparison reached with spaCy and random forest an accuracy of 43%. The two models, which performed best, are based on stylistic differences rather than neural networks.

SQ5 To what extent can we identify meaningful language markers for having a mental disorder

It is too early to quantify to what extent we can identify the markers. However, as seen in section 4.5.3, there are classifiers to find out if a person has a mental disorder or not. LIME was applied to the two deep learning models and explained why a few of the interviews were classified as mental disorder and others were not. However, this could not be generalised and thus no language markers were found with LIME. These results were also discussed with domain experts. They said that the strength of this research does not lie in the exact results, but it is a stepping stone for other potential questions which could be answered by the different models used in this research.

6.2 Main research question

The goal of this research was to explore language markers in Dutch psychiatric interview transcriptions. We found that research thus far mainly focused on LIWC. So, first task was to investigate several traditional and deep learning models and spaCy, fastText and RobBERT were chosen. Next, the prediction performances of LIWC, spaCy, fastText and RobBERT were compared. The best performing technique to find out if a person has a mental disorder is LIWC in combination with the classification algorithm random forest which reached an accuracy-score of 0.952 and a Cohen's kappa of 0.889. spaCy in combination with random forest predicted best which men-

tal disorder a person has with an accuracy-score of 0.429 and a Cohen's kappa of 0.304. Furthermore, as could be seen in table 5 in section 4.5.3, several language markers were found. With these markers, the LIWC-decision tree and an interview transcription, there could be determined if a person has a mental disorder or not.

References

- Addawood, A., Badawy, A., Lerman, K. & Ferrara, E. (2019). Linguistic cues to deception: Identifying political trolls on social media, In *Proceedings of the international aaai conference on web and social media*.
- Bohnet, B. (2010). Top accuracy and fast dependency parsing is not a contradiction, In *Proceedings of the 23rd international conference on computational linguistics (coling 2010)*.
- Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146.
- Bosch, A. v. d., Busser, B., Canisius, S. & Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for dutch. *LOT Occasional Series*, 7, 191–206.
- Bosnjak, Z., Grljevic, O. & Bosnjak, S. (2009). Crisp-dm as a framework for discovering knowledge in small and medium sized enterprises' data, In *2009 5th international symposium on applied computational intelligence and informatics*. IEEE.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. Et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Calvo, R. A., Milne, D. N., Hussain, M. S. & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5), 649–685.
- Cameron, R. P. & Gusman, D. (2003). The primary care ptsd screen (pc-ptsd): Development and operating characteristics. *Primary care psychiatry*, 9(1), 9–14.
- Carter, P. E. & Grenyer, B. F. (2012). Expressive language disturbance in borderline personality disorder in response to emotional autobiographical stimuli. *Journal of personality disorders*, 26(3), 305–321.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (1999). The crisp-dm user guide, In *4th crisp-dm sig workshop in brussels in march*.

- Choi, J. D. & McCallum, A. (2013). Transition-based dependency parsing with selectional branching, In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)*.
- Choi, J. D., Tetreault, J. & Stent, A. (2015). It depends: Dependency parser comparison using a web-based evaluation tool, In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37–46.
- Coppersmith, G., Dredze, M., Harman, C. & Hollingshead, K. (2015). From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses, In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*.
- Corcoran, C. M. & Cecchi, G. (2020). Using language processing and speech analysis for the identification of psychosis and other disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*.
- Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V. & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.
- Davcheva, E. (2018). Text mining mental health forums—learning from user experiences.
- de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G. & Nissim, M. (2019). Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.
- Delobelle, P., Winters, T. & Berendt, B. (2020). Robbert: A dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*.
- Deng, L., Yu, D. Et al. (2014). Deep learning: Methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4), 197–387.
- Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

- Evans, S., Ferrando, S., Findler, M., Stowell, C., Smart, C. & Haglin, D. (2008). Mindfulness-based cognitive therapy for generalized anxiety disorder. *Journal of anxiety disorders*, 22(4), 716–721.
- Fairburn, C. G., Cooper, Z. & O'Connor, M. (1993). The eating disorder examination. *International Journal of Eating Disorders*, 6, 1–8.
- Fava, M. & Kendler, K. S. (2000). Major depressive disorder. *Neuron*, 28(2), 335–341.
- Feng, S., Fu, H., Zhou, H., Wu, Y., Lu, Z. & Dong, H. (2021). A general and transferable deep learning framework for predicting phase formation in materials. *npj Computational Materials*, 7(1), 1–10.
- Fineberg, S., Deutsch-Link, S., Ichinose, M., McGuinness, T., Bessette, A., Chung, C. & Corlett, P. (2015). Word use in first-person accounts of schizophrenia. *The British Journal of Psychiatry*, 206(1), 32–38.
- Forgeard, M. (2008). Linguistic styles of eminent writers suffering from unipolar and bipolar mood disorder. *Creativity Research Journal*, 20(1), 81–92.
- Friedman, J., Hastie, T. & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432–441.
- Gamallo, P. (2017). Sense contextualization in a dependency-based compositional distributional model, In *Proceedings of the 2nd workshop on representation learning for nlp*.
- Groom, C. J. & Pennebaker, J. W. (2002). Words. *Journal of Research in Personality*, 36(6), 615–621.
- Guevara, E. R. (2010). A regression model of adjective-noun compositionality in distributional semantics, In *Proceedings of the 2010 workshop on geometrical models of natural language semantics*.
- Hagiwara, M. (2018). Improving a sentiment analyzer using elmo — word embeddings on steroids, In *Real-world natural language processing*.
- Hermann, K. M. (2014). Distributed representations for compositional semantics. *arXiv preprint arXiv:1411.3146*.

- Honnibal, M. (2016). Embed, encode, attend, predict: The new deep learning formula for state-of-the-art nlp models. *Blog, Explosion*, November, 10.
- Howard, J. & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.
- Huang, K., Altosaar, J. & Ranganath, R. (2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.
- Insel, T. R. (2010). Rethinking schizophrenia. *Nature*, 468(7321), 187–193.
- Khattak, F. K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C. & Rudzicz, F. (2019). A survey of word embeddings for clinical text. *Journal of Biomedical Informatics: X*, 4, 100057.
- Kim, K., Lee, S. & Lee, C. (2015). College students with adhd traits and their language styles. *Journal of attention disorders*, 19(8), 687–693.
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.
- Le, N. Q. K., Yapp, E. K. Y., Nagasundaram, N. & Yeh, H.-Y. (2019). Classifying promoters by interpreting the hidden information of dna sequences via deep learning and combination of continuous fasttext n-grams. *Frontiers in bioengineering and biotechnology*, 7, 305.
- Lei, T., Xin, Y., Zhang, Y., Barzilay, R. & Jaakkola, T. (2014). Low-rank tensors for scoring dependency structures, In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*.
- Liang, P. & Potts, C. (2015). Bringing machine learning and compositional semantics together. *Annu. Rev. Linguist.*, 1(1), 355–376.
- Lieb, K., Zanarini, M. C., Schmahl, C., Linehan, M. M. & Bohus, M. (2004). Borderline personality disorder. *The Lancet*, 364(9432), 453–461.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

- Lopez, M. M. & Kalita, J. (2017). Deep learning applied to nlp. *arXiv preprint arXiv:1703.03091*.
- Lyons, M., Aksayli, N. D. & Brewer, G. (2018). Mental distress and language use: Linguistic analysis of discussion forum posts. *Computers in Human Behavior, 87*, 207–211.
- March, J. S. & Mulle, K. (1998). *Ocd in children and adolescents: A cognitive-behavioral treatment manual*. Guilford Press.
- McIntosh, A. M., Stewart, R., John, A., Smith, D. J., Davis, K., Sudlow, C., Corvin, A., Nicodemus, K. K., Kingdon, D., Hassan, L. Et al. (2016). Data science for mental health: A uk perspective on a global challenge. *The Lancet Psychiatry, 3*(10), 993–998.
- Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality, In *Advances in neural information processing systems*.
- Miotto, R., Wang, F., Wang, S., Jiang, X. & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in bioinformatics, 19*(6), 1236–1246.
- Moraes, R., Valiati, J. F. & Neto, W. P. G. (2013). Document-level sentiment classification: An empirical comparison between svm and ann. *Expert Systems with Applications, 40*(2), 621–633.
- Nguyen, T., Phung, D. & Venkatesh, S. (2013). Analysis of psycholinguistic processes and topics in online autism communities, In *2013 IEEE international conference on multimedia and expo (icme)*. IEEE.
- Otter, D. W., Medina, J. R. & Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*.
- Pan, S. J. & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering, 22*(10), 1345–1359.
- Papini, S., Yoon, P., Rubin, M., Lopez-Castro, T. & Hien, D. A. (2015). Linguistic characteristics in a non-trauma-related nar-

