applied sciences

MDPI

Article

Exploring Language Markers of Mental Health in Psychiatric Stories

Marco Spruit [1,2,*], Stephanie Verkleij [3], Kees de Schepper [4] and Floortje Scheepers [4]

[1] Leiden University Medical Center (LUMC), Campus The Hague, Leiden University, Turfmarkt 99, 2511 DC The Hague, The Netherlands
[2] Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
[3] Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands; stephanieverkleij@hotmail.com
[4] University Medical Center Utrecht (UMCU), Utrecht University, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands; c.w.m.deschepper@umcutrecht.nl (K.d.S.); f.e.scheepers-2@umcutrecht.nl (F.S.)
* Correspondence: m.r.spruit@lumc.nl

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 to diagnose a person. Research thus far on language markers and the associated mental disorders has been done mainly with the Linguistic Inquiry and Word Count (LIWC) program. In order to improve on this research, we employed a range of Natural Language Processing (NLP) techniques using LIWC, spaCy, fastText and RobBERT to analyse Dutch psychiatric interview transcriptions with both rule-based and vector-based approaches. Our primary objective was to predict whether a patient had been diagnosed with a mental disorder, and if so, the specific mental disorder type. Furthermore, the second goal of this research was to find out which words are language markers for which mental disorder. LIWC in combination with the random forest classification algorithm performed best in predicting whether a person had a mental disorder or not (accuracy: 0.952; Cohen's kappa: 0.889). SpaCy in combination with random forest predicted best which particular mental disorder a patient had been diagnosed with (accuracy: 0.429; Cohen's kappa: 0.304).

Keywords: language marker; mental disorder; deep learning; LIWC; spaCy; RobBERT; fastText; LIME

Citation: Spruit, M.; Verkleij, S.; de Schepper, K.; Scheepers, F. Exploring Language Markers of Mental Health in Psychiatric Stories. *Appl. Sci.* **2022**, *12*, 2179. https://doi.org/10.3390/app12042179

Academic Editors: Arturo Montejo-Ráez and Salud María Jiménez-Zafra

Received: 21 September 2021
Accepted: 15 February 2022
Published: 19 February 2022

1. Introduction

Mental disorders make up a major portion of the global burden of disease [1], and in 2017, 10.7% of the global population reported having or having had a mental disorder [2]. This prevalence is not staying stead, but is rising mainly in developing countries [1]. Furthermore, mental disorders have a substantial long term impact on individuals, caregivers and society [3]. The challenge of diagnosing a mental disorder is the complexity of multiple genetic, environmental and psychological contributors and individual risk factors [4].

Research has shown that people with mental health difficulties use distinctive language patterns [5]. Until now, the Language Inquiry and Word Count (LIWC) toolkit has been the main focus for identifying language markers [6]. This toolkit of Natural Language Processing (NLP) techniques calculates the number of words of certain categories that are used in a text based on a dictionary [7]. LIWC is a traditional programme in the sense that it analyses texts with symbolic (i.e., deterministic and rule-based) techniques, predominantly at the word level. LIWC does not use subsymbolic (i.e., probabilistic and vector-based) NLP techniques such as word vector representations within neural networks.

The objective of our research was to compare the performance of LIWC with the performances of other NLP techniques in the quest to provide useful insights into Dutch psychiatric stories. In this paper, we compare the performances of LIWC [6], spaCy [8],

fastText [9] and RobBERT [10] when applied to psychiatric interview transcriptions. SpaCy provides, among other things, a dependency grammar parser to syntactically process texts. This NLP technique can provide insights by unravelling the grammatical structure of each sentence, and it will provide information about the grammatical relationships between words [11]. By using this technique, we aimed to uncover the different uses of grammar by patients with different mental illnesses. This provides further insights into the stylistic differences between people with and without mental disorders. fastText and RobBERT were selected because both techniques employ deep learning models. Deep learning exploits layers of non-linear information processing for both supervised and unsupervised tasks [12]. We hypothesise that deep learning techniques can provide more insights than other methods into these complex mental health disorders.

2. Related Work

This research is not the first to attempt to identify language markers associated with mental disorders. Several researchers already compared mental disorders using the LIWC tool [5,13]. We introduce and compare several state-of-the-art alternative NLP approaches to identifying language markers' associations with mental health disorders.

2.1. Language Markers for Mental Health Disorders

A literature study was performed to review earlier work related to language markers for mental health disorders. The snowballing method was used to find the relevant literature. Both backward snowballing and forward snowballing were employed [14]. A curated set of recent papers on language markers in mental healthcare was used as the starting point [5,6,13,15,16]. Then, one or two levels deep were snowballed back and forth. The number of levels snowballed depended on whether new relevant literature was found. Whenever a dead end was reached, the snowballing procedure was stopped. We selected Google Scholar (with a proxy from Utrecht University) to execute the following search queries:

- "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

Table 1 summarises our findings related to ten different mental disorders, highlighting their uses of language. These include mainly characteristic use of pronouns (Pron), the degree ([n]ormal/[i]mpaired) of semantic coherence (SC) and usage of topical words. We only list the disorders that appear in our dataset as the main diagnosis; the N column shows the number of patients.

We found that people with attention deficit hyperactivity disorder (ADHD) use more third-person plural (3pl) pronouns, less words of relativity [13] and more sentences, but less clauses per sentence [17] than normal. Autism is strongly linked to motion, home, religion

and death features [18]. Furthermore, people with autism are more self-focused, because they use more first-person singular (1sg) pronouns [18]. People who are bipolar are also more self-focused and use more words related to death [19]. The use of more swear words, words related to death and third-person singular (3sg) pronouns, and less use of cognitive emotive words are associated with borderline personality disorder (BPD) [5]. 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 emotive words, self-focused words and cognitive process words [13]. People with generalised anxiety disorder (GAD) produce more sentences which lack semantic coherence [20]. Furthermore, they use more tentative words and impersonal pronouns, and they use more words related to death and health [13]. Major depressive disorder (MDD) has a strong appearance of being more self-focused, involving more past tense and repetitive words and producing short, detached and arid sentences [21]. Obsessive compulsive disorder (OCD) is associated with words related to anxiety and cognitive words. Researchers do not yet agree on the language cues associated with post-traumatic stress disorder (PTSD). One study showed that there were no cues [13], yet another study showed that people with PTSD use more singular pronouns and words related to death and less cognitive words [22]. Finally, research shows that a lack of semantic cohesion [23], usage of words related to religion and hearing voices and sounds are associated with schizophrenia [5]. Further details are available in [24].

Table 1. Overview of associated language markers for ten mental health disorders.

