Utrecht University

UMC Utrecht

Applying machine learning in the classification of psychosis using syntactic, semantic and phonological features of speech

Fleur Slegers (5605407)

15 ECTS
Bachelor Thesis Artificial Intelligence

Practical supervisor: Drs. A. E. Voppel
1st supervisors: Dr. H. G. Schnack
2nd supervisors: Dr. D. Paperno

November 8th, 2019

Abstract

A variety of neurological and psychiatric illnesses are characterized by verbal communication disorders. Recently, there has been growing interest in automated speech-based techniques for screening mental disorders. Schizophrenia spectrum disorders are characterized by diminished effective expression and disturbances in thought and language, which results in disorganized speech. Diagnosing schizophrenia is often a challenging process prone to subjectiveness, as deviancies in speech are subtle and follow each other rapidly. Schizophrenia is the most common disorder in psychosis, which is a set of related conditions. As speech contains markers for schizophrenia, we believe that automated speech-based techniques may also be used to improve and simplify the process of diagnosing this disorder.

In this study, we implement multiple machine learning algorithms to examine the extent to which psychosis can be classified using syntactic, semantic and phonological features of speech. These features were extracted from speech samples using the tools T-Scan, Word2Vec and OpenSMILE, resulting in three separate data sets. Speech samples were collected by interviewing 50 psychotic patients and 50 healthy controls. We investigate the suitability of five different classification algorithms, namely Logistic Regression, Naïve Bayes, Random Forest, Stochastic Gradient Descent and Support Vector Machines on the separate data sets for classifying psychosis.

Our results show that distinguishing psychotic patients from healthy controls is possible using speech-derived features and techniques. Reasonably high accuracy scores can be achieved by using syntactic, semantic or phonological information about speech. This research adds to the field of clinical language analysis and has implications for future use of speech-based analytics in the clinical diagnostic process.

Keywords: psychosis, machine learning, automated speech-based techniques, syntax, semantics, phonology, T-Scan, Word2Vec, OpenSMILE.

Table of contents

Chapter 1

Introduction

Various neurological and psychiatric illnesses are characterized by verbal communication disorder. In schizophrenia spectrum disorders, mental resources are taxed, which results in diminished effective expression and disturbances in thought and language. This results in impaired social communication and disorganized speech, sometimes to the extent that speech is completely incomprehensible (Cohen, Kim, & Najolia, 2013). Many of the language abnormalities in schizophrenic spectrum disorders fall within formal thought disorder. Both positive and negative thought disorders pose restrictions on speech.

Positive thought disorder is characterized by an excess of normal function, which leads to derailment – a pattern of speech that tends to slip off track and in which remarkably unrelated concepts are expressed. Some patients with positive thought disorder produce neologisms – words that are self-invented - or use common words in an unusual way. In severe cases, this can lead to unintelligible speech in which neither sentences nor individual words seem to convey an overall meaning.

In negative thought disorder, normal behavior is partly absent, impaired or delayed, which leads to poverty of speech. These abnormalities are most evident in the syntactic, semantic and production aspects of language (Cokal, Zimmerer, Varley, Watson, & Hinzen, 2019). A detailed overview of the exact ways in which speech of schizophrenia patients differs from normal speech is provided in chapter 2.

It is evident that speech is an important aspect of the disorder profile of schizophrenia, so it should not come as a surprise that speech collected during a psychiatric interview is a crucial guideline for establishing a diagnosis. It is a relatively easily accessible measure that provides insight into the underlying clinical and cognitive aspects of schizophrenia (Cohen & Elvevag, 2014). Other guidelines for establishing a diagnosis are interviewer-based rating scales. However, these scales often contain relatively few response options and ambiguous operational definitions (Cohen, Kim, & Najolia, 2013). Besides, the ability and desire of a patient to communicate their symptoms have a large influence on the results of these scales (Cummins, et al., 2015). Identifying signals that indicate the presence of thought disorders is often challenging and subjective. Speech abnormalities are normally subtle and can succeed each other rapidly. In schizophrenia, the process is especially difficult when patients are not undergoing an acute psychotic episode at the time of the interview (Bar, 2019). The term psychosis covers a set of related conditions, of which the commonest is schizophrenia, and includes schizoaffective disorder, schizophreniform disorder, delusional disorder and non-affective psychosis (National Collaborating Centre for Mental Health, 2014). As the diagnosis of schizophrenia spectrum disorders is established purely through clinical observation (Kuperberg, 2010; M. J. Kas, 2019), a full understanding of speech deviancies in psychotic individuals should be established. Furthermore, diagnosis would benefit from an objective screening mechanism that is sensitive to subtle abnormalities in speech that are normally imperceptible to clinicians.

Recently, there has been growing interest in automated speech-based techniques for screening psychiatric disorders. For instance, Marmar, Brown, Quan, Laska, Siegel, Li, Abu-Amara, Tsiartas, Richey, Smith & Knoth (2019) established a speech-based algorithm that can objectively differentiate posttraumatic stress disorder (PTSD) cases from controls using phonological features extracted from clinical interviews. It demonstrates the usability of speech in the classification process of mental disorders. We believe that these or similar automated speech-based techniques are also suitable for classifying psychosis. However, unlike the PTSD study, phonology is not the only domain within language we are interested in with regard to psychosis. Fortunately, it is possible to extract syntactic, semantic and phonological quantified features of speech using different automated speech-based techniques in the realm of natural language processing (NLP).

Using semantic space models, it is possible to objectively extract precise and detailed information from a speech sample. The models represent words as points in an abstract multidimensional space. The models are designed so that words with similar meanings appear in similar contexts; the distance between word points is a measure for difference in semantic meaning. A method of semantic space models that can both capture semantic as well as some syntactic features is Word2Vec. Word2Vec uses neural networks to measure semantic and syntactic regularities in a large data set of words (de Boer J. N., et al., 2018). Phonological features, for example loudness, can be extracted from a speech sample using the toolkit OpenSMILE (Eyben, Wöllmer, & Schuller, 2010). The complexity of speech can be extracted using the software tool T-Scan, which yields mostly syntactical information. Using these toolkits, a multitude of variables can be extracted from speech samples that form a comprehensive syntactic, semantic and phonologic representation of a sample. To utilize these features in classification of psychosis, multiple machine learning methods will be implemented.

1.1 - Goals

In this research, we explore the possibilities for classification of psychosis using features of speech. We aim to answer the following sub-questions:

- How well can psychosis be predicted based on syntactic, semantic or phonologic information?
- What are the most important features of speech in the classification process within each domain (i.e. syntax, semantics and phonology)?
- How well can psychosis be classified combining the three domains?
- Which features of the combinatory dataset play an important role in the classification process?

The eventual aim of this study is to investigate the extent to which psychosis can be detected from controls using NLP and classification algorithms. A successful classification model can be used to guide clinicians in the process of diagnosing schizophrenia and increase our understanding of the disorder. Using NLP tools, subtle deviancies in speech that currently go unnoticed that could help to identify individuals at higher risk of psychosis can be detected at an early stage, as some of the cognitive impairments of psychosis are already detectable before the onset of the disease (Magaud, et al., 2010).