- rative task are associated with ptsd diagnosis and symptom severity. *Psychological Trauma: Theory, Research, Practice, and Policy*, 7(3), 295.
- Pennebaker, J. W., Boyd, R. L., Jordan, K. & Blackburn, K. (2015). *The development and psychometric properties of liwc2015* (tech. rep.).
- Pennebaker, J. W., Francis, M. E. & Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001), 2001.
- Perera, G., Broadbent, M., Callard, F., Chang, C.-K., Downs, J., Dutta, R., Fernandes, A., Hayes, R. D., Henderson, M., Jackson, R. Et al. (2016). Cohort profile of the south london and maudsley nhs foundation trust biomedical research centre (slam brc) case register: Current status and recent enhancement of an electronic mental health record-derived data resource. *BMJ open*, 6(3), e008721.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.
- Polanczyk, G., De Lima, M. S., Horta, B. L., Biederman, J. & Rohde, L. A. (2007). The worldwide prevalence of adhd: A systematic review and metaregression analysis. *American journal of psychiatry*, 164(6), 942–948.
- Prizant, B. M. (1983). Language acquisition and communicative behavior in autism: Toward an understanding of the” whole” of it. *Journal of speech and hearing disorders*, 48(3), 296–307.
- Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. (2018). Improving language understanding by generative pre-training.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 9.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.
- Remmers, C. & Zander, T. (2018). Why you don’t see the forest for the trees when you are anxious: Anxiety impairs intuitive decision making. *Clinical Psychological Science*, 6(1), 48–62.

- Ribeiro, M. T., Singh, S. & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier, In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*.
- Ritchie, H. & Roser, M. (2020). Mental health [<https://ourworldindata.org/mental-health>]. *Our World in Data*.
- Russ, T. C., Woelbert, E., Davis, K. A., Hafferty, J. D., Ibrahim, Z., Inkster, B., John, A., Lee, W., Maxwell, M., McIntosh, A. M. Et al. (2019). How data science can advance mental health research. *Nature human behaviour*, 3(1), 24–32.
- Sarhan, I. & Spruit, M. (2020). Can we survive without labelled data in nlp? transfer learning for open information extraction. *Applied Sciences*, 10(17), 5758.
- Shearer, C. (2000). The crisp-dm model: The new blueprint for data mining. *Journal of data warehousing*, 5(4), 13–22.
- Spruit, M. & Lytras, M. (2018). Applied data science in patient-centric healthcare: Adaptive analytic systems for empowering physicians and patients. Elsevier.
- Tausczik, Y. R. & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1), 24–54.
- Torrey, L. & Shavlik, J. (2010). Transfer learning, In *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques*. IGI global.
- Trautmann, S., Rehm, J. & Wittchen, H.-U. (2016). The economic costs of mental disorders: Do our societies react appropriately to the burden of mental disorders? *EMBO reports*, 17(9), 1245–1249.
- Trifu, R. N., NEMEŞ, B., BODEA-HAŢEGAN, C. & Cozman, D. (2017). Linguistic indicators of language in major depressive disorder (mdd). an evidence based research. *Journal of Evidence-Based Psychotherapies*, 17(1).
- Van der Beek, L., Bouma, G., Malouf, R. & Van Noord, G. (2002). The alpine dependency treebank, In *Computational linguistics in the netherlands 2001*. Brill Rodopi.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need, In *Advances in neural information processing systems*.

- Wakefield, J. C. (1992). Disorder as harmful dysfunction: A conceptual critique of dsm-iii-r's definition of mental disorder. *Psychological review*, *99*(2), 232.
- Wang, W., Bi, B., Yan, M., Wu, C., Bao, Z., Xia, J., Peng, L. & Si, L. (2019). Structbert: Incorporating language structures into pre-training for deep language understanding. *arXiv preprint arXiv:1908.04577*.
- Webster, J. & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly*, xiii–xxiii.
- Whiteford, H. A., Degenhardt, L., Rehm, J., Baxter, A. J., Ferrari, A. J., Erskine, H. E., Charlson, F. J., Norman, R. E., Flaxman, A. D., Johns, N. Et al. (2013). Global burden of disease attributable to mental and substance use disorders: Findings from the global burden of disease study 2010. *The lancet*, *382*(9904), 1575–1586.
- Wieringa, R. J. (2014). *Design science methodology for information systems and software engineering*. Springer.
- Wirth, R. & Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining, In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Springer-Verlag London, UK.
- Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering, In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. Et al. (2019). Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
- Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R. & Le, Q. V. (2019). Xlnet: Generalized autoregressive pre-training for language understanding, In *Advances in neural information processing systems*.
- Young, T., Hazarika, D., Poria, S. & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *ieee Computational intelligence magazine*, *13*(3), 55–75.

Zimmerer, V. C., Watson, S., Turkington, D., Ferrier, I. N. & Hinzen, W. (2017). Deictic and propositional meaning—new perspectives on language in schizophrenia. *Frontiers in psychiatry*, 8, 17.