Disorder	Pron	SC	Word Use	More	N
ADHD	3pl	-	-	Relativity, more sentences, less clauses	4
Autism	1sg	-	Motion, home, religion and death	-	5
Bipolar	1sg	-	Death	-	7
BPD	3sg	n	Death	Swearing, less cognitive emotion words	5
Eating	1sg	-	Body	Negative emotion words	10
GAD	imprs	i	Death and health	Tentative words	4
MDD	1sg	i	-	Inverse word-order and repetitions	11
OCD	1sg	-	Anxiety	More cognitive words	4
PTSD	sg	-	Death	Less cognitive words	6
Schizophrenia	3pl	i	Religion	Hearing voices and sounds	16

2.2. NLP Techniques for Identifying Language Markers

We investigated the following four basic approaches in NLP for identification of language markers: lexical processing from a lexical semantics perspective, dependency parsing from a compositional semantics viewpoint, shallow neural networks in a stochastic paradigm and deep neural networks employing a transformer-based architecture.

2.2.1. Lexical Processing

Research so far on exploring language markers in mental health has been done mainly with Linguistic Inquiry and Word Count (LIWC) [6]. LIWC is a computerised text-analysis tool and has two central features: a processing component and dictionaries [15]. 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. For example, the word "had" can be put in the categories verbs, auxiliary verbs and past tense verbs. Next, the program calculates the percentage for each category in the text; for example, 17% of the words may be verbs. A disadvantage of the LIWC program is that it ignores context, idioms, sarcasm and irony. Furthermore, the 89 different categories are based on language research. However, this does not guarantee that these categories represent reality, because categories could be missing.

2.2.2. Dependency Parsing

The syntactic processing of texts is called dependency parsing [25]. This processing is valuable because it forms transparent lexicalised representations and it is robust [25]. Furthermore, it also gives insights into the compositional semantics, i.e., the meanings of a sentence's individual words or phrases [26]. Small changes in the syntactic structure of a sentence can change the whole meaning of the sentence. For example, *John hit Mary* and *Mary hit John* contain the same words, but have different meanings. It is said that compositionality is linked to our ability to interpret and produce new remarks, because once one has mastered the syntax of a language, its lexical meanings and its modes of composition, one can interpret new combinations of words [27]. Compositionality is the semantic relationship combined with a syntactic structure [28]. Compositional semantics is driven by syntactic dependencies, and each dependency forms, from the contextualised sense of the two related lemmas, two new compositional vectors [29]. Therefore, the technique required for extracting the compositional semantics needs to contain a dependency parser and a lemmatizer. Choi et al. [25] compared the ten leading dependency parsers based on the speed/accuracy trade-off. Although Mate [30], RBG [31] and ClearNLP [32] perform best in unlabeled attachment score (UAS), none of them includes a Dutch dictionary, which was needed for this research. However, spaCy does include a Dutch dictionary. Other Dutch dependency parsers are Frog [33] and Alpino [34]. Both Frog (https://github.com/LanguageMachines/frog/releases/, accessed on 17 October 2021) and spaCy (https://spaCy.io/models/nl, accessed on 17 October 2021) include the Dutch dictionary corpus of Alpino, but due to equipment constraints, we selected spaCy for the dependency parsing task.

2.2.3. Shallow Neural Networks

Features made for traditional NLP systems are frequently handcrafted, time consuming and incomplete [35]. Neural networks, however, can automatically learn multilevel features and give better results based on dense vector representations [16]. The trend toward neural networks has been caused by the success of deep learning applications and the concept of word embeddings [16]. Word embeddings, such as the skip-gram model and the continuous bag-of-words (CBOW) model [36], distribute high-quality vector representations and are often used in deep learning models as the first data processing layer [16]. The word2vec algorithm uses neural networks to learn vector representations [37]. It can use the skip-gram model or the CBOW model, and it works for both small and large datasets [37]. However, out-of-vocabulary (OOV) words, also referred to as unknown words, are a common issue for languages with large vocabularies [16]. The fastText model overcomes this problem by handling each word as a bag-of-character n-gram. This is achieved by using the skip-gram model from word2vec as an extension. These n-grams are used to represent the sums of the n-gram vectors [9]. Finally, it is worth noting that both Word2vec and fastText are said to employ a shallow neural network architecture; i.e., their neural networks only define one hidden layer, which explains why these models are known to be many orders of magnitude faster in training and evaluation than other deep learning classifiers, while often performing as well as those classifiers in terms of accuracy [38].

2.2.4. Deep Neural Networks

In 2017 the transformer neural network architecture was introduced [39], which much improved NLP tasks such as text classification and language understanding [40]. Bidirectional encoder representations from transformers (BERT) is an immensely popular transformer-based language representation model designed to pretrain, from unlabelled text, deep bidirectional representations [41]. The multilingual version of BERT is simply called mBERT. A more recent and improved version of BERT is RoBERTa, which stands for robustly optimised BERT approach [42]. The main changes are that RoBERTa trains for longer, on more data, with bigger batches and on longer sequences [42].

2.2.5. Neural Networks for Dutch

In Table 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 were chosen for this research, one based on words and one based on sentences. Furthermore, the neural networks had to have a Dutch model. Thus, the choice was between word2vec and fastText at the word-level and between BERT, mBERT and RoBERTa at the sentence level. Other models, such as ClinicalBERT, could also be used in combination with a transfer learning model such as the Cross-lingual Language Model (XLM) to tackle the Dutch data. However, these models have not yet been used extensively in the medical domain [43]. This could be because the interpretability and performance of a model are equally important in the medical domain. Even though deep learning models can perform better than the more traditional models, they are hard to explain or understand [44]. Hence, this approach was not used for this research. Furthermore, fastText has proven that it results in better performance in comparison to Word2vec [45] and it is able to handle OOV words as well, because of the n-grams.

Table 2. Overview of neural network models under consideration for identifying language markers in Dutch.

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/RobBERT	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

The Dutch version of BERT is called BERTje [46], the Dutch version of RoBERTa is called RobBERT [10] and mBERT is the multilingual BERT with support for more than 100 languages, including Dutch [41]. A choice between the three BERTs was made by looking at their performances with respect to the classification task, because that was the focus of this research. The research of Delobelle et al. [10] shows that RobBERT (ACC = 95.1%) performs best on classification tasks compared to mBERT (ACC = 84.0%) and BERTje (ACC = 93.0%) with a full dataset. Therefore, the neural networks selected for this research were fastText and RobBERT.

3. Methodology

3.1. Dataset and Preprocessing

The dataset used for this research was obtained from the Verhalenbank ("Storybank") of the University Medical Centre Utrecht (UMCU) in The Netherlands. Its psychiatry department has been collecting stories about mental illness of people who have or had psychiatric issues or were in contact with people with psychiatric issues. Interviews were conducted with (ex-)patients, caregivers and medical employees to gain new leads which could benefit the recovery of patients. The interviews were then transcribed into anonymous stories and put on the website of the Verhalenbank (https://psychiatrieverhalenbank.nl/, accessed on 17 October 2021). The dataset consists of 108 interviews with 11 diagnostic

labels; 36 are without mental disorder labels. The diagnoses were assigned by multiple doctors and based on other material than the interviews. The interviews were all between 60 and 90 min long, and the corresponding transcripts are between 6782 and 9531 words in length. The split used for this research was 80% training and 20% testing. There were not enough data to have a validation set. Source code for the data analysis is available at: https://github.com/StephanieVx/ExploringLinguisticMarkers, accessed on 17 October 2021.