This is an exploratory study; the use of semantic space models and other NLP methods for classifying psychosis is a relatively new field. The results of this study can be used as a guideline for future research; if the predictability of psychosis proves to benefit from the use of features of speech, the feature importances found in this study can be used and validated in data sets acquired from different participants.

1.2 - Thesis structure

In chapter 2, a more detailed account of the conducted literature study on deviancies in speech of schizophrenia patients will be given. Chapter 3 and 4 contain the methods and results. Chapter 5 consists of the discussion.

Chapters 3 and 4 are built up by the following reoccurring structure: the three domains of speech - syntax, semantics and phonology - are first discussed separately (in that particular order), followed by a combinatory approach.

Chapter 2

Literature review

A literature review of the abnormalities in speech of schizophrenia patients was conducted. Based on this review, we identified features of speech that are abnormal in schizophrenia patients. In this chapter, an account of the most interesting findings will be given for the syntactical, semantical and phonological domains of speech. It needs to be stated that the reported findings are generalized findings on the group level of schizophrenia patients. These deviancies need not be present in each psychotic individual, for the patient group is heterogeneous. However, statements apply to the majority of the psychotic population and are thus reported as generally valid for the patient group. The findings presented in this chapter show that speech contains a multitude of markers for psychosis, which makes it a suitable domain for the classification of the disorder.

2.1 - Syntax

Syntax refers to the way words are arranged together (Jurafsky & Martin, 2009). Syntactic dysfunction disturbs the structure of language on all levels. Even when semantics and discourse organization of speech are impaired, the syntax of speech of schizophrenia patients can still be intact. Still, several aspects of syntax are abnormal in schizophrenia patients.

First of all, schizophrenia is characterized by reduced syntactic complexity and comprehension of speech (Covington, et al., 2005; Stanislawski, 2019; Kuperberg, 2010). Patients produce utterances which are syntactically less complex, and which contain more syntactic errors (Çokal, Zimmerer, Varley, & Watson, 2019). Patients also have difficulties with interpreting long and grammatically complex sentences (Kuperberg, 2010).

Furthermore, the use of parts of speech (nouns, verbs, pronouns, prepositions, adverbs, conjunctions, participles and articles) is deviant in schizophrenia patients (Jurafsky & Martin, 2009). Stanislawski (2019) showed that the use of determiner pronouns such as “which” and “that”, which introduce dependent clauses, is negatively correlated with negative symptoms in schizophrenia patients. In addition, usage of nouns is lower in schizophrenia patients compared to controls. Patients also use less adjectives and different adjectives to modify certain nouns compared to healthy controls (Obrebska & Obrebski, 2007; Bar, 2019). Pronoun use is also decreased in schizophrenia, particularly the personal pronoun “I” (Deutsch-Link, 2016). Lastly, patients use more verbs compared to controls (Obrebska & Obrebski, 2007).

Formal thought disorder in schizophrenia results in the diminished use of embedded clauses in spontaneous speech (Çokal, Zimmerer, Varley, & Watson, 2019). An embedded clause is a clause that in itself is not a complete sentence, but which is placed in the middle of another clause. Patients suffering from formal thought disorder also show a deficit in comprehending embedded clauses (Çokal, Zimmerer, Varley, & Watson, 2019).

2.2 - Semantics

Lexical semantics is the study of the meaning of words. Semantic impairment affects the ability to map thoughts onto language and pursue a communicative goal (Covington, et al., 2005). In schizophrenia patients, semantic coherence and word use is deviant.

Words that appear within the same context are usually more semantically related than words appearing in different contexts. Speakers with schizophrenia often jump from one subject to another based on the sounds or associations of words they have uttered earlier. Patients produce a greater number of associations between words than healthy controls, which results in a higher number of shifts between topics that are not or only remotely related to previous topics. This difference is particularly noticeable in units of text greater than fifteen words, which suggests that associations between words stretch over longer periods in patients than in controls (Kuperberg, 2010). Reductions in semantic coherence are a predictor of psychosis onset in clinical high-risk individuals (Stanislawski, 2019).

In conversation, patients with schizophrenia often use words that are incompatible with preceding sentences (Kuperberg, 2010). Discourse often includes neologisms and rare words, indicating the presence of a large and intact vocabulary. Schizophrenia patients use different adjectives from controls to modify certain nouns (Bar, 2019). A recent study conducted by Rezaii, Walker & Wolff (2019) found that besides low semantic density, conversion to psychosis is also signaled by increased usage of words related to voices and sounds.

2.3 - Phonology

Phonology is concerned with the systematic organization and production of sounds in speech. The main phonologic feature deviant in speech of schizophrenia patients is pause length; schizophrenic speech contains abnormal pauses. Furthermore, pause length and percentage of pauses is significantly correlated with negative symptom severity (Stanislawski, 2019; Cohen, Kim, & Najolia, Psychiatric Symptom versus Neurocognitive Correlates of Diminished Expressivity in Schizophrenia and Mood Disorders, 2013). Increasing severity of negative symptoms is also associated with less prosody (the patterns of stress and intonation) (Cohen, Kim, & Najolia, 2013). In a meta-analysis of 46 articles conducted by Alberto, Arndis, Vibeke & Riccardo (2019), weak atypicalities in pitch variability related to flat affect and stronger atypicalities in proportion of spoken time, speech rate and pauses related to alogia and flat affect were found. However, these effects are modest compared to perceptual and clinical judgements and characterized by heterogeneity between studies.

The features of speech discussed in this chapter can be quantitatively extracted from speech samples. Based on our findings in this literature review, we expect these features to be valuable for classification purposes.

Chapter 3

Methods

In this chapter, the methods and approaches used in this thesis are described, including a description of the participants and the collection of speech samples. NLP tools used to extract features from the samples of speech will be discussed, followed by a description of feature selection methods. Because NLP tools and feature selection methods differ in the syntactic, semantic and phonological domain, this section will be partitioned into these 3 domains. Next, the process of classifying psychosis using the selected features of speech will be clarified. This includes a description of feature selection methods and the machine learning algorithms that were implemented. This chapter ends with a description of how the results from the machine learning models were analyzed and compared. Lastly, a description of the statistical analysis conducted on the 3 data sets will be given.

3.1 - Participants

50 patients with a schizophrenia spectrum disorder and 50 healthy controls were included. Healthy controls were screened for former or current mental illness. This was done either by a neuropsychologist using the Comprehensive Assessment of Symptoms and History (CASH) or with the use of a modified psychiatric history screener. In case of former or current mental illness, controls were excluded. For healthy participants that underwent CASH screener, family history of psychotic symptoms was also a criterion for exclusion. Patients were diagnosed by their treating psychiatrist. In addition, the diagnosis was confirmed using the outcome of the CASH or the Mini International Neuropsychiatric Interview 5.0.0. (M.I.N.I. Plus (Sheenan, et al., 1998)). The severity of psychotic symptoms was assessed by means of the Positive and Negative Syndrome Scale (PANSS (Leucht, et al., 2005)). All participants were age eighteen or above and had Dutch as their first language. To be included, patients also required a DSM-IV diagnosis of 295.x (schizophrenia, schizophreniform disorder, schizoaffective disorder) or 298.9 (psychotic disorder NOS) (Millon & Davis, 1996). Exclusion criteria were the presence of uncorrected hearing disabilities and speech impediments, for example excessive stutter. Informed consent was obtained from all participants before study participation. Participants received a small monetary reward.