Appendix A LIWC

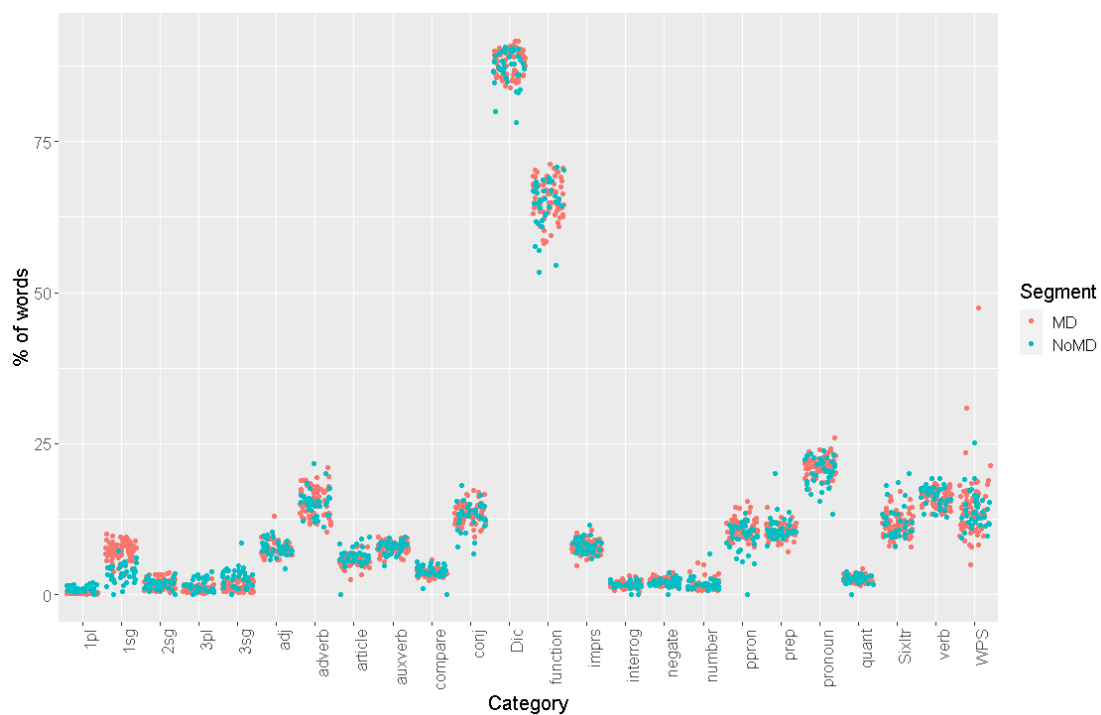


Figure A.1: LIWC output part 1

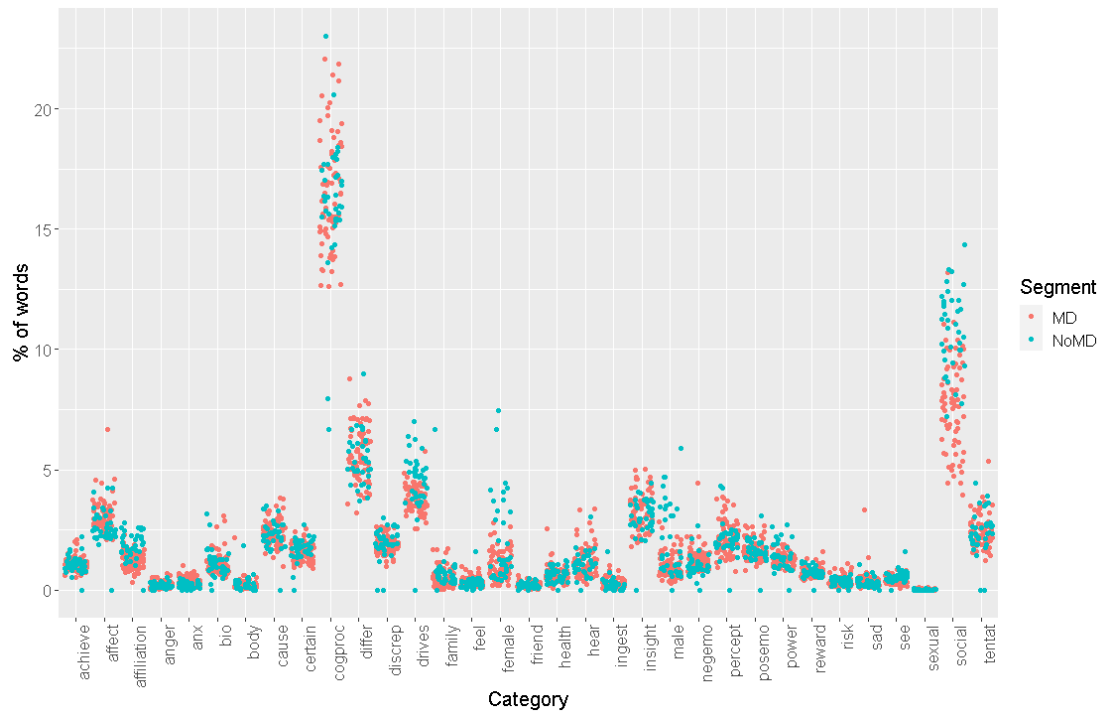


Figure A.2: LIWC output part 2

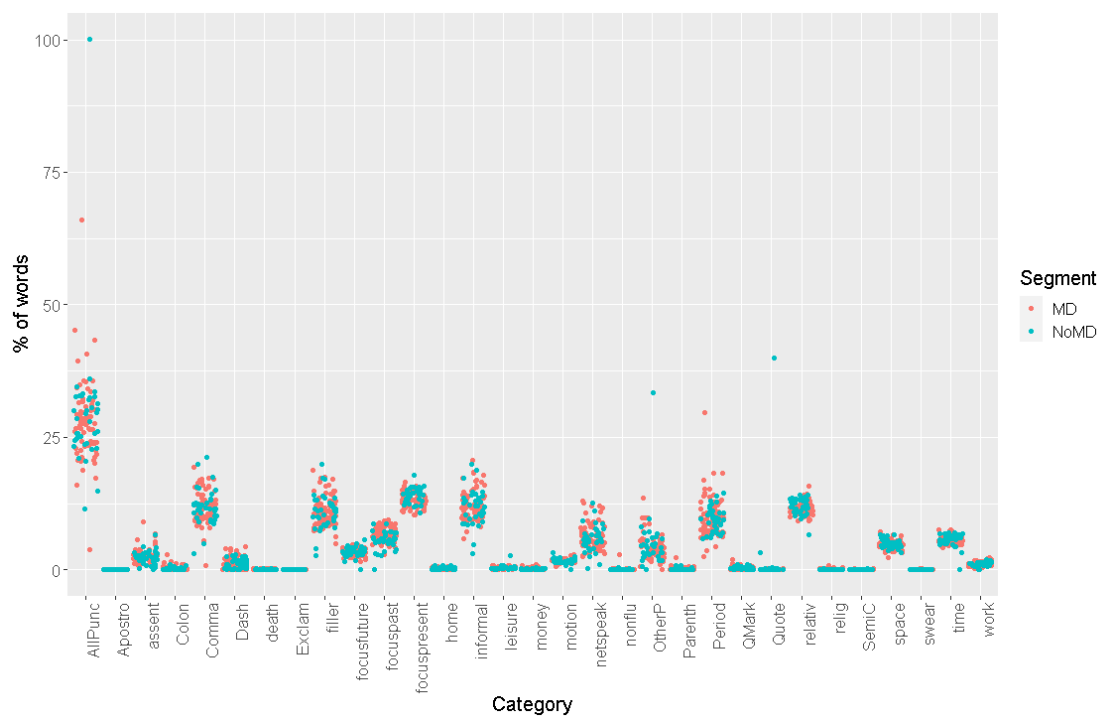


Figure A.3: LIWC output part 3

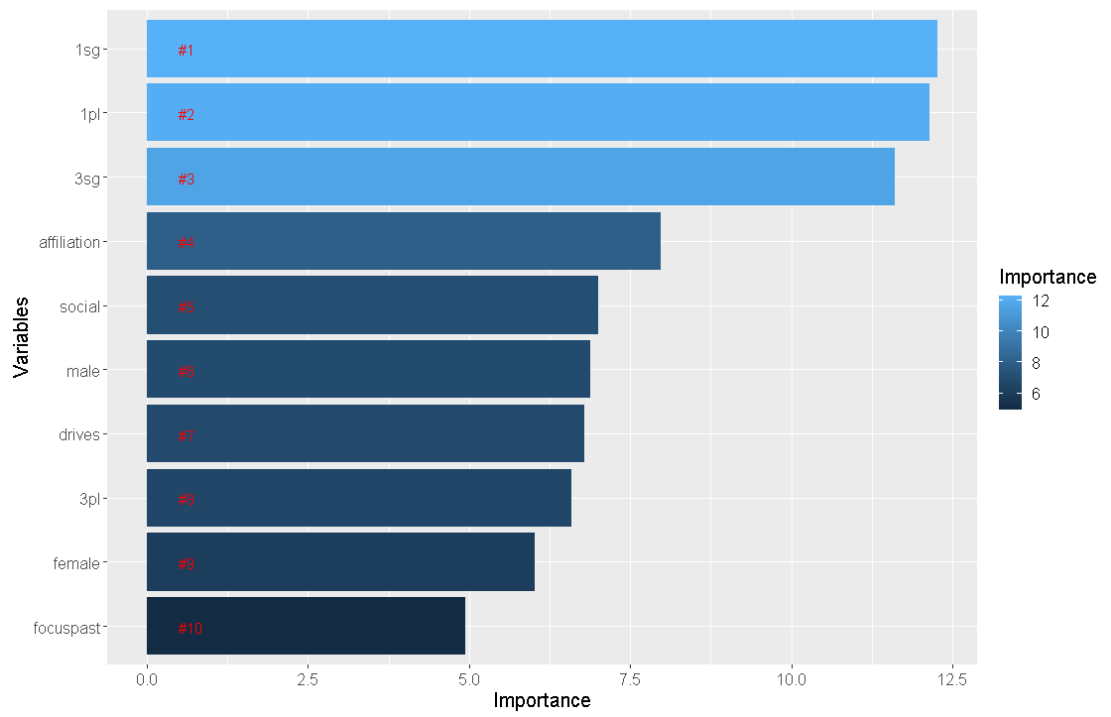


Figure A.4: LIWC feature importance top 10 for the binary classification

Category	Disorder	n	mean	sd	se
affiliation	MD	72	1.186944	0.369113	0.0435
	NoMD	36	1.849722	0.575462	0.09591
drives	MD	72	3.805972	0.546836	0.064445
	NoMD	36	4.573056	1.193685	0.198948
female	MD	72	1.065278	0.594782	0.070096
	NoMD	36	2.047778	1.798682	0.29978
focuspast	MD	72	6.708611	1.310005	0.154386
	NoMD	36	5.453333	1.72206	0.28701
i	MD	72	7.14375	1.46491	0.172641
	NoMD	36	3.431389	1.496495	0.249416
male	MD	72	1.017639	0.532865	0.062799
	NoMD	36	2.048611	1.503014	0.250502
shehe	MD	72	1.269306	0.655989	0.077309
	NoMD	36	2.855556	1.577011	0.262835
social	MD	72	7.729583	1.860704	0.219286
	NoMD	36	10.82139	1.661339	0.27689
they	MD	72	0.864583	0.466053	0.054925
	NoMD	36	1.404722	0.965145	0.160858
we	MD	72	0.345972	0.199391	0.023498
	NoMD	36	0.991667	0.450006	0.075001

Figure A.5: LIWC summary top 10 features

Category	group1	group2	n1	n2	t	p	p.adj	p.adj.signif
affiliation	MD	NoMD	72	36	380	2.42E-09	6.05E-09	****
drives	MD	NoMD	72	36	568	2.12E-06	3.53E-06	****
female	MD	NoMD	72	36	937	0.0195	0.0195	*
focuspast	MD	NoMD	72	36	1856	0.000266	0.00038	***
i	MD	NoMD	72	36	2487	8.57E-15	8.57E-14	****
male	MD	NoMD	72	36	767	0.000572	0.000715	***
shehe	MD	NoMD	72	36	454	4.15E-08	8.3E-08	****
social	MD	NoMD	72	36	281	3.8E-11	1.27E-10	****
they	MD	NoMD	72	36	882	0.00704	0.007822	**
we	MD	NoMD	72	36	217.5	2.12E-12	1.06E-11	****

Figure A.6: LIWC Mann-Whitney U-test top 10 features

Appendix B spaCy

spaCy variable	Example
ik_doen_nsubj I_do_nsubj	Ik doe normaal, haal mijn studie en gebruik geen drugs en ben niet irritant I do normal, get my degree and don't use drugs and am not irritating
ik_gaan_nsubj I_go_nsubj	ik ben meer waard dan dit, ik ga voor mezelf opkomen. I am worth more than this, I'm going to stand up for myself
ik_hebben_nsubj I_have_nsubj	Ik heb ook behandelingen gehad, of een behandeling gehad I have also gotten treatments, or got a treatment
ik_komen_nsubj I_come_nsubj	Ja, ik kwam in de bijstand. Yes, I came into welfare.
er_zijn_advmod there_are_advmod	Er zijn zo veel vrouwelijke sociotherapeuten in heel [naam][centrum] die opgeroepen kunnen worden There are so many female sociotherapists in [name][centre] who can be called
ze_hebben_nsubj they_have_nsubj	Al een tijdje maar ze hebben nooit wat aan mij verteld For some time, but they have never told me anything