3.2. Data Analysis

This exploratory study compares the classification 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 disorder. The four different techniques were applied to the two tests. The first test consisted of deciding between mental disorder and no mental disorder; and the second one consisted of deciding between the different mental disorders. After applying the techniques, predictions were made. For LIWC and spaCy, the classification algorithms decision tree, random forest and support vector machine (SVM) were used by means of the default configurations of the R packages *rpart*, *randomForest* and *e1071*, respectively. The deep learning techniques used their default prediction models without incorporating a transfer learning step [47]. Next, the techniques and predictions were applied again after removing the stop words, as listed in the Dutch portion of the NLTK Python package [48], after which the interviews and the predictions were compared. Furthermore, to gain further insight into the predictions of fastText and RobBERT, LIME (Local Interpretable Model Agnostic Explanation) was applied [49].

4. Results

4.1. Descriptive Statistics

An overview of the number of people per mental disorder in our dataset is shown in Figure 1. The group with dissociation (a disconnection between a person's memories, feelings, perceptions and sense of self) contains the least number of people in this dataset; the group with psychosis is the largest. Furthermore, there are two labels about personality. Personality includes obsessive-compulsive personality disorder, avoidant personality disorder, dependent personality disorder and unspecified personality disorders. Personality+ in this research only includes borderline personality disorder (BPD). Figure 2 shows a boxplot of the number of words per mental disorder, which indicates that people with eating disorders use less words than people without eating disorders.

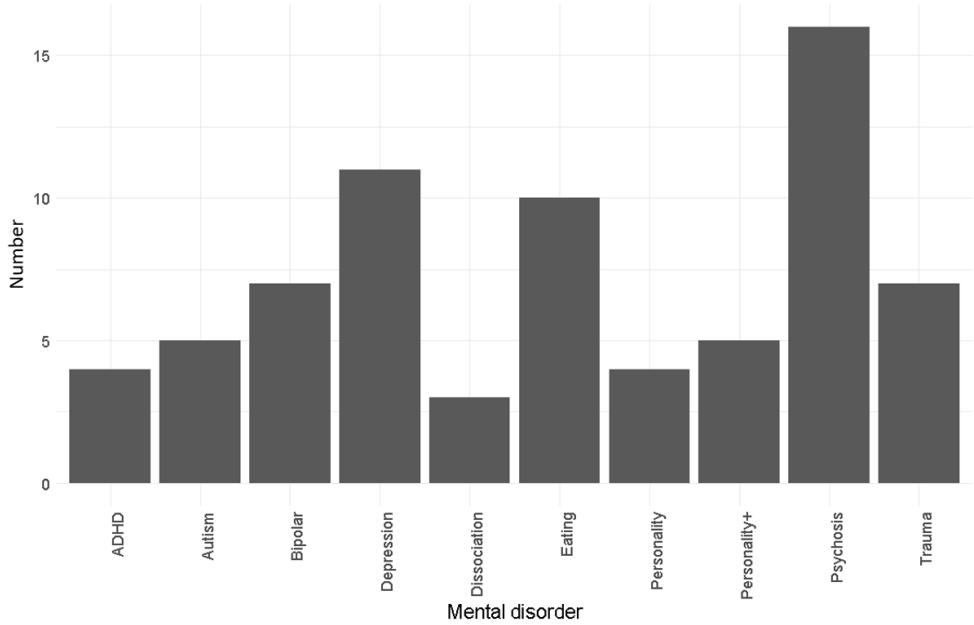

Figure 1. Columnchart of number of people per mental disorder in the dataset.

4.2. Predictions

Table 3 shows the accuracies in the two tests and Cohen's Kappa per prediction. The best performing classifiers are highlighted in bold text. The LIWC program in combination with the random forest algorithm achieved the highest accuracy when comparing mental disorder to no mental disorder (accuracy: 0.952). SpaCy reached the highest accuracy when comparing the different kinds of mental disorder (accuracy: 0.429).

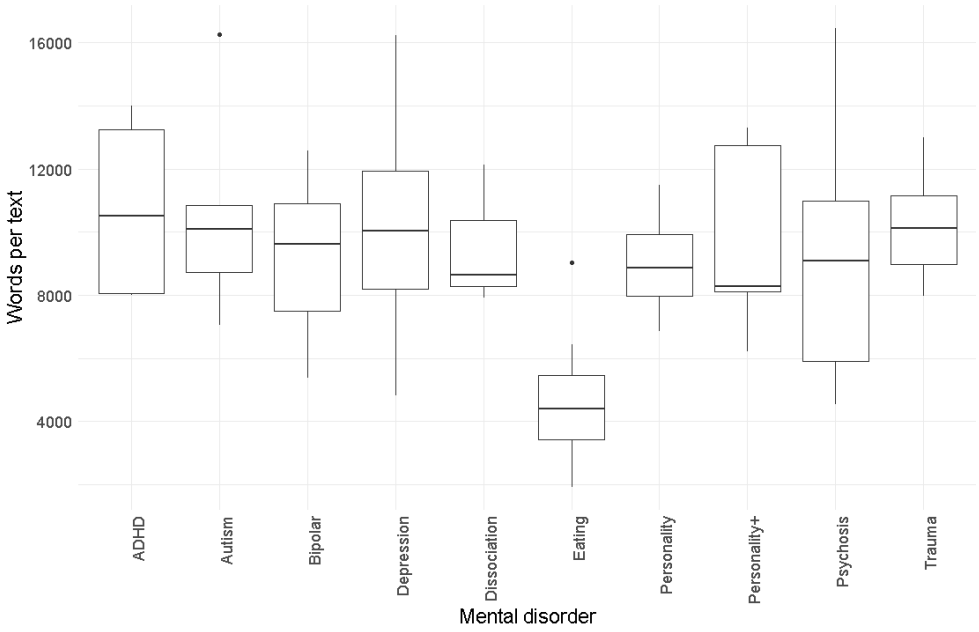

Figure 2. Boxplot of number of words per mental disorder in the dataset.