3.2 - Data collection

Samples of spontaneous speech were collected in semi-structured interviews of 5 to 30 minutes. In these interviews, a set of questions about general life experiences was used, with the aim to avoid topics that could be expected to have noticeably different emotional charge for patients and healthy controls. Potential variations in speech due to the discussed topics were controlled for using this approach. If a participant did not want to answer a question, the interviewer would move on to the next question.

The subject’s speech was recorded using an AKG-C544l head-worn cardioid microphone. Speech was digitally recorded onto a Tascam DR40 solid state recording device at a sampling rate of 44,100 kHz with 16-bit quantization. Per interview, two audio tracks were recorded; one for the participant and one for the interviewer to aid in speaker separation. Each segment of speech was coded, using the Praat software, as belonging either to the participant or the interviewer. When both speakers spoke at the same time, the segment was coded as belonging to both speakers. The pause resulting from a switch between speakers was attributed to the speaker following the pause. For every individual, a new audio file was created consisting only of the speech segments in which the participant in question was speaking or pausing. Data files were blinded for diagnosis to prevent bias in separating the speaker. Inter-rater reliability for tier separation was 97.7 percent. All files were set to an average sound pressure level of 60dB to avoid differences in the analyses based on speaking volume. Transcription of the interviews were produced using the tools CLAN (MacWhinney & Wagner, 2014) and CHILDES (MacWhinney B. , 2014).

3.3 – Data preprocessing

Features of speech were extracted from the speech samples using three different linguistic tools; namely T-Scan, Word2Vec and OpenSMILE. These tools extract mainly syntactical, semantical and phonological features, respectively. For all three domains, feature selection was executed to remove irrelevant and redundant features that can otherwise negatively impact model performance in terms of accuracy and time to build the model (Gnana, Balamurugan, & Leavline, 2016). The feature selection procedures used are clarified for each domain in their respective sections. Prior to training of classification algorithms, features were standardized using min-max scaling, which brings each value between 0 and 1.

3.4 – Syntactic data

Syntactic features of speech were extracted from the transcriptions of the interview using T-Scan. T-Scan is a software tool for text analysis that mainly captures information about text complexity by measuring word difficulty, sentence complexity, referential and relational coherence, verbiage, use of semantic classes, personal elements, use of names, probability measures and usage of part-of-speech tagging. A text is analyzed on the level of words, sentences, paragraphs and in its entirety. The Output of T-Scan gives a syntactical representation of a text (Maat, Kraf, & Dekker, 2017). The output of T-scan consists of 457 variables in total.

To create a more suitable data set for our sample size, feature selection guided by literature research was carried out. An overview of previous findings on syntactical deviations in schizophrenic speech can be found in chapter 2. Using these findings, 24 corresponding T-Scan variables were identified and selected to be used by the machine learning models. The selected variables are given in table 3.1.

Table 3.1

Selected variables from t-scan output with description.

	Variable name	Description*
1	D_level	Measure for syntactic complexity
2	Wrd_per_morf	Number of words per morpheme, measures word length
3	Al_gem	Measure for the distance between two clauses that are dependent of each other
4	Lem_over_buf_dz	Measures the number of referential repetitions
5	Onbep_nwg_dz	Number of indefinite noun groups
6	Conn_temp_dz	Number of temporal connectives (e.g. "*before", "formerly")* per clause
7	Conn_reeks_zin_dz	Number of words that connect clauses (e.g. *"and", "furthermore"*) per sentence
8	Conn_contr_dz	Number of oppositive connectives (e.g. "still", "even so") per clause
9	Conn_caus_dz	Number of causal connectives (e.g. "*when*", "*because*") per clause
10	Ww_tt_p	Proportion of present tense verbs
11	Vd_vrij_dz	Number of free-standing past participles per clause
12	Inhwrd_d	Density of content words
13	Pv_Frog_d	Density of verbs
14	Ontk_tot_d	Density of refutations
15	Pers_vnw1_d	Density of first-person personal and possessive pronouns

16	Pers_vnw3_d	Density of third-person personal and possessive pronouns
17	Bvnw_d	Density of adjectives
18	Vg_d	Density of conjunctives
19	Vnw_d	Density of pronouns
20	Lidw_d	Density of articles
21	Tuss_d	Density of interjections
22	Int_bvnw_d	Density of intensifying adverbs
23	Alg_bijw_d	Density of general adverbs
24	Spec_bijw_d	Density of specific adverbs

* these descriptions are translations from the Dutch descriptions as given in the T-Scan manual (Maat, Kraf, & Dekker, 2017).

3.5 – Semantic data

The Word2Vec model is a tool for learning word embeddings using neural networks. The model contains a dictionary in which each word is represented by a feature vector. This feature vector captures syntactic and semantic relationships between words and can thus be used for quantitative examination of words. The feature vector is established during training of the model. Training takes places either according to the continuous bag of words (CBOW) or Skip-Gram architecture. In CBOW, the model uses the context of a word, e.g. its neighboring words, to predict that word. In Skip-Gram, the model uses a word to predict its context. The limit on the number of words in a context is determined by a parameter called window size. When the feature vector assigned to a word cannot be used to accurately predict the word's context, or the other way around, the components of the vector are adjusted. A well-trained set of word vectors will place similar words close to each other. The resulting representations of words exhibit a linear structure that makes precise analogical reasoning possible (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013).

We trained our Word2Vec model on the corpus *gesproken Nederlands* (REF) (Oostdijk, 2000) using the skip-gram architecture, with a 300-dimensional semantic model. Following model creation, each word of each participant's transcript was semantically compared to other words in the window size, using the cosine of the angle between vectors as a metric for distance in 300-dimensional semantic space. Average, minimum, maximum and mean distances over a given window are indications of semantic coherences. These measures are calculated for a window of fixed size that moves over the words one by one (we will call this *simple*) and a partition of the words according to a fixed window size (we will call this *summary*). We repeated these measures for window sizes ranging from 2 to 20. After having been run on the speech samples, Word2Vec generated 159 output variables.

Because the output from Word2Vec takes on a different form than results found in similar studies, see chapter 2, our literature review could not be used as a method for feature selection. Instead, the Random Forest algorithm (RF) was used for this purpose. Feature selection based on RF has been found to provide multivariate feature importance measures which are relatively cheap to obtain, and which have been successfully applied to high dimensional data. Multiple studies indicate the efficiency of the importance measures for a RF classifier in an explicit feature selection (Menze, et al., 2009), and the method was also applied to create a subset of features for the classification of PTSD (Marmar, et al., 2019). A more in-depth description of RF is given in section 3.7. Running the RF classifier on the output of Word2Vec results in a feature importance measure for all variables. The 20 features with the highest measure were selected to be used for classification using other algorithms.