Table 6: Examples of spaCy variables

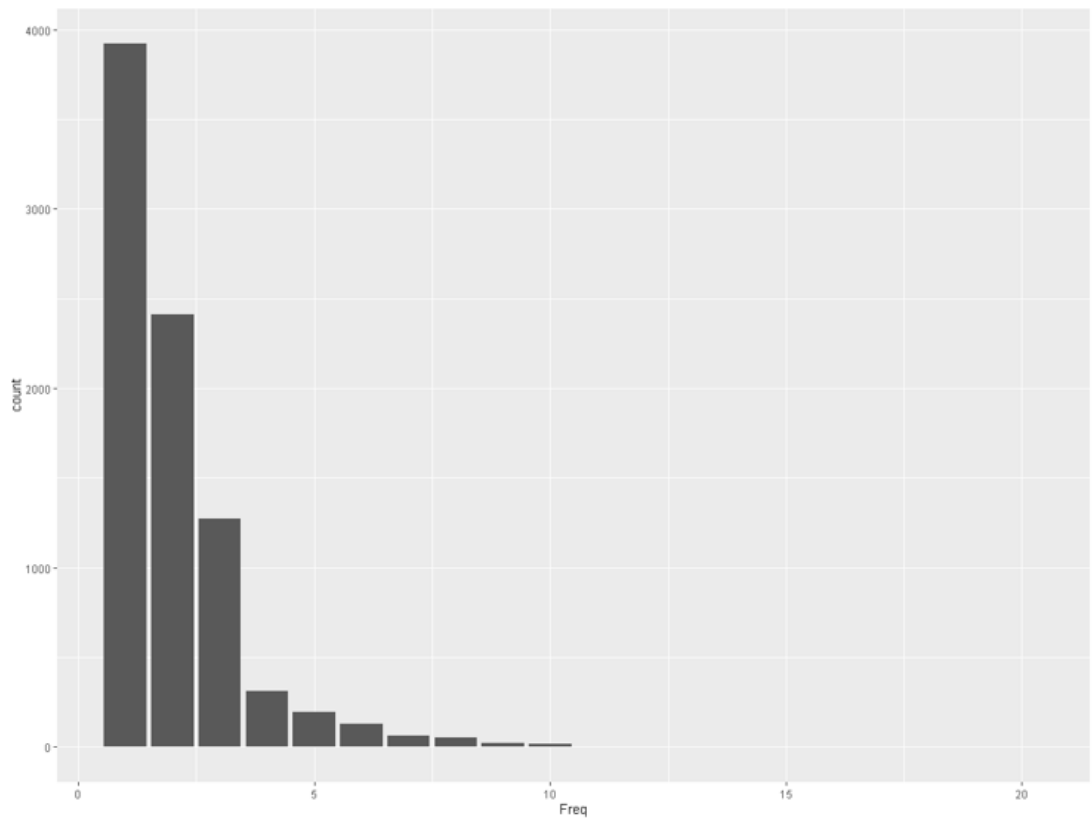


Figure B.1: Frequency of the spaCy variables

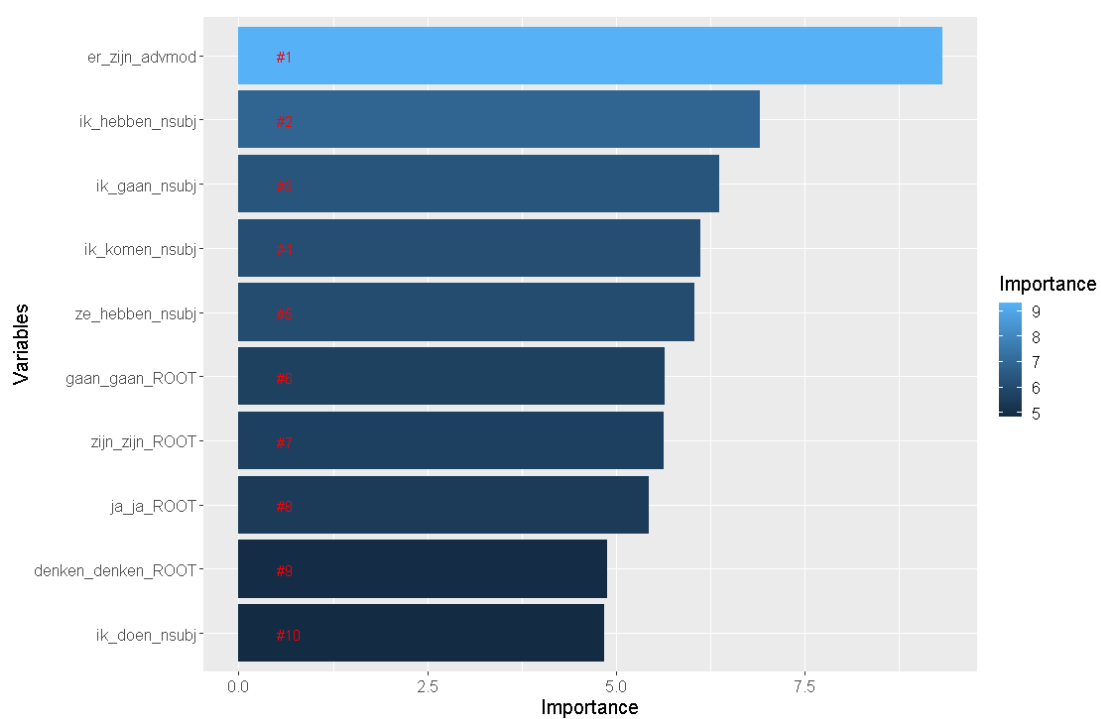


Figure B.2: spaCy feature importance top 10 for the binary classification

ngram	Disorder	n	mean	sd	se
denken_denken_ROOT	MD	72	19.375	17.03057524	2.00707254
	NonMD	36	20.22222222	13.26889189	2.211481982
er_zijn_advmod	MD	72	13	12.07255064	1.422763737
	NonMD	36	18.80555556	9.080023425	1.513337237
gaan_gaan_ROOT	MD	72	30.30555556	22.68749043	2.673746389
	NonMD	36	36.77777778	27.19675518	4.532792529
ik_doen_nsubj	MD	72	11.47222222	10.76898549	1.269137111
	NonMD	36	5.75	8.083051051	1.347175175
ik_gaan_nsubj	MD	72	17.29166667	14.12463386	1.664604064
	NonMD	36	9.638888889	11.58854799	1.931424664
ik_hebben_nsubj	MD	72	49.34722222	38.89355634	4.583649572
	NonMD	36	25.61111111	20.6013561	3.433559349
ik_komen_nsubj	MD	72	6.75	7.084500042	0.834916337
	NonMD	36	1.777777778	4.427905736	0.737984289
ja_ja_ROOT	MD	72	27.33333333	25.80206346	3.04080234
	NonMD	36	30.58333333	24.41940095	4.069900159
ze_hebben_nsubj	MD	72	2.902777778	5.39994711	0.63638987
	NonMD	36	11.58333333	15.08144555	2.513574259
zijn_zijn_ROOT	MD	72	16.90277778	13.56690227	1.598874766
	NonMD	36	20.72222222	13.15753149	2.192921915

Figure B.3: spaCy summary top 10 features

ngram	group1	group2	n1	n2	W	p	p.adj	p.adj.signif
denken_denken_ROOT	MD	NonMD	72	36	1186.5	0.475	0.475	ns
er_zijn_advmod	MD	NonMD	72	36	849	0.00337	0.008425	**
gaan_gaan_ROOT	MD	NonMD	72	36	1110.5	0.226	0.2825	ns
ik_doen_nsubj	MD	NonMD	72	36	1700.5	0.00673	0.011216667	*
ik_gaan_nsubj	MD	NonMD	72	36	1726	0.00439	0.00878	**
ik_hebben_nsubj	MD	NonMD	72	36	1796.5	0.00106	0.003533333	**
ik_komen_nsubj	MD	NonMD	72	36	1852.5	0.0000681	0.000565	***
ja_ja_ROOT	MD	NonMD	72	36	1170	0.412	0.457777778	ns
ze_hebben_nsubj	MD	NonMD	72	36	768.5	0.000113	0.000565	***
zijn_zijn_ROOT	MD	NonMD	72	36	1091	0.181	0.258571429	ns

Figure B.4: spaCy Mann-Whitney U-test top 10 features

Appendix C LIWC categories

The 2 pages below show the different categories of LIWC, the abbreviations and a few examples per category (Pennebaker et al., 2015).

Each of the default LIWC2015 categories is composed of a list of dictionary words that define that scale. Table 1 provides a comprehensive list of the default LIWC2015 dictionary categories, scales, sample scale words, and relevant scale word counts.

Table 1. LIWC2015 Output Variable Information

Category	Abbrev	Examples	Words in category	Internal Consistency (Uncorrected α)	Internal Consistency (Corrected α)
Word count	WC	-	-	-	-
Summary Language Variables					
Analytical thinking	Analytic	-	-	-	-
Clout	Clout	-	-	-	-
Authentic	Authentic	-	-	-	-
Emotional tone	Tone	-	-	-	-
Words/sentence	WPS	-	-	-	-
Words > 6 letters	Sixltr	-	-	-	-
Dictionary words	Dic	-	-	-	-
Linguistic Dimensions					
Total function words	funct	it, to, no, very	491	.05	.24
Total pronouns	pronoun	I, them, itself	153	.25	.67
Personal pronouns	ppron	I, them, her	93	.20	.61
1st pers singular	i	I, me, mine	24	.41	.81
1st pers plural	we	we, us, our	12	.43	.82
2nd person	you	you, your, thou	30	.28	.70
3rd pers singular	shehe	she, her, him	17	.49	.85
3rd pers plural	they	they, their, they'd	11	.37	.78