Cohen's kappa was used to assess the inter-classifier agreement [50]. 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. If the coefficient is below 0.4, there is a slight correlation between the models (and with a negative kappa it is even below chance level). A kappa of above 0.6 means that the classifiers have a substantial agreement; for example, see the LIWC-output with the SVM model in the MD (mental disorder) vs. control group comparison. When the kappa is between 0.8 and 1.0, this indicates that the classifiers have almost perfect agreement. This applies to the LIWC-output with the random forest model in the second comparison with a kappa of 0.889. Care should be taken when interpreting Cohen's kappa [51], but the fact that the item with the highest kappa also has the highest accuracy is reassuring. The low accuracy of the second comparison can be explained due to a dataset having only 72 interviews from people with mental disorders and 10 different kinds of mental disorders.

What also can be seen in Table 3 in the sixth and seventh columns is that without stop words spaCy performed less accurately, while LIWC, fastText and RobBERT performed almost the same in both comparisons.

Table 3. Accuracy and Cohen's Kappa for the model predictions (with and without stop words).

Comparison	Input	Model	Accuracy	Kappa	Accuracy No Stopwords	Kappa No Stopwords
Mental Disorder vs. No Mental Disorder	LIWC-output	decision tree	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
	spaCy	decision tree	0.810	0.391	0.444	−0.309
	spaCy	random-Forest	0.762	0.173	0.389	−0.370
	spaCy	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	decision tree	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
	spaCy	decision tree	0.143	−0.0120	0.071	−0.052
	spaCy	random-Forest	**0.429**	**0.304**	0.214	0.078
	spaCy	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

4.3. Interpretation

In this section, we elaborate on our findings regarding the performances of the LIWC, SpaCy, fastText and RobBERT approaches to NLP for language marker identification.

4.3.1. Lexical Processing with LIWC

Figure 3 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 being of a person with a mental disorder. If not and if less than 8.5% of the words were related to social concepts, then the interview was classified as being of a person with no mental disorder. Furthermore, the decision tree categories of the LIWC tool were visualised in a stripchart (jitter) plot, a fragment of which is shown in Figure 4. In particular, this plot effectively illustrates the potential to identify people with and without a mental disorder based on the empirical frequencies of hypothesised LIWC category occurrences, such as first-person singular pronoun (1sg), further strengthening the rationale behind this feature being the root decision of the LIWC decision tree shown in Figure 3.

Furthermore, we investigated the LIWC's feature importance using a random forest classifier to determine which variables added the most value to our binary predictions. Figure 5 shows the top 10 variables that impacted the classification.

4.3.2. Dependency Parsing with SpaCy

Similarly, we investigated the SpaCy feature importance using a random forest classifier to determine which n-grams added the most value to our binary predictions. Figure 6 shows the top 10 variables that impact the classification. In addition, we present the mean, standard deviation (sd) and standard error (se) for each n-gram in Figure 7. A Mann–Whitney U test revealed no significant difference between people with and without mental disorders in their usage of the following four spaCy variables: denken_denken_ROOT, gaan_gaan_ROOT, ja_ja_ROOT and zijn_zijn_ROOT. Finally, we provide example sentences for each of the identified SpaCy language markers in Table 4.

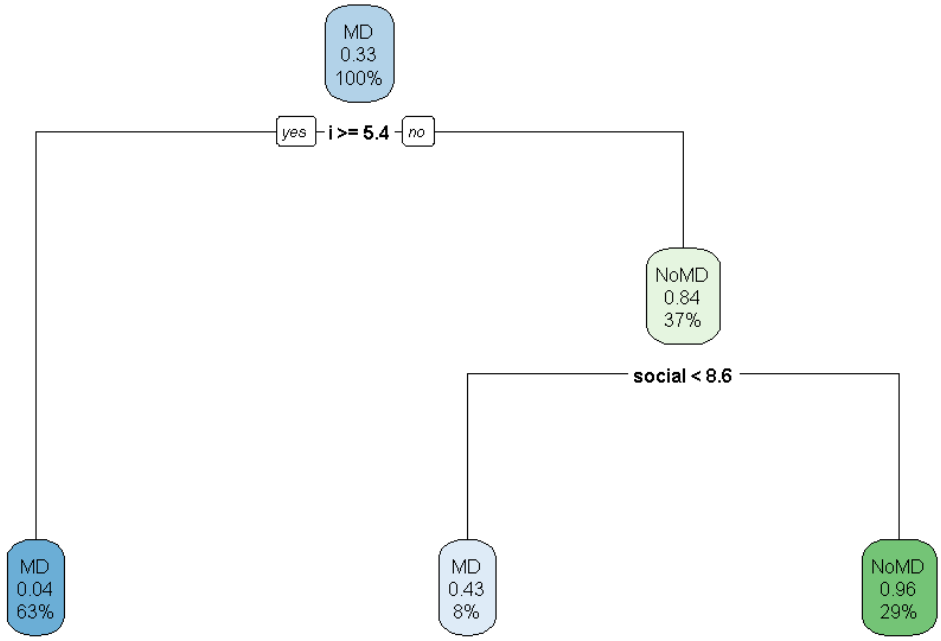

Figure 3. Example decision tree with two LIWC parameters (parameter *i* means the percentage of first-person pronouns and parameter *social* the percentage of words referring to others, such as *they*; each box lists the choice between mental disorder or not, the chance of the class being no mental disorder and the percentage of the data that fall in this box).

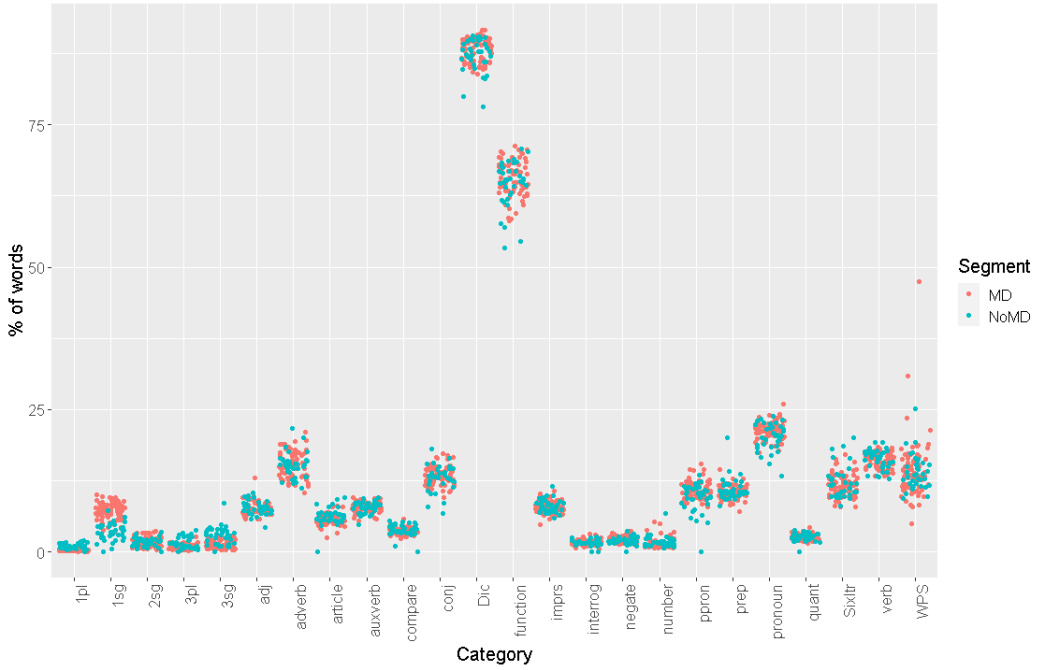

Figure 4. This stripchart plot illustrates the potential to identify people with and without a mental disorders based on the empirical frequencies of hypothesised LIWC category occurrences, e.g., first-person singular pronoun (1sg).