3.6 – Phonological data

OpenSMILE is an audio analysis toolkit that retrieves phonological information from a segment of speech. It enables explorative analysis of audio segments by combining feature extraction and pattern recognition. OpenSMILE extracts some Low-Level Descriptors (LLD) and applies various functionals and transformations to these. These LLD's include waveform, signal energy, loudness, pitch and voice quality. For a complete overview of the OpenSMILE feature set, we refer to Eybe, Weninger, Wöllmer & Schuller (2014). The functionals that can be applied to the LLD's are extremes, means, moments, percentiles, regression, peaks, segments, sample values, times/durations, onsets, coefficients of the Discrete Cosine Transformation (DCT) and zero-crossings. Functionals can be applied repeatedly in a hierarchical structure. This results in a a large quantity of variables; the *emo_large* feature extraction setting that is used in this study produces 6557 variables.

In order to create a suitable data set of phonological features to be used in classification, RF was used. The 19 features with highest feature importance coefficients were selected before being used by other classification algorithms.

3.7 – Machine learning models

For every domain, multiple machine learning algorithms were used to classify psychosis in order to find out which model is most capable in distinguishing psychosis patients from controls using the respective data sets. The models that were implemented are Support Vector Machines (SVM's), Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF) and Stochastic Gradient Descent (SGD). The models were implemented using the open source python module Scikit-learn. In this section, a description of these algorithms will be given. Leave-two-out cross-validation (L2O-CV) was implemented to optimize parameter settings for every model while minimizing risk of overfitting. A description of this process is also given.

3.7.5 - Support Vector Machine

When used for classification, Support Vector Machines (SVMs) separate a given set of binary labeled training data with a hyperplane that is maximally distant from these data points. When SVMs are combined with kernel functions, they become applicable even when linear separation is not possible. A kernel function maps the dataset into a higher dimensional space where a hyperplane can then be found that does separate the samples. In this study, the Radial Basis Function (RBF) is applied, whose value depends only on the distance between two points. The hyperplane found by an SVM corresponds to a decision boundary in the input space. The location of a data point from a test set in respect to this hyperplane then determines which class is assigned to that point. A benefit of SVMs is that they can be used to identify instances whose established classification is incorrect. SVMs are well suited to working with high dimensional data and are remarkably proof against sparse and noisy data (Furey, et al., 2000).

3.7.1 – Logistic Regression

Logistic Regression (LR) is one of the most used Machine Learning algorithms for binary classification. The LR classifier assigns a class to a data point based on the logistic function whose values lie between 0 and 1. The LR hypothesis is

$$h_\theta(\vec{x}) = g(\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n)$$

where $\vec{x}$ is the feature vector of a data point and g is the logistic function. Based on the value of $h_\theta(\vec{x})$ in respect to a certain threshold, one of two classes will be assigned to the data point $\vec{x}$. The optimization problem of a LR classifier consists of computing the optimal parameter vector θ. The LR classifier is suitable for dichotomous data. It can handle both dense and sparse input. An advantage of LR is its simplicity; it is highly interpretable, relatively easy to implement and training is efficient. However, solving non-linear problems with LR is more difficult to implement (James, Witten, Hastie, & Tibshirani, 2013).

3.7.2 - Naive Bayes

Bayesian classifiers assign the most likely class to an example that is described by its feature vector. The Naive Bayes classifier (NB) simplifies learning by assuming that features are independent, given a class. Although independence is an unrealistic assumption, NB often competes well with more sophisticated classifiers in practice; its classification decision is often correct even when its probability estimates are inaccurate. The NB classifier has proven effective in many practical applications, including text classification and medical diagnosis (Rish, 2001).

3.7.3 - Random Forest classifier

The Random Forest classifier (RF) consists of a combination of tree classifiers where each classifier is generated using a randomly selected combination of features. Each tree in RF casts a unit vote for the most popular class to classify an input vector. RF uses averaging over the tree classifiers, which improves the accuracy of the overall model and controls overfitting (Pal, 2005). A benefit of RF is that its decision trees can be seen as a collection of if-statements, which makes the results highly interpretable (James, Witten, Hastie, & Tibshirani, 2013).

3.7.4 – Stochastic Gradient Descent classifier

The Stochastic Gradient Descent (SGD) classifier implements linear classification models (e.g. SVM, logistic regression) with SGD learning; at each iteration the gradient is estimated on a single data point that is randomly selected, and weights are updated accordingly. SGD decreases the error with respect to one data point at a time. The algorithm does not remember which data points were visited during previous iterations (Bottou, 2010). Applying SGD learning to linear classifiers speeds up computation time (Wijnhoven & de With, 2010).

3.7.5 – Evaluation of classification models

All classification algorithms in this study have parameters that can be tuned. Tuning of these parameters has a great impact on the performance of the models; parameter tuning is often more important than the choice of algorithm (Lavesson & Davidsson, 2006). In order to find the optimal parameter setting for each combination of classifier and data set, leave-two-out cross-validation (L2O-CV) is implemented. L2O-CV is a method for estimating pointwise out-of-sample prediction accuracy from a trained classification model. It requires refitting the model with *n/2* different training sets, where *n* is the number of available data points. In our case, this means that the model is trained 50 times in total. Each time the data set is partitioned by withholding two data points, one psychosis patient and one healthy control, of which the classes are predicted by the model that is trained on the remaining 98 data points. All data points are used as testing point exactly once (Vehtari, Gelman, & Gabry, 2016). An average accuracy score over all 50 partitions is then calculated. This process is repeated for all possible combinations of parameter settings of a model.

Performance of a model is measured not only by accuracy, but by precision and recall as well. A data point can be labeled in 4 different ways which are given in the confusion matrix in table 3.2.

Table 3.2

Confusion matrix

		True value	
		Control	**Psychotic**
Predicted value	**Control**	True negative	False negative
	Psychotic	False positive	True positive

For every fold (one instance of the 50 times the data set is partitioned during L2O-CV), the labels of two data points are predicted. The entries of the confusion matrix for these predictions are recorded. When the cross-validation process is completed, accuracy, precision and recall scores are calculated. Accuracy score is

$$\frac{true\ positives + true\ negatives}{100}$$

and measures the percentage of correctly labeled data points. Precision is the ratio

$$\frac{true\ positives}{true\ positives + false\ positives}$$

and is a measure for the model's ability to not label negative samples as positive. Recall is the ratio

$$\frac{true\ positives}{true\ positives + false\ negatives}$$

and measures the model's ability to find all the positive samples of a data set (Powers, 2011). Only accuracy scores are used to determine the optimal parameter setting. Recall and precision scores are used to gain insight of the prediction abilities of the classifier and are used to compare the performance of different classifiers.

After the optimal parameter setting of a model has been established, training is repeated on the full training set using this setting. After the model has been trained on the training data set in its entirety, a test set can be used to validate in-training scores. Due to time limits, validation of the models has not yet been carried out.

To increase our understanding of the differences between normal and psychotic speech, it is informative to know which features of speech play an important role in the classification process; these measures are important for separating healthy controls from psychotic patients. Using the Scikit-learn tools, feature importances for both SVM with linear kernel can be accessed and visualized.