Impersonal pronouns	ipron	it, it's, those	59	.28	.71
Articles	article	a, an, the	3	.05	.23
Prepositions	prep	to, with, above	74	.04	.18
Auxiliary verbs	auxverb	am, will, have	141	.16	.54
Common Adverbs	adverb	very, really	140	.43	.82
Conjunctions	conj	and, but, whereas	43	.14	.50
Negations	negate	no, not, never	62	.29	.71
Other Grammar					
Common verbs	verb	eat, come, carry	1000	.05	.23
Common adjectives	adj	free, happy, long	764	.04	.19
Comparisons	compare	greater, best, after	317	.08	.35
Interrogatives	interrog	how, when, what	48	.18	.57
Numbers	number	second, thousand	36	.45	.83
Quantifiers	quant	few, many, much	77	.23	.64
Psychological Processes					
Affective processes	affect	happy, cried	1393	.18	.57
Positive emotion	posemo	love, nice, sweet	620	.23	.64
Negative emotion	negemo	hurt, ugly, nasty	744	.17	.55
Anxiety	anx	worried, fearful	116	.31	.73
Anger	anger	hate, kill, annoyed	230	.16	.53
Sadness	sad	crying, grief, sad	136	.28	.70
Social processes	social	mate, talk, they	756	.51	.86
Family	family	daughter, dad, aunt	118	.55	.88

Category	Abbrev	Examples	Words in category	Internal Consistency (Uncorrected α)	Internal Consistency (Corrected α)
Friends	friend	buddy, neighbor	95	.20	.60
Female references	female	girl, her, mom	124	.53	.87
Male references	male	boy, his, dad	116	.52	.87
Cognitive processes	cogproc	cause, know, ought	797	.65	.92
Insight	insight	think, know	259	.47	.84
Causation	cause	because, effect	135	.26	.67
Discrepancy	discrep	should, would	83	.34	.76
Tentative	tentat	maybe, perhaps	178	.44	.83
Certainty	certain	always, never	113	.31	.73
Differentiation	differ	hasn't, but, else	81	.38	.78
Perceptual processes	percept	look, heard, feeling	436	.17	.55
See	see	view, saw, seen	126	.46	.84
Hear	hear	listen, hearing	93	.27	.69
Feel	feel	feels, touch	128	.24	.65
Biological processes	bio	eat, blood, pain	748	.29	.71
Body	body	cheek, hands, spit	215	.52	.87
Health	health	clinic, flu, pill	294	.09	.37
Sexual	sexual	horny, love, incest	131	.37	.78
Ingestion	ingest	dish, eat, pizza	184	.67	.92
Drives	drives		1103	.39	.80
Affiliation	affiliation	ally, friend, social	248	.40	.80
Achievement	achieve	win, success, better	213	.41	.81
Power	power	superior, bully	518	.35	.76
Reward	reward	take, prize, benefit	120	.27	.69
Risk	risk	danger, doubt	103	.26	.68
Time orientations	TimeOrient				
Past focus	focuspast	ago, did, talked	341	.23	.64
Present focus	focuspresent	today, is, now	424	.24	.66
Future focus	focusfuture	may, will, soon	97	.26	.68
Relativity	relativ	area, bend, exit	974	.50	.86
Motion	motion	arrive, car, go	325	.36	.77
Space	space	down, in, thin	360	.45	.83
Time	time	end, until, season	310	.39	.79
Personal concerns					
Work	work	job, majors, xerox	444	.69	.93
Leisure	leisure	cook, chat, movie	296	.50	.86
Home	home	kitchen, landlord	100	.46	.83
Money	money	audit, cash, owe	226	.60	.90
Religion	relig	altar, church	174	.64	.91
Death	death	bury, coffin, kill	74	.39	.79
Informal language	informal		380	.46	.84
Swear words	swear	fuck, damn, shit	131	.45	.83
Netspeak	netspeak	btw, lol, thx	209	.42	.82
Assent	assent	agree, OK, yes	36	.10	.39
Nonfluencies	nonflu	er, hm, umm	19	.27	.69
Fillers	filler	I mean, you know	14	.06	.27