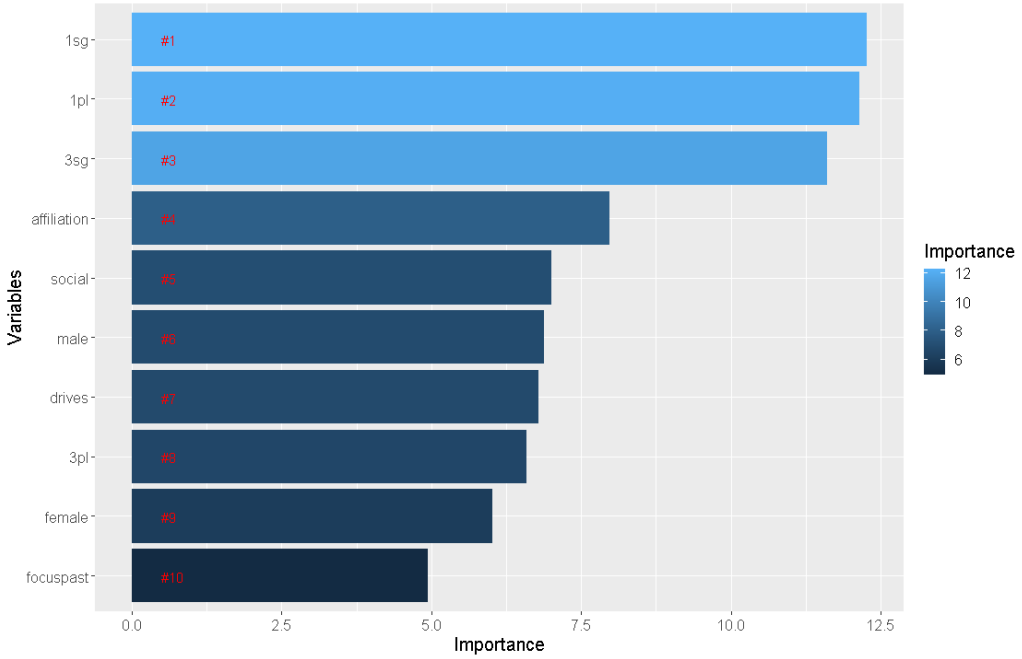

Figure 5. Top 10 LIWC features by importance in binary classification.

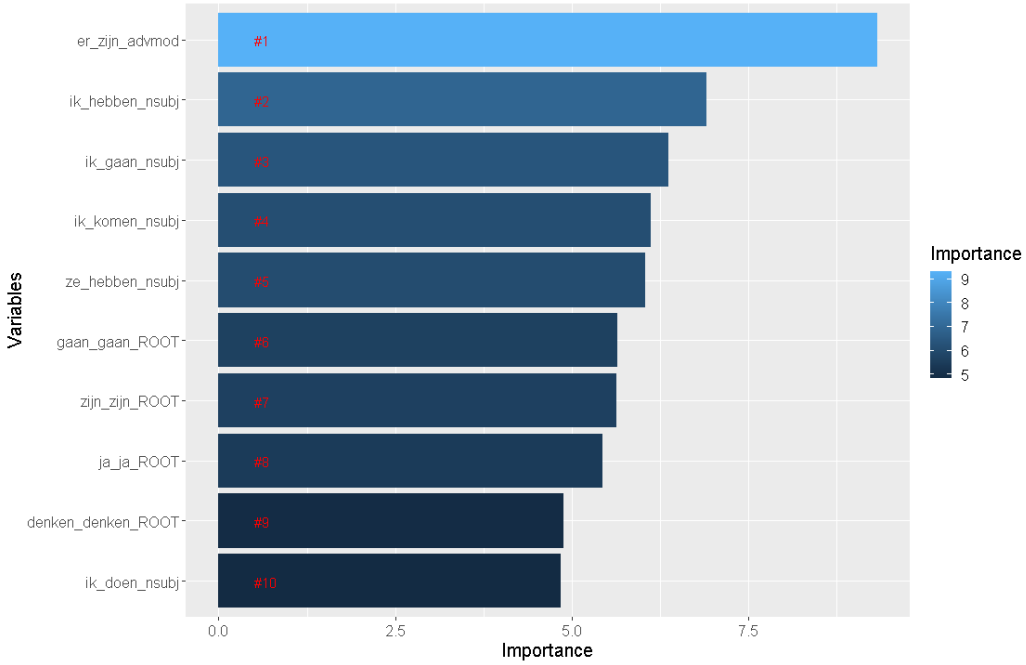

Figure 6. Top 10 SpaCy features by importance in binary classification.

4.3.3. Neural Networks with fastText and RobBERT

LIME was applied to both fastText and RobBERT to gain further insight into the black-box neural network models. LIME is a well-known and well-understood surrogate model-based approach to help explain model predictions by learning surrogate models using an operation called input perturbation [49]. For each sentence, subsamples of words were generated and fed to the model, so for each word the predictions for subsamples with and without this word could be compared, and subsequently the contribution of this word could be assessed. For example, quote 1 was from someone who had been diagnosed with schizophrenia, and the text was labelled by RobBERT as mental disorder. The word "eh" has been highlighted because it explains according to LIME why it was labelled as mental disorder (class = 0). Note that the original quote is in Dutch, but for convenience we

provide English translations here. In addition, "[silence]" means a pause that was judged as meaningful by the transcriber of the interview. In Figure 8, the ten words with the highest usage can be seen. Some words appear multiple times in the figure. This is because LIME 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, especially for context sensitive algorithms like RobBERT.

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 7. Top 10 SpaCy n-gram features in binary classification.