3.8 - Combining the domains

Combining the results from the different classifiers and domains of speech makes analyzing the results on a participant level possible. We are interested in which participants are relatively often misclassified and who receive a correct label most of the time. This information can be extracted from the accuracy, precision and recall scores in the following way: per fold in the cross-validation process two subjects (one from the patient and one from the control group) receive a predicted label. For each classifier, the test set in a specific fold includes the same subjects, which makes comparison of each subject over the different classifiers possible. As there is only one patient and one control per fold, accuracy and precision scores must be either 0, ½ or 1 and recall must be either 0 or 1. Because of how these scores are calculated, there are only 4 possible combinations of these values, which makes the identification of a predicted label a deterministic process. Whether the participant from the patient group and the healthy control received a correctly predicted class label can be read off table 3.3.

Table 3.3.
Table to read out the correctness of a predicted label for the patient and control in the test set of a fold of L2O-CV using the accuracy, precision and recall scores of that fold.

Accuracy	Precision	Recall	Correctly classified?	
			Patient	Control
1	1	1	Yes	Yes
0.5	0.5	1	Yes	No
0.5	0	0	No	Yes
0	0	0	No	No

Using table 3.3, we can calculate whether a subject's predicted class was correct and derive the predicted class for each classifier and domain. The mode of the predicted classes for a subject can then be calculated over all classifiers and domains. The mode of a set of values is the value that appears most often. This metric is a form of ensemble learning and provides a classification of a subject based on multiple classifiers. The mode is expected to have lower variance and thereby higher accuracy than a single classifier would have (Polikar, 2012).

3.9 - Statistical analysis

As mentioned in section 3.7, feature importances are only visualized for the SVM model. Due to time limitations, no further methods for data visualizations are implemented. To gain more insight into the exact ways syntax, semantics and phonology of psychotic speech differs from normal speech, statistical analysis is applied. Independent-samples t-tests are conducted to compare the selected features of the T-Scan, Word2Vec and OpenSMILE data sets for the patient group and controls. As this is an exploratory study and the results from the statistical analyses are complementary the evaluation of the classification models, we have chosen not to correct for multiple testing. We are however aware of the consequential increased risk of finding correlations by chance. The results of these analyses can help to understand the classification processes described in the previous chapter; findings can explain why certain features are important in the classification process. It improves interpretability of the machine learning models and underpins why a certain decision is made, which is valuable as these models are designed for use in clinical diagnosis. Besides, results from the statistical analyses can be used in the feature selection process of future comparable studies.

Chapter 4

Results

In this chapter, performance of the classification algorithms is presented per domain of speech after a general report of our findings has been given. Figure 4.1 shows in-training accuracy scores for each classification algorithm on the T-Scan, Word2Vec and OpenSMILE data sets.

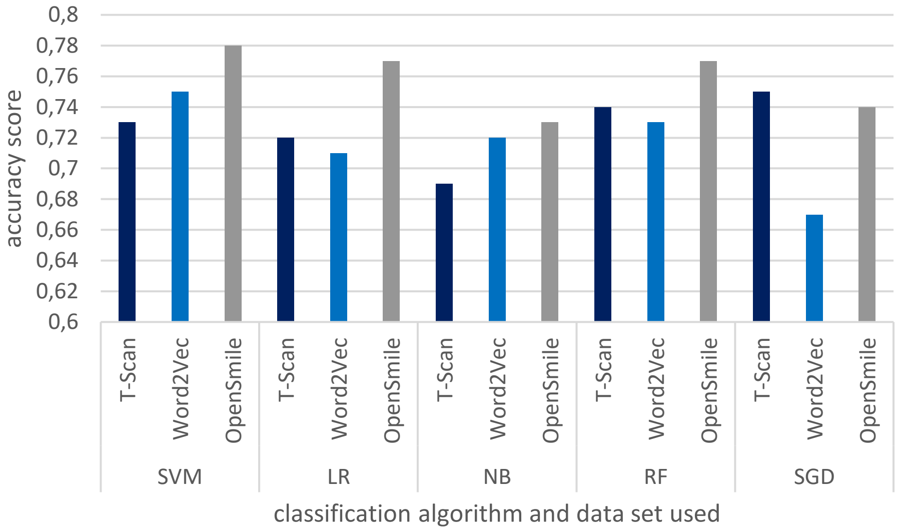

Figure 4.1. In-training accuracy scores of classifiers trained on T-Scan, Word2Vec and OpenSMILE data sets. SVM stands for Support Vector Machine, LR for Logistic Regression, NB for Naïve Bayes, RF for Random Forest and SGD for Stochastic Gradient Descent.

Figure 4.1 shows that all possible combination of classifier and data sets achieved accuracy scores between 0.67 and 0.78. As a reminder, these scores are the averages of all leave-one-out cross validation scores. Note that model performance strongly depends on the data set that is used for classification. No single classifier can be pointed out that consistently attains highest accuracy. After calculating average accuracy scores per classifier, we find that the SVM and RF classification algorithms achieve the highest overall accuracy. The NB model performs worst on average. Looking at the separate data sets, it is striking that OpenSMILE consistently achieves relatively high accuracy. Average accuracy scores per data set were also computed. For T-Scan, average accuracy was 0.726; for Word2Vec 0.716 and for OpenSMILE 0.758.

4.1 - T-Scan

Table 4.1 contains the performance scores (accuracy, precision and recall) of the classification models trained on the T-Scan data set. Models were trained using the optimized parameter setting as found using leave-one-out cross validation. The optimized parameter settings per classification algorithm are given in table 4.2.

Table 4.1

Performance scores of the optimized models trained on the T-Scan data set.

Model	Accuracy score	Precision score	Recall score
Support Vector Machine	0.73	0.58	0.64
Logistic Regression	0.72	0.58	0.66
Naïve Bayes	0.69	0.53	0.60
Random Forest	0.74	0.59	0.66
Stochastic Gradient Descent	0.75	0.62	0.70

Note that all performance scores are above 0.5, which means that all models are better at classifying psychosis from controls than a procedure that randomly assigns classes would be. Precision scores lie relatively close to 0.5 in comparison to the other performance measures. As precision measures the ratio of true positives to the total of positively classified cases, this means that relatively many control cases are classified as psychotic; as recall scores give us the proportion of psychotic cases that were classified as such. In figure 4.2, precision score is plotted against recall score for each model.

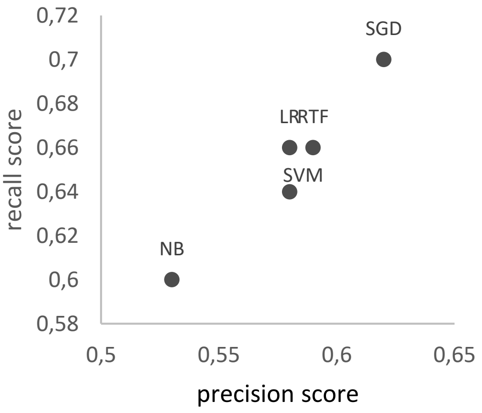

Figure 4.2. Recall and precision scores of the optimized models trained on the T-Scan data.

Regarding the overall performance on the T-Scan data set, we deduce from table 4.1 and figure 4.2 that the SGD model performed best, and the NB model performed worst. The optimal parameter settings found in this process are given in table 4.2.