Appendix D Draft Paper

Below the draft paper in preparation for submission to a scientific conference or journal.

Deep & Dutch: Exploring linguistic markers in psychiatric stories

Stephanie Verkleij¹, Marco Spruit^{1,2,3}[0000-0002-9237-221X], Kees de Schepper⁴,
and Floortje Scheepers⁴[0000-0002-1801-6153]

¹ Department of Information and Computing Sciences, Utrecht University, 3584 CC Utrecht, The Netherlands

² Department of Public Health and Primary Care, Leiden University Medical Centre, 2511 DP, The Hague, The Netherlands; m.r.spruit@lumc.nl

³ Leiden Institute of Advanced Computer Science, Leiden University, 2333 CA, Leiden, The Netherlands

⁴ Department of Psychiatry, University Medical Centre Utrecht, Utrecht, The Netherlands

Abstract. Diagnosing mental disorders is complex due to the genetic, environmental and psychological contributors and the individual risk factors. Linguistic markers for mental disorders can help diagnose a person. Research thus far on linguistic markers and the associated mental disorders has been done mainly with the Linguistic Inquiry and Word Count program (LIWC). In order to improve on this research, we apply, next to LIWC, spaCy and the deep learning models fastText and RobBERT to Dutch psychiatric interview transcriptions. This is analysed with the goal to find out if a person has a mental disorder and if so which one. Furthermore, the second goal of this research is to find out which words are linguistic markers for which mental disorder. LIWC in combination with the classification algorithm random forest performed best in predicting if a person has a mental disorder (AUC: 0.888; accuracy: 0.952; Cohen’s kappa: 0.889) and spaCy in combination with random forest predicted best which mental disorder a person has (accuracy: 0.429; Cohen’s kappa: 0.304).

Keywords: Linguistic marker · Mental disorder · Deep learning · LIWC · spaCy · RobBERT · fastText · LIME

1 Introduction

The contribution of mental disorders are a major part of the global burden of disease [22] and in 2017 accounted for 10.7% of the global population [19]. This contribution is not keeping an even position, but is rising mainly in developing countries [22]. Furthermore, mental disorders have a substantial long term impact on individuals, caregivers and society [13]. The challenge of diagnosing a mental disorder is the complexity of the multiple genetic, environmental and psychological contributors and the individual risk factors [20].

Research has shown that people with mental health difficulties use distinctive linguistic patterns [12]. The program Language Inquiry and Word Count (LIWC) is the main focus for identifying linguistic markers [2]. This technique calculates the number of words of certain categories that are used in a text based on a dictionary [16]. LIWC is a traditional program that analyses at word level and it does not use neural networks. The goal of this research is to compare LIWC with other NLP techniques to provide more useful insights into psychiatric stories.

In this paper we compare the performance of spaCy [10], fastText [1] and RobBERT [7] when applied to the psychiatric interview transcriptions. spaCy is a dependency parser and syntactically processes text. This technique can provide insights because it will show the grammatical structure of the sentences and it will provide information about the grammatical relationships between words [6]. By using this technique, the different uses of grammar between mental illnesses are uncovered. This will give further insight into the stylistic differences between people with and without a mental disorder. fastText and RobBERT are chosen because both techniques employ deep learning models. Deep learning exploits layers of non-linear information processing for both supervised and unsupervised tasks [8]. We hypothesise that deep learning techniques could provide for more insight into these complex disorders.

2 Related work

This research is not the first looking into linguistic markers of people with a mental disorder. A few researchers compared mental disorders using the LIWC-tool [4] [12]. Table 1 summarises ten different mental disorders and the highlights of their use of language. This includes mainly their pronoun use, semantic coherence (SC) and word use.

People with Attention Deficit Hyperactivity Disorder (ADHD) use more third person plural (3pl) pronouns, less words of relativity [4] and more sentences, but less clauses per sentence [11]. Autism had a strong display of motion, home, religion and death features [14]. Furthermore, people with autism are more self-focused, because they use more first person singular (1sg) pronouns [14]. People who are bipolar, are also more self-focused and use more words related to death [9]. The use of more swearing words, words related to death, third person singular (3sg) pronouns and less use of cognitive emotion words were associated with borderline personality disorder (BPD) [12]. Eating disorders, consisting of bulimia, anorexia and eating disorders not otherwise specified, are associated with the use of the words related to the body, negative emotion words, self-focused words and cognitive process words [4]. People with generalised anxiety disorder (GAD) produce more sentences which lack semantic coherence [17]. Furthermore, they also use more tentative words, impersonal pronouns and they use more words related to death and health [4]. Major depressive disorder (MDD) had a strong display of being more self-focused, using more past tense and repetitive words and producing more short, detached and arid sentences [21]. Obsessive compulsive disorder (OCD) is associated with words related to anxiety and more

cognitive words. Researchers do not yet agree on the linguistic cues of PTSD, short for post-traumatic stress disorder. One study showed that there were no cues [4] and another study showed that people with PTSD use more singular pronouns, words related to death and less cognitive words [15]. Finally, research shows that the lack of semantic cohesion [5], usage of words related to religion, hearing voices and sounds is associated with schizophrenia [12].

Table 1: Linguistic markers per disorder

Disorder	Pronoun	SC	Word use	More
ADHD	3pl	-	-	Relativity, more sentences, less clauses
Autism	1sg	-	Motion, home, religion and death	-
Bipolar	1sg	-	Death	-
BPD	3sg	Normal	Death	Swearing, less cognitive emotion words
Eating	1sg	-	Body	Negative emotion words
GAD	imprs	Impaired	Death and health	Tentative words
MDD	1sg	Impaired	-	Inverse word-order and repetitions
OCD	1sg	-	Anxiety	More cognitive words
PTSD	sg	-	Death	Less cognitive words
Schizophrenia	3pl	Impaired	Religion	Hearing voices and sounds

3 Methodology

3.1 Dataset and Preprocessing

The data used for this research is obtained from the Verhalenbank of the University Medical Centre Utrecht (UMCU) in the Netherlands. The Psychiatry department has been collecting stories about mental illness of people who have, had or were in contact with psychiatric complaints. Interviews are conducted with (ex-)patients, caregivers and medical employees to gain new leads which could benefit the recovery of patients. The interviews are then transcribed into anonymous stories and put on the website of the Verhalenbank⁵. The dataset consists of 108 interviews with 11 labels of which 36 are without mental disorder. The split used for this research is 80% training and 20% testing. Source code for the data analyses is available at: <https://github.com/StephanieVx/ExploringLinguisticMarkers>.