Table 4. Example sentences containing the top 6 spaCy variables.

spaCy Variable	Example Sentence
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 do not 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'

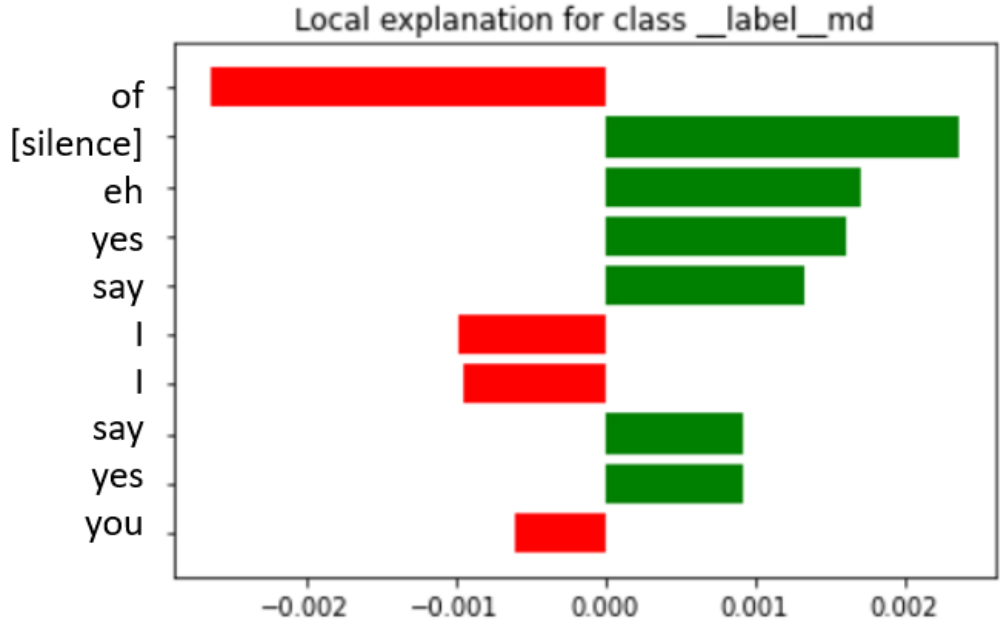

Figure 8. LIME explanation for quote 1 (top 10 words and how much they approximately contribute to the classification decision).

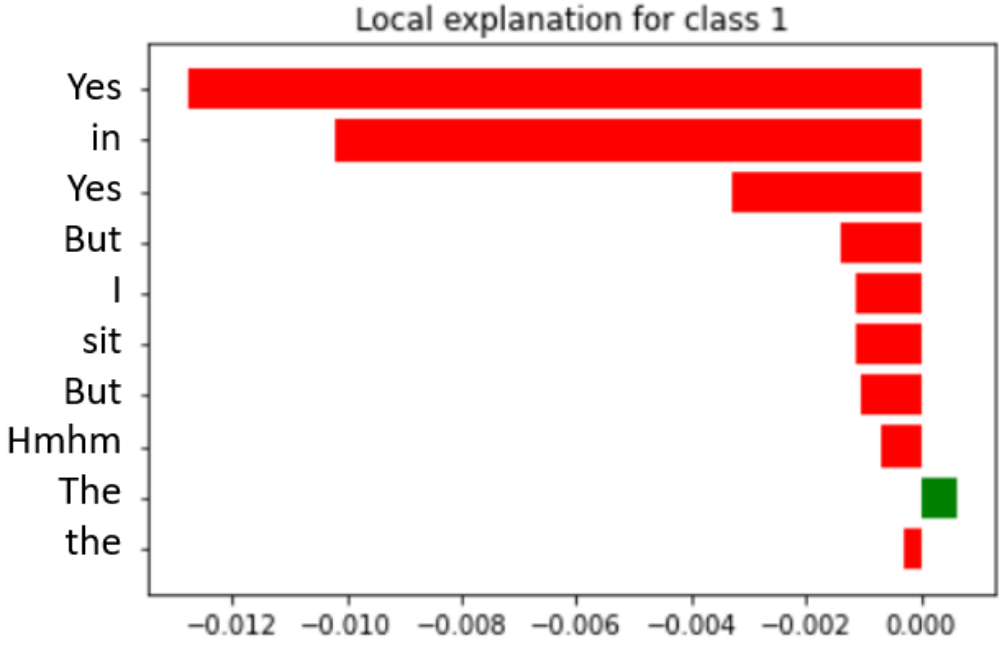

Figure 9. LIME explanation for quote 2 (top 10 words and how much they approximately contribute to the classification decision).

> 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_of_wife] and I lay together in bed. Furthermore, nothing happens there. As I do not need that, haha. However, I cannot even feel that I love her. I know it, that I love her. Furthermore, I know that my wife is and I, and I. However, that is all in here eh, but I do not feel it. Furthermore, that is the biggest measure which you can set· · · Yes. Furthermore, I talked about it with her."

Quote 2 is from someone with an eating disorder and was analysed with fastText. The word "eh" was highlighted because it explained why the transcription was labelled as

coming from a patient with a mental disorder (class = _label_md). Figure 9 shows the ten words with the highest probabilities from that transcription.

> 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. Then, you think oh I am good at something. Then, yes I don't know. Then 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. Furthermore, 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."

Other text also heavily featured conversational words such as "eh," "well," and "yes" in the LIME analyses. This suggests that perhaps for these interviews the difference between mental disorder and no disorder was more prevalent in the manner of speaking than in the topics they addressed.

Table 5 shows samples of eight interviews whose words resulted in the assignment of the mental disorder (MD) label or the no mental disorder (noMD) label. The first four interviews were analysed with stop words, and as can be seen, most of the words are stop words or "generally not meaningful" words. They could, however, be related to insightful words, which are also showen in the quotes. This could be supposedly because RobBERT looks both left and right in the context of a word in all layers of the transcription and then conditions it. Apparently, some words appear both in the mental disorder column and in the no mental disorder column, simply because these words appear in different contexts. Such words can contribute to a mental disorder classification in some language contexts, whereas in another context they do not. To further investigate, we removed all stop words from the last four interviews to determine whether LIME found more meaningful words. For example, in interview 7 with the fastText model, LIME found the words "psychiatrics" and "performance" as markers for a mental disorder, whereas in interview 8 LIME found the words "healing" and "job". In conclusion, without stop words we tended to find moderately more insightful words than with stop words. However, the words found by LIME are different for almost every interview and thus not yet applicable to support analyses of other interviews.

Table 5. LIME output of fastText and RobBERT for a sample of eight interviews.

ID	MD	SW	RobBERT	fastText	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, Furthermore, 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, Furthermore, whole, whole, she, one, also, eh, again	whole, completely, · · · , further, times	stood, sick, selfish, and, ehm
7	N	N	0.41	0.46	eh, nineteen ninety seven, of, notices of objection, say, team	car, ehm, team, through, However,	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

4.4. Summary of Findings: Language Markers

Table 6 shows an overview of the uncovered language markers for LIWC and spaCy. The 1SG LIWC pronoun notably came out as a language marker for a person with a mental disorder. In spaCy, 1SG was also the basis for labelling a mental disorder. The **W; $p < 0.05$** caption of the rightmost column refers to the Mann–Whitney two-tailed U tests that were performed to determine whether the means of the two groups per variable were equal to each other.