Table 4.2

Optimized parameter values for the classification algorithms on T-Scan data.

Classification algorithm	Optimized parameter values
Support Vector Machine	C = 10
	gamma = 0.1
	kernel = rbf
Logistic Regression	C = 1
	penalty = l2
	tol = 1 x 10^{-6}
Naïve Bayes	var_smoothing = 1 x 10^{-10}
Random Forest	criterion = entropy
	min_samples_leaf = 2
	min_samples_split = 3
Stochastic Gradient Descent	alpha = 0.01
	loss = log
	penalty = elastic net

To gain insight in which features play an important role in the classification process, an SVM with linear kernel was run on the T-Scan data. The feature importances of the 20 most important features in classification for the SVM model with linear kernel are shown in figure 4.3. A clarification of the variable names in this figure can be found in table 3.1.

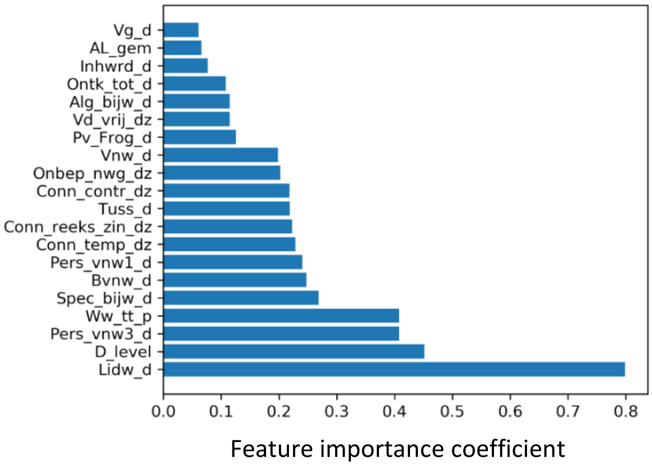

Figure 4.3. Feature importances of the 20 most important features of T-Scan data of a Support Vector Machine with linear kernel.

4.2 - Word2Vec

To create a smaller data set more suitable for classification, the top 20 features of the Word2Vec data for the Random Forest classifier were selected. These 20 features were then used by the other classifiers. Table 4.3 lists these features.

Table 4.3

Selected features from the Word2Vec data set.

1	Var_summary_5
2	Min_summary_9
3	Min_summary_17
4	Min_simple_8
5	Min_summary_20
6	Mean_summary_14
7	Max_simple_18
8	Max_simple_17
9	Max_summary_10
10	Mean_summary_2
11	Min_simple_3
12	Var_simple_17
13	Var_simple_19
14	Var_simple_10
15	Var_simple_14
16	Max_summary_12
17	Var_simple_9
18	Max_summary_17
19	Var_simple_5
20	Min_simple_20

Var stands for variance, min for minimum, max for maximum and *n* for window size. Summary and simple are explained in section 3.5.

Table 4.4 contains the performance scores of the classification models trained on the Word2Vec data subset. Models were trained using the optimized parameter setting found using cross validation. These parameter settings are given in table 4.5.

Table 4.4

Performance scores of the optimized models on the Word2Vec data subset.

Model	Accuracy score	Precision score	Recall Score
Support Vector Machine	0.75	0.59	0.64
Logistic Regression	0.71	0.57	0.66
Naïve Bayes	0.72	0.54	0.58
Random Forest	0.73	0.58	0.64
Stochastic Gradient Descent	0.67	0.48	0.58

Table 4.5

Optimized parameter values for the classification algorithms on Word2Vec data subset.

Classification algorithm	Optimized parameter values

Support Vector Machine	C = 10
	gamma = 0.1
	kernel = rbf

Logistic Regression	C = 1
	penalty = l1
	tol = 0.1

Naïve Bayes	var_smoothing = 0.01

Random Forest	criterion = entropy
	min_samples_leaf = 2
	min_samples_split = 4

Stochastic Gradient Descent	alpha = 1 x 10^{-7}
	loss = squared hinge
	penalty = elastic net

Note again that all performance scores are above 0.5. As was the case for the T-Scan data set, precision scores are low in comparison to the other performance measures. In figure 4.4, precision score is plotted against recall score for each model.

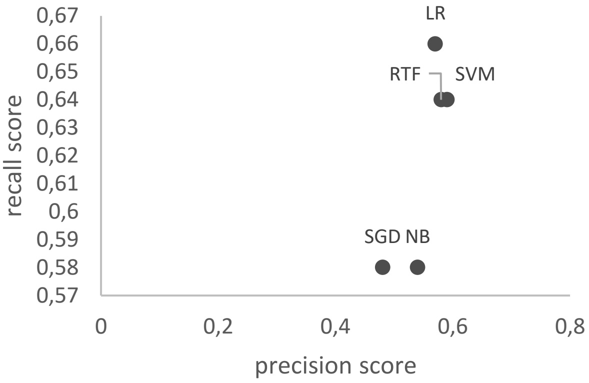

Figure 4.4. Recall and precision scores of the optimized models on Word2Vec data subset.

Note that as opposed to the T-Scan data set, the SGD classifier has lowest performance scores on the Word2Vec data. The SVM classifier has highest performance scores.

Figure 4.5 shows the feature importance coefficients for the 20 variables that were selected using RF. Notice that the label '*simple'* appears relatively often in the list of variables in comparison to the label '*summary*'. Recall that for the *simple* variables minimum, maximum, mean and variance were calculated on subsets of the data resulting from a moving window size of size *n,* while for the *summary* variables the data was split in *n* disjoint sets. Especially variables of the form *'var_simple_n'*, which measure the variance of coherence of all instances of a moving window of size *n*, appear relatively high in the list.

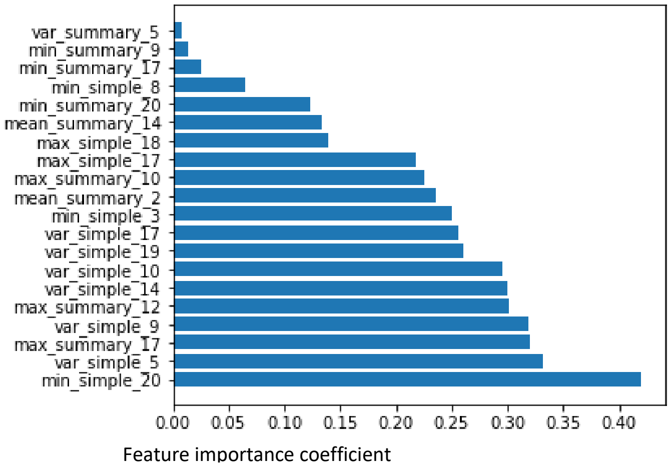

Figure 4.5. Feature importance coefficients for the top 20 variables of the Word2Vec data for a Support Vector Machine with linear kernel.

4.3 – OpenSMILE

As was the case with the Word2Vec data, a subset of features from the OpenSMILE data was selected using the Random Forest classifiers. The selected features are enumerated in table 4.6.

Table 4.6

Selected features from the OpenSMILE data set.