⁵ <https://psychiatrieverhalenbank.nl/>

3.2 Data analysis

This exploratory study compares performances of different NLP techniques and looks at which language cues could predict if a person has a mental disorder and if so, which kind of mental disorders. The four different techniques were applied to the two test cases. The first test case is mental disorder vs mental disorder and the second is mental disorder vs no mental disorder. After applying the techniques, predictions are made. For LIWC and spaCy, the classification algorithms decision tree, random forest and support vector machine (SVM) were used and the deep learning techniques used their default prediction models without incorporating a transfer learning step. Next, the techniques and predictions were applied again after removing the stop words from the interviews and the predictions were compared. Furthermore, to gain further insight in the predictions of fastText and RobBERT, LIME (Local Interpretable Model Agnostic Explanation) was applied [18].

4 Results

4.1 Predictions

Table 2 shows the accuracy of the two comparisons and Cohen’s Kappa per prediction. The LIWC program in combination with the random forest algorithm reached the highest accuracy when comparing mental disorder versus no mental disorder. spaCy reached the highest accuracy when comparing the different kinds of mental disorder. Cohen’s kappa was used to assess the inter-classifier agreement [3]. This metric takes the probability that the 10 different labels in this case, agree by chance into consideration when quantifying how much they agree. Cohen’s kappa was calculated for each model and prediction algorithm. Some of the models have classifiers below 0.4 which means that there is a slight agreement. A kappa of above 0.6 means that the classifiers have a substantial agreement, for example the n-grams input with the SVM model in the MD (mental disorder) vs Control group comparison. When the kappa is between 0.8 and 1 it means that the classifiers have an almost perfect agreement. This applies to the LIWC-output with the random forest model in the second comparison with a kappa of 0.889. The low accuracy of the second comparison can be explained due to a dataset of only 72 interviews from people with mental disorders and 10 different kinds of mental disorders.

What also can be seen in Table 2 in the sixth and seventh column, is that without stop words spaCy perform less accurate while LIWC, fastText and RobBERT perform almost similar in both comparisons.

Figure 1 shows the decision tree for the LIWC-output. If an interview transcription consisted of more than 5.4% of the first person singular pronoun, than it was classified as mental disorder. If not and if less than 8.5% of the words related to social concepts were included, than the interview was classified as no mental disorder. Furthermore, the plot in Figure 2 shows the AUC-ROC of LIWC with the random forest classification algorithm for the comparison between mental disorder and no mental disorder.

Table 2: Predictions

Comparison	Input	Model	Accuracy	Kappa	Accuracy no SW	Kappa no SW
Mental Disorder vs no Mental Disorder	LIWC-output	rpart	0.857	0.667	0.857	0.674
	LIWC-output	random-Forest	0.952	0.889	0.952	0.877
	LIWC-output	SVM	0.857	0.64	0.905	0.738
	n-grams	rpart	0.810	0.391	0.444	-0.309
	n-grams	random-Forest	0.762	0.173	0.389	-0.370
	n-grams	SVM	0.714	0.115	0.528	-0.275
	raw data	fastText	0.643	0.172	0.607	0.072
	raw data	RobBERT	0.607	0.000	0.607	0.000
Mental Disorder multiclass	LIWC-output	rpart	0.286	0.157	0.286	0.177
	LIWC-output	random-Forest	0.214	0.120	0.214	0.144
	LIWC-output	SVM	0.286	0.114	0.143	0.0718
	n-grams	rpart	0.143	-0.0120	0.071	-0.052
	n-grams	random-Forest	0.429	0.304	0.214	0.078
	n-grams	SVM	0.357	0.067	0.143	0.091
	raw data	fastText	0.286	0.000	0.200	0.000
	raw data	RobBERT	0.200	0.000	0.267	0.120

4.2 LIME

LIME was applied to both fastText and RobBERT to gain further insight into the black-box models. For example, quote 1 is from someone who has been diagnosed with schizophrenia and the text is labelled by RobBERT as a mental disorder. The word 'eh' has been highlighted because it explains according to LIME why it was labelled as mental disorder (class = 0). In Figure 3, the ten words with the highest probabilities can be seen. Some words appear multiple times in the figure and this is because it looks locally at a text and every word appears in a different context. This also means that sometimes a word will be an explanation for a mental disorder and other times not, based on the context. The second quote is from someone with an eating disorder and analysed by fastText. The word 'Eh' was highlighted because it explained why the transcription was

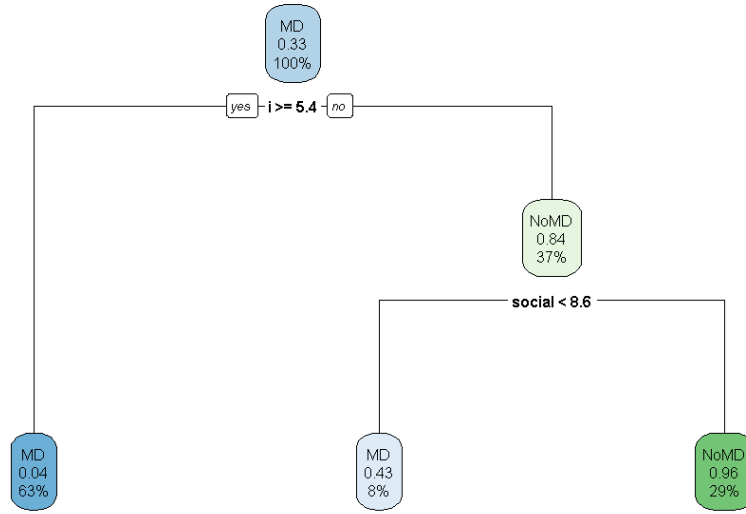


Fig. 1. LIWC decision tree

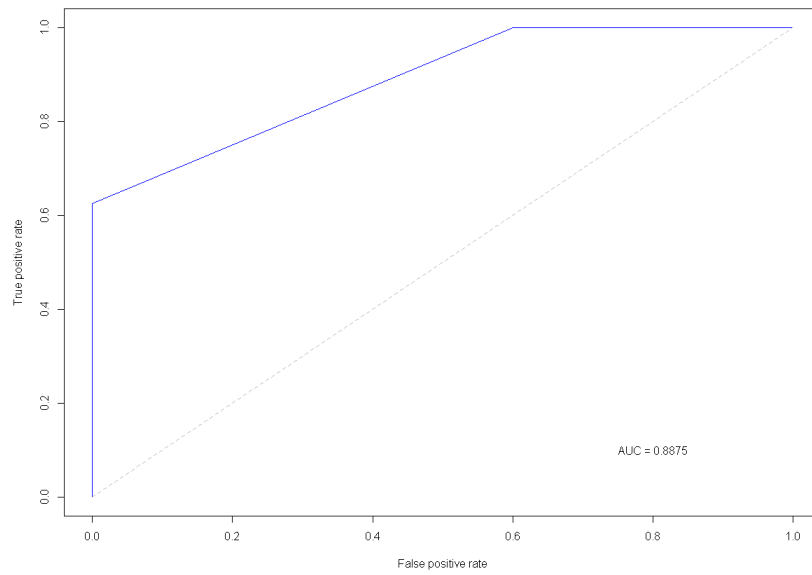


Fig. 2. LIWC random forest

labelled as a mental disorder (class = _label_md). Figure 4 shows the ten words with the highest probabilities from that transcription.