Unfortunately, we did not uncover clear patterns in the LIME results of the RobBERT and fastText neural network-based models, as different words were found for every interview to indicate either a mental disorder or no mental disorder.

Table 6. Summary of language markers uncovered by LIWC and spaCy.

	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

4.5. Focus Group

Furthermore, the results of the different models were discussed in a qualitative focus group session with UMCU data scientists, researchers and psychologists to better understand the outcomes. We discussed three key observations. First, the data scientists noted that the data used for this research are static data—i.e., somebody told their story and that was it. No new data from this particular person were added at a later time. The group hypothesised that following a person in their healing process, including their language usage, over a longer period of time, would result in additional relevant datapoints, and therefore could reveal additional interesting outcomes.

Second, the language markers found by LIWC and spaCy were discussed. The data originated from both people with mental disorders who told their own personal stories and from medical employees and family members who talked about people with mental disorders. This dual data origin situation likely influenced the outcome of this research. When an individual tells his own personal story, he will probably use more 1sg pronouns. Furthermore, when a health professional discusses an experience with a patient, he will likely use more 3sg and 3pl pronouns. Finally, people with mental disorders also shared their personal stories when they were not in an acute phase, and then, they could talk more about a completed story in their past. Therefore, the uncovered language markers actually make a lot of sense, according to the experts.

Third, rigid classifications are being abandoned in psychiatry, because they do not really help a person, according to some psychologists. However, if the current outcome classification will be changed depending on how far someone is in their healing process, one could find additional interesting results. The models discussed in this research could be applied for this new direction. To exemplify this, it was hypothesised that a person who is

further into his healing process will tell a more integrated story about his past than a person who is less far. In other words, "focuspast" could be a marker for someone being further into the healing process. Another proposition was that this research could be used to look at symptoms instead of being used for diagnostic assistance: what kind of treatment will help a person based on how he speaks? Another idea is to look at suicidality or aggression: what can a text tell us about that? Put differently, find out what a person is not explicitly saying, by analysing the deeper layers to find possible patterns or symptoms. One domain expert concluded: "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. Discussions and Conclusions

We have explored language markers in Dutch psychiatric interview transcriptions. We particularly focused on comparing the performances of traditional machine learning algorithms trained on LIWC and spaCy inputs with neural network approaches such as fastText and RobBERT, in predicting mental disorders. We found that the best performing technique in terms of determining whether a person has a mental disorder based on their word choices was LIWC in combination with random forest as the classification algorithm, which reached an accuracy of 0.952 and a Cohen's kappa of 0.889. Our hypothesis that the neural network approaches of fastText and RobBERT would perform best was not borne out. Several reasons may be posited. First, the pretrained language models of fastText and RobBERT did not for the most part consist of (transcribed) interview data. Second, the dataset was rather small (108 interviews) and the concept under consideration (mental illness) is not immediately apparent from a text. This suggests that for similar tasks with small datasets it may be best to use a dedicated algorithm such as LIWC, as it uses only a small selection of curated variables.

With regard to differentiating between mental illnesses, spaCy in combination with random forest predicted best which mental disorder each person had with an accuracy-score of 0.429 and a Cohen's kappa of 0.304. This moderate accuracy score can be explained due to the fact that the dataset of people with mental disorders only included 72 interview transcriptions and yet 10 mental disorder labels.

Finally, stop words did not appear to have that much influence on the performance of the classifiers except when employed using spaCy. We presume that is ws due to spaCy analysing the text from a grammatical point of view. When stop words are missing, spaCy cannot deduce the correct syntactic dependencies. Further work will focus on exploring additional model explainability techniques with differing explainability mechanisms and visualisation techniques in comparison to LIME, and investigating alternative NLP models in combination with an expanded data collection.

Ultimately, we argue that better understanding of a person's language use through the identification of language markers will result in better diagnosis of that person's mental health state, similar to the identification of a person's biomarkers. The impressive recent advancements within the field of Natural Language Processing are now allowing us to recalibrate our ambitions regarding language marker identification in informal patient narratives.

Author Contributions: Conceptualization, M.S. and F.S.; Data curation, K.d.S.; Formal analysis, S.V.; Funding acquisition, M.S.; Investigation, S.V., M.S. and K.d.S.; Methodology, M.S. and S.V.; Project administration, S.V.; Resources, M.S., K.d.S. and F.S.; Software, S.V.; Supervision, M.S., K.d.S. and F.S.; Validation, F.S.; Writing—original draft, S.V.; Writing—review and editing, M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Computing Visits Data (COVIDA) research programme of the Strategic Alliance Fund of Utrecht University, University Medical Center Utrecht and Technical University Eindhoven (round 2019).

Institutional Review Board Statement: The UMC Utrecht Medical Research Ethics Committee on 5 October 2016 confirmed with reference number WAG/mb/16/030724 that the Medical Research Involving Human Subjects Act (WMO) does not apply in the context of the Psychiatrieverhalenbank project with reference 16/626.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the Verhalenbank study.