1.	Pcm_fftMag_melspec_sma_de[4]_percentile95.0
2.	Mfcc_sma[9]_stddev
3.	Mfcc_sma[9]_kurtosis
4.	Mfcc_sma_de_de[2]_peakMean
5.	Pcm_fftMag_spectralCentroid_sma_amean
6.	Pcm_fftMag_melspec_sma_de_de[6]_qregerrA
7.	Pcm_fftMag_dband0-650_sma_variance
8.	Pcm_fftMag_melspec_sma_de_de[14]_maxPos
9.	Pcm_fftMag_melspec_sma[19]_qregc3
10.	Pcm_fftMag_melspec_sma[24]_nzqmean
11.	Mfcc_sma[12]_linregc1
12.	Pcm_fftMag_melspec_sma_de_de[12]_iqr1-2
13.	Pcm_fftMag_melspec_sma_de[17]_variance
14.	Mfcc_sma_de[4]_variance
15.	Pcm_fftMag_spectralFlux_sma_amean
16.	Pcm_fftMag_melspec_sma[15]_iqr1-3
17.	F0env_sma_qregc3
18.	Pcm_fftMag_spectralRollOff50.0_sma_de_de_iqr2-3
19.	Pcm_fftMag_melspec_sma[6]_iqr1-3

A description of the variable names can be found in Eyben, Wöllmer & Schuller, 2010.

Table 4.7 contains the performance scores of the classification models trained on the OpenSMILE data subset. were trained using the optimized parameter setting found using cross validation. These parameter settings are given in table 4.8.

Table 4.7

Performance scores of the optimized models on the OpenSMILE data subset.

Model	Accuracy score	Precision score	Recall Score
Support Vector Machine	0.78	0.64	0.70
Logistic Regression	0.77	0.65	0.72
Naïve Bayes	0.73	0.68	0.84
Random Forest	0.77	0.66	0.72
Stochastic Gradient Descent	0.74	0.63	0.76

Table 4.8

Optimized parameter values for the classification algorithms on OpenSMILE data subset.

Classification algorithm	Optimized parameter values

Support Vector Machine	C = 100
	gamma = 0.01
	kernel = rbf

Logistic Regression	C = 10
	penalty = l2
	tol = 1 x 10^{-6}

Naïve Bayes	var_smoothing = 0.1

Random Forest	criterion = gini
	min_samples_leaf = 2
	min_samples_split = 2

Stochastic Gradient Descent	alpha = 0.01
	loss = squared_hinge
	penalty = l2

We notice that classification using the OpenSMILE data performs consistently better than when T-Scan or Word2Vec data is used; all performance scores for each classifier is higher for the OpenSMILE data subset than for the T-Scan or Word2Vec data sets. Recall scores for OpenSMILE lie within the range from 0.70 to 0.84, whereas recall scores for T-Scan lie between 0.60 and 0.70 and for Word2Vec between 0.58 and 0.66. OpenSMILE also generates better precision scores; namely scores ranging from 0.63 to 0.68, compared to scores from 0.53 to 0.59

for T-Scan and scores from 0.48 to 0.59 for Word2Vec. In figure 4.6, precision score is plotted against recall scores for the results from the OpenSMILE data subset.

In figure 4.7, the feature importance coefficients of the selected features from the OpenSMILE data are depicted.

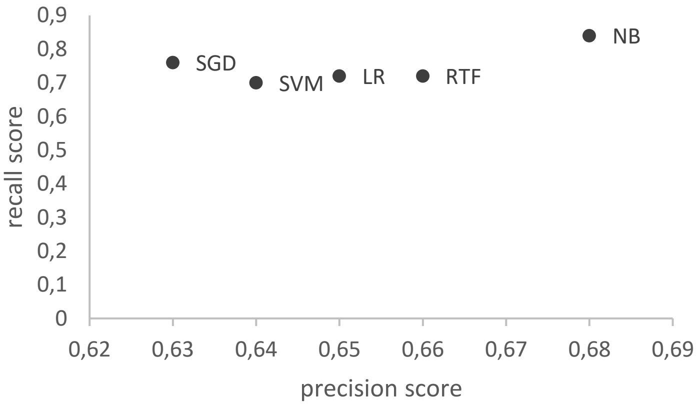

Figure 4.6. Recall and precision scores of the optimized models on OpenSMILE data subset.

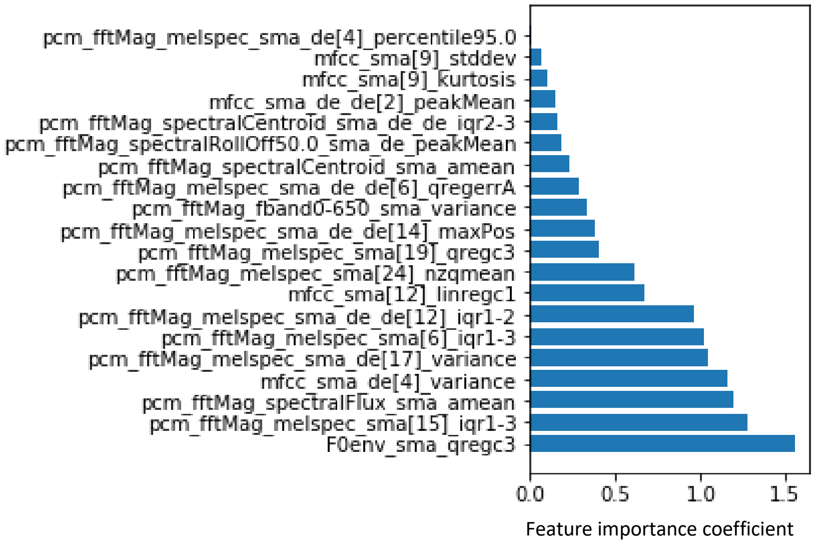

Figure 4.7. Feature importance coefficients for the top 20 variables of the OpenSMILE data of a Support Vector Machine with linear kernel.

4.4 – Combining the domains

The accuracy, precision and recall scores for each classifier with optimal parameter setting can be found in appendix 4.A. This information was used to identify the predicted class of each subject for each classifier, of which the results can be found in appendix 4.B. From these results, we inferred for each participant the percentage of correctly predicted class per domain, which is given in table 4.9.

Table 4.9

Percentage of correctly predicted class and modus of predicted classes per subject.