Quote 1: "I ehm, [silence] the most poignant I will you- Yes, the most poignant what I can tell you is that, I have weekend leave on the weekend and than [name][wife] and I lay together in bed. And nothing happens there. Because I don't need that, haha. But I can't even feel that I love her. I know it, that I love her. And I know that my wife is and I, and I. But that's all in here eh, but I don't feel it. And that is the biggest measure which you can set.. Yes. And I talked about it with her. "

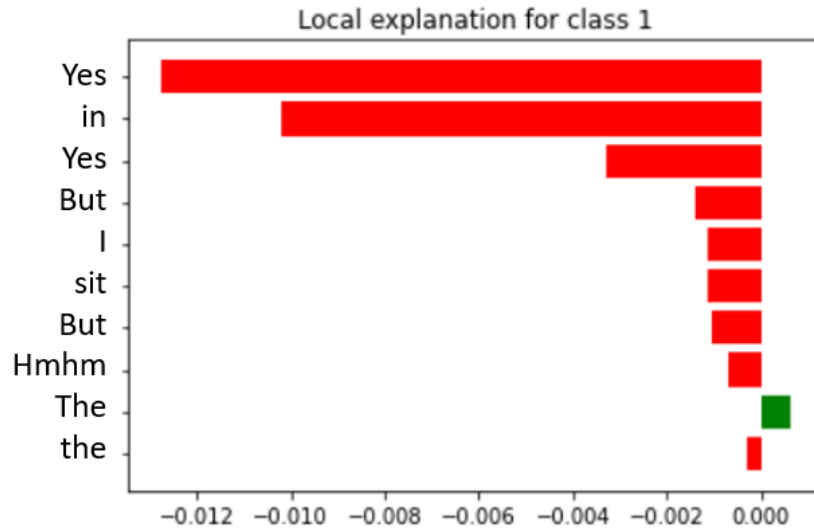


Fig. 3. LIME explanation quote 1

Quote 2: "Yes it gives kind of a kick or something to go against it and to see that people you really eh yes I don't know . That your that your eating disorder is strong and people find that than. And then you think oh I am good at something . And than yes I don't know. Than you want there that you want to be doing something you are good at . . Eh I am able to walk again since two months. Before I eh stayed in bed and in a wheelchair around half a year, because I eh could not walk myself. And I was just to weak to do it. and eh yes I still cannot do quite a lot of things. I am really happy that I can walk again by myself."

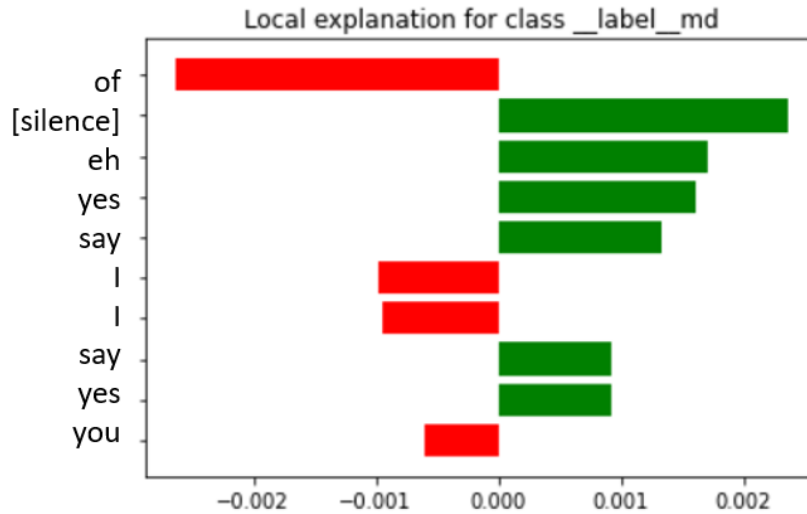


Fig. 4. LIME explanation quote 2

5 Discussions and Conclusion

We explored the linguistic markers in Dutch psychiatric interview transcriptions. We particularly focused on comparing the performances of LIWC, spaCy, fast-Text and RobBERT, to find linguistic markers for the different kinds of mental disorders. The best performing technique to find out if a person has a mental disorder is LIWC in combination with the classification algorithm random forest which reached an AUC of 0.888, an accuracy-score of 0.952 and a Cohen's kappa of 0.889. spaCy in combination with random forest predicted best which mental disorder a person has with an accuracy-score of 0.429 and a Cohen's kappa of 0.304. The moderate accuracy score could be explained due to the fact that the dataset of people with a mental disorder only included 72 interview transcriptions and 10 mental disorder labels. Furthermore, stop words do not appear to have that much influence on the performance of the classifications except when employed using spaCy. We presume that is may be due to spaCy analyses the text from a grammatical point of view and if stop words are missing it cannot form the correct syntactic dependencies anymore. Further work will focus on exploring more advanced methods to explain blackbox-models and investigating alternative NLP-models in combination with an expanded data collection.

References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* **5**, 135–146 (2017)
2. Calvo, R.A., Milne, D.N., Hussain, M.S., Christensen, H.: Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering* **23**(5), 649–685 (2017)
3. Cohen, J.: A coefficient of agreement for nominal scales. *Educational and psychological measurement* **20**(1), 37–46 (1960)
4. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K.: From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In: *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. pp. 1–10 (2015)
5. Corcoran, C.M., Cecchi, G.: Using language processing and speech analysis for the identification of psychosis and other disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* (2020)
6. Davcheva, E.: Text mining mental health forums - learning from user experiences. In: Bednar, P.M., Frank, U., Kautz, K. (eds.) *26th European Conference on Information Systems: Beyond Digitization - Facets of Socio-Technical Change, ECIS 2018, Portsmouth, UK, June 23-28, 2018*. p. 91 (2018), https://aisel.aisnet.org/ecis2018_rp/91
7. Delobelle, P., Winters, T., Berendt, B.: Robbert: a dutch roberta-based language model. *arXiv preprint arXiv:2001.06286* (2020)
8. Deng, L., Yu, D., et al.: Deep learning: methods and applications. *Foundations and Trends® in Signal Processing* **7**(3–4), 197–387 (2014)
9. Forgeard, M.: Linguistic styles of eminent writers suffering from unipolar and bipolar mood disorder. *Creativity Research Journal* **20**(1), 81–92 (2008)
10. Honnibal, M., Johnson, M.: An improved non-monotonic transition system for dependency parsing. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. pp. 1373–1378. Association for Computational Linguistics, Lisbon, Portugal (September 2015), <https://aclweb.org/anthology/D/D15/D15-1162>
11. Kim, K., Lee, S., Lee, C.: College students with adhd traits and their language styles. *Journal of attention disorders* **19**(8), 687–693 (2015)
12. Lyons, M., Aksayli, N.D., Brewer, G.: Mental distress and language use: Linguistic analysis of discussion forum posts. *Computers in Human Behavior* **87**, 207–211 (2018)
13. McIntosh, A.M., Stewart, R., John, A., Smith, D.J., Davis, K., Sudlow, C., Corvin, A., Nicodemus, K.K., Kingdon, D., Hassan, L., et al.: Data science for mental health: a uk perspective on a global challenge. *The Lancet Psychiatry* **3**(10), 993–998 (2016)
14. Nguyen, T., Phung, D., Venkatesh, S.: Analysis of psycholinguistic processes and topics in online autism communities. In: *2013 IEEE International Conference on Multimedia and Expo (ICME)*. pp. 1–6. IEEE (2013)
15. Papini, S., Yoon, P., Rubin, M., Lopez-Castro, T., Hien, D.A.: Linguistic characteristics in a non-trauma-related narrative task are associated with ptsd diagnosis and symptom severity. *Psychological Trauma: Theory, Research, Practice, and Policy* **7**(3), 295 (2015)

16. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates **71**(2001), 2001 (2001)
17. Remmers, C., Zander, T.: Why you don't see the forest for the trees when you are anxious: Anxiety impairs intuitive decision making. *Clinical Psychological Science* **6**(1), 48–62 (2018)
18. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
19. Ritchie, H., Roser, M.: Mental health. *Our World in Data* (2020), <https://ourworldindata.org/mental-health>
20. Russ, T.C., Woelbert, E., Davis, K.A., Hafferty, J.D., Ibrahim, Z., Inkster, B., John, A., Lee, W., Maxwell, M., McIntosh, A.M., et al.: How data science can advance mental health research. *Nature human behaviour* **3**(1), 24–32 (2019)
21. Trifu, R.N., NEMEŞ, B., BODEA-HAŢEGAN, C., Cozman, D.: Linguistic indicators of language in major depressive disorder (mdd). an evidence based research. *Journal of Evidence-Based Psychotherapies* **17**(1) (2017)
22. Whiteford, H.A., Degenhardt, L., Rehm, J., Baxter, A.J., Ferrari, A.J., Erskine, H.E., Charlson, F.J., Norman, R.E., Flaxman, A.D., Johns, N., et al.: Global burden of disease attributable to mental and substance use disorders: findings from the global burden of disease study 2010. *The lancet* **382**(9904), 1575–1586 (2013)