Data Availability Statement: See Section 3.1 for more information on the Verhalenbank ("Storybank") dataset which is available at https://psychiatrieverhalenbank.nl/, accessed on 17 October 2021.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. 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. *Lancet* **2013**, *382*, 1575–1586. [CrossRef]
2. Ritchie, H.; Roser, M. Mental Health. In *Our World in Data*; 2020. Available online: https://ourworldindata.org/mental-health (accessed on 17 October 2021).
3. 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. *Lancet Psychiatry* **2016**, *3*, 993–998. [CrossRef]
4. 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. *Nat. Hum. Behav.* **2019**, *3*, 24–32. [CrossRef] [PubMed]
5. Lyons, M.; Aksayli, N.D.; Brewer, G. Mental distress and language use: Linguistic analysis of discussion forum posts. *Comput. Hum. Behav.* **2018**, *87*, 207–211. [CrossRef]
6. Calvo, R.A.; Milne, D.N.; Hussain, M.S.; Christensen, H. Natural language processing in mental health applications using non-clinical texts. *Nat. Lang. Eng.* **2017**, *23*, 649–685. [CrossRef]
7. Pennebaker, J.W.; Francis, M.E.; Booth, R.J. Linguistic inquiry and word count: LIWC 2001. *Mahway Lawrence Erlbaum Assoc.* **2001**, *71*, 2001.
8. 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, Lisbon, Portugal, 17–21 September 2015; pp. 1373–1378.
9. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
10. Delobelle, P.; Winters, T.; Berendt, B. RobBERT: A dutch RoBERTa-based language model. *arXiv* **2020**, arXiv:2001.06286.
11. Davcheva, E. Text Mining Mental Health Forums—Learning from User Experiences. In Proceedings of the 26th European Conference on Information Systems: Beyond Digitization—Facets of Socio-Technical Change, ECIS 2018, Portsmouth, UK, 23–28 June 2018; Bednar, P.M., Frank, U., Kautz, K., Eds.; AIS eLibrary: Atlanta, GA, USA, 2018; p. 91.
12. Deng, L.; Yu, D. Deep learning: Methods and applications. *Found. Trends® Signal Process.* **2014**, *7*, 197–387. [CrossRef]
13. 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, Denver, CO, USA, 5 June 2015; pp. 1–10.
14. Webster, J.; Watson, R.T. Analyzing the past to prepare for the future: Writing a literature review. *MIS Q.* **2002**, *26*, xiii–xxiii.
15. Tausczik, Y.R.; Pennebaker, J.W. The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* **2010**, *29*, 24–54. [CrossRef]
16. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent trends in deep learning based natural language processing. *IEEE Comput. IntelligenCe Mag.* **2018**, *13*, 55–75. [CrossRef]
17. Kim, K.; Lee, S.; Lee, C. College students with ADHD traits and their language styles. *J. Atten. Disord.* **2015**, *19*, 687–693. [CrossRef] [PubMed]
18. Nguyen, T.; Phung, D.; Venkatesh, S. Analysis of psycholinguistic processes and topics in online autism communities. In Proceedings of the 2013 IEEE International Conference on Multimedia and Expo (ICME), San Jose, CA, USA, 15–19 July 2013; pp. 1–6.
19. Forgeard, M. Linguistic styles of eminent writers suffering from unipolar and bipolar mood disorder. *Creat. Res. J.* **2008**, *20*, 81–92. [CrossRef]
20. Remmers, C.; Zander, T. Why you don't see the forest for the trees when you are anxious: Anxiety impairs intuitive decision making. *Clin. Psychol. Sci.* **2018**, *6*, 48–62. [CrossRef]
21. Trifu, R.N.; Nemeş, B.; Bodea-Hategan, C.; Cozman, D. Linguistic indicators of language in major depressive disorder (MDD). An evidence based research. *J. Evid.-Based Psychother.* **2017**, *17*, 105–128. [CrossRef]
22. 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. *Psychol. Trauma Theory Res. Pract. Policy* **2015**, *7*, 295. [CrossRef] [PubMed]

23. Corcoran, C.M.; Cecchi, G. Using language processing and speech analysis for the identification of psychosis and other disorders. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **2020**, *5*, 770–779. [CrossRef] [PubMed]
24. Verkleij, S. Deep and Dutch NLP: Exploring Linguistic Markers for Patient Narratives Analysis. Master's Thesis, Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands, 2021.
25. Choi, J.D.; Tetreault, J.; Stent, A. 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), Beijing, China, 15 July 2015; pp. 387–396.
26. Hermann, K.M. Distributed representations for compositional semantics. *arXiv* **2014**, arXiv:1411.3146.
27. Liang, P.; Potts, C. Bringing machine learning and compositional semantics together. *Annu. Rev. Linguist.* **2015**, *1*, 355–376. [CrossRef]
28. Guevara, E.R. A regression model of adjective-noun compositionality in distributional semantics. In Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics, Uppsala, Sweden, 16 July 2010; pp. 33–37.
29. Gamallo, P. Sense Contextualization in a Dependency-Based Compositional Distributional Model. In Proceedings of the 2nd Workshop on Representation Learning for NLP, Vancouver, BC, Canada, 3 August 2017; pp. 1–9.
30. Bohnet, B. Top accuracy and fast dependency parsing is not a contradiction. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, China, 23–27 August 2010; pp. 89–97.
31. Lei, T.; Xin, Y.; Zhang, Y.; Barzilay, R.; Jaakkola, T. Low-rank tensors for scoring dependency structures. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, MD, USA, 14 June 2014; pp. 1381–1391.
32. Choi, J.D.; McCallum, A. Transition-based dependency parsing with selectional branching. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Sofia, Bulgaria, 13 August 2013; pp. 1052–1062.
33. Van den Bosch, A.; Busser, B.; Canisius, S.; Daelemans, W. An efficient memory-based morphosyntactic tagger and parser for Dutch. *LOT Occas. Ser.* **2007**, *7*, 191–206.
34. Van der Beek, L.; Bouma, G.; Malouf, R.; Van Noord, G. The Alpino dependency treebank. In *Computational Linguistics in The Netherlands 2001*; Brill Rodopi: Amsterdam, The Netherlands, 2002; pp. 8–22.
35. Otter, D.W.; Medina, J.R.; Kalita, J.K. A survey of the usages of deep learning for natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 604–624. [CrossRef] [PubMed]
36. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *2*, 3111–3119.
37. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
38. Joulin, A.; Grave, E.; Mikolov, P.B.T. Bag of Tricks for Efficient Text Classification. *EACL* **2017**, *2017*, 427.
39. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
40. Wolf, T.; Chaumond, J.; Debut, L.; Sanh, V.; Delangue, C.; Moi, A.; Cistac, P.; Funtowicz, M.; Davison, J.; Shleifer, S.; et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 16–20 November 2020; pp. 38–45.
41. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
42. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
43. Khattak, F.K.; Jeblee, S.; Pou-Prom, C.; Abdalla, M.; Meaney, C.; Rudzicz, F. A survey of word embeddings for clinical text. *J. Biomed. Inform.* **2019**, *4*, 100057. [CrossRef]
44. Miotto, R.; Wang, F.; Wang, S.; Jiang, X.; Dudley, J.T. Deep learning for healthcare: Review, opportunities and challenges. *Briefings Bioinform.* **2018**, *19*, 1236–1246. [CrossRef]
45. Le, N.Q.K.; Yapp, E.K.Y.; Nagasundaram, N.; Yeh, H.Y. Classifying promoters by interpreting the hidden information of DNA sequences via deep learning and combination of continuous fasttext N-grams. *Front. Bioeng. Biotechnol.* **2019**, *7*, 305. [CrossRef] [PubMed]
46. de Vries, W.; van Cranenburgh, A.; Bisazza, A.; Caselli, T.; van Noord, G.; Nissim, M. Bertje: A dutch bert model. *arXiv* **2019**, arXiv:1912.09582.
47. Sarhan, I.; Spruit, M. Can we survive without labelled data in NLP? Transfer learning for open information extraction. *Appl. Sci.* **2020**, *10*, 5758. [CrossRef]
48. Loper, E.; Bird, S. NLTK: The Natural Language Toolkit. *arXiv* **2002**, arXiv:cs/0205028.
49. 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, New York, NY, USA, 13–17 August 2016; pp. 1135–1144.
50. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
51. McHugh, M.L. Interrater reliability: The kappa statistic. *Biochem. Medica* **2012**, *22*, 276–282. [CrossRef]