Fold	subjects	Percentage of correct class predictions			M
	class	T-Scan	Word2Vec	OpenSMILE	
1	p	100	40	100	p
	c	100	80	20	c
2	p	100	100	100	p
	c	80	0	60	c
3	p	100	100	100	p
	c	100	100	100	c
4	p	100	60	100	p
	c	20	100	100	c
5	p	100	60	100	p
	c	100	100	100	c
6	p	20	0	100	c
	c	80	100	60	c
7	p	0	100	100	p
	c	100	100	80	c
8	p	80	60	100	p
	c	100	80	80	c
9	p	100	100	100	p
	c	100	100	100	c
10	p	0	40	20	c
	c	20	40	100	p
11	p	60	20	20	c
	c	20	80	0	p
12	p	100	80	100	p
	c	0	100	100	c
13	p	100	100	100	p
	c	80	100	100	c
14	p	40	60	40	c
	c	20	0	80	p
15	p	40	20	40	c
	c	100	100	100	c
16	p	100	100	100	p
	c	100	100	100	c
17	p	0	0	80	c
	c	100	100	100	c
18	p	100	80	100	p
	c	40	100	100	c
19	p	100	100	40	p
	c	100	100	80	c
20	p	100	80	80	p
	c	100	100	100	c
21	p	0	20	0	c
	c	20	100	100	c
22	p	20	80	40	c
	c	60	100	100	c
23	p	40	100	40	c

	c	100	60	100	c
24	p	60	0	80	p
	c	100	100	80	c
25	p	0	0	20	c
	c	100	60	80	c
26	p	100	100	100	p
	c	80	80	80	c
27	p	100	80	100	p
	c	60	80	0	c
28	p	60	0	20	c
	c	100	80	60	c
29	p	100	100	60	p
	c	100	40	80	c
30	p	40	80	20	c
	c	60	100	80	c
31	p	100	80	100	p
	c	100	100	80	c
32	p	0	60	0	c
	c	100	80	100	c
33	p	100	60	100	p
	c	20	100	100	c
34	p	100	100	80	p
	c	100	60	80	c
35	p	40	20	100	c
	c	100	100	20	c
36	p	100	100	100	p
	c	100	100	100	c
37	p	100	60	60	p
	c	100	20	60	c
38	p	0	40	80	c
	c	100	60	40	c
39	p	100	100	100	p
	c	100	100	100	c
40	p	60	40	40	c
	c	100	40	60	c
41	p	60	20	100	p
	c	100	80	100	c
42	p	0	40	20	c
	c	100	80	60	c
43	p	0	0	100	c
	c	100	100	100	c
44	p	100	100	100	p
	c	80	100	80	c
45	p	100	60	100	p
	c	80	100	100	c
46	p	80	100	100	p
	c	100	80	40	c
47	p	40	80	100	p
	c	100	80	0	c
48	p	100	100	100	p

	c	60	60	80	c
49	p	20	80	100	p
	c	100	20	20	p
50	p	100	100	100	p
	c	20	100	60	c

p stands for patient group, c for control group. M stands for the mode has as value the class that was predicted most often over all classifiers and domains for that subject.

Table 4.9 shows that for 10 subjects from the psychosis group and 11 subjects from the control group, the predicted class is correct for every classifier in each domain. There are 3 subjects, all from the patient group, that received only 0 to 40 percent correctly predicted classes for each domain of speech. It is striking that a low percentage for one domain of speech often occurs together with a high percentage for another domain of speech. Accuracy, recall and precision scores are calculated for the mode of all classifiers taken over the three domains; accuracy is 0.78, recall is 0.64 and precision is 0.89.

4.5 – Statistical analysis

4.5.1 - Syntax

In table 4.10, the results of independent t-tests on the selected T-Scan features are reported.

Table 4.10

Results of independent t-tests on T-Scan data for patients and control group

Variable name	Patients group		Control group			
	M	SD	M	SD	t-score	p-value
D_level	1.825	0.819	2.577	0.747	4.796	<0.001*
Al_gem	0.970	0.317	1.302	0.306	5.332	<0.001*
Onbep_nwg_dz	0.140	0.068	0.175	0.552	2.856	0.005*
Conn_temp_dz	0.102	0.067	00113	0.040	1.010	0.315
Conn_reeks_zin_dz	0.068	0.036	0.076	0.035	1.010	0.315
Conn_contr_dz	0.068	0.031	0.089	0.029	3.581	0.001*
Conn_caus_dz	0.126	0.057	0.163	0.046	3.647	<0.001*
Ww_tt_p	82.894	13.785	87.597	12.812	1.767	0.080
Vd_vrij_dz	0.126	0.050	0.129	0.045	0.356	0.723
Inhwrd_d	473.256	25.508	478.914	16.421	1.319	0.190
Pv_Frog_d	127.776	10.326	128.257	11.363	0.222	0.825
Ontk_tot_d	25.289	8.987	22.266	7.202	1.856	0.066

Pers_vnw1_d	92.599	16.349	85.637	13.264	2.338	0.021*
Pers_vnw3_d	8.544	5.429	7.701	4.280	0.863	0.390
Bvnw_d	88.291	15.487	93.219	15.320	1.599	0.113
Vg_d	73.245	14.300	77.388	11.802	1.580	0.117
Vnw_d	192.313	16.013	193.234	11.552	0.330	0.742
Lidw_d	45.889	12.137	50.795	8.546	2.337	0.021*
Tuss_d	58.354	26.856	33.404	13.236	5.892	<0.001*
Int_bvnw_d	18.998	8.550	20.405	9.154	0.795	0.429
Alg_bijw_d	114.466	23.403	116.462	20.048	0.458	0.648
Spec_bijw_d	26.488	8.635	27.778	6.943	0.823	0.412

M stands for mean, SD for standard deviation. Scores for independent t-test where equal variances for patient group and control group are assumed are reported under t-score. Degrees of freedom is 98 for each entry. P-values smaller than 0.05 are highlighted.

Significant differences between groups are found for *D_level, Al_gem, Onbep_nwg_dz, Conn_contr_dz*, *Conn_caus_dz, Lidw_d*, and *Tuss_d*.

The *D_level* feature measures syntactic complexity and is the second most important feature in the classification process of an SVM with linear kernel. From our literature review, we expected schizophrenia patients to utter syntactically less complex sentences. Our findings are in accordance with this expectation, as *D_level* is significantly lower for the patient group (M = 1.825, SD = 0.819) than for the control group (M = 2.577, SD = 0.747; t(98) = 4.736, p < 0.001), indicating that psychotic patients use less complex sentence structures.

The density of articles, measured by *Lidw_d*, is the most important classification feature for an SVM with linear kernel. Density of articles is significantly lower for the patient group (M = 0.970, SD = 0.317) than for the control group (M = 1.302, SD = 0.306; t(98) = 5.332, p < 0.001).

We also expected differences in the use of other parts of speech, namely decreased use of determiner pronouns, especially first-person, nouns and adjectives. Although psychotic patients use significantly more personal pronouns (M = 92.599, SD = 16.349) than controls (M = 85.637, SD = 13.264; t(38) = 2.338, p = 0.021), no significant difference is found for third-person pronouns (t(38) = 0.863, p = 0.390). Patients use significantly more interjections (M = 58.354, SD = 26.856) than controls (M = 33.404, SD = 13.236; t(98) = 5.892, p < 0.001).

Lastly, *Al_gem* is significantly lower for psychotic speech (M = 0.970, SD = 0.317) than for normal speech (M = 1.302, SD = 0.306; t(98) = 5.332, p < 0.001). As *Al_gem* measures the distance between to clauses that are dependent of each other, this implies that psychotic patients utter sentences of a simpler structure, which is conform the view that psychotic speech is syntactically less complex.

4.5.2 – Semantics

Table 4.11 contains the results of independent t-tests conducted on the selected Word2Vec features.

Table 4.11

Results of independent t-tests on Word2Vec data for patients and control group

Variable name	Patients group		Control group			
	M	SD	M	SD	t-score	p-value
max_simple_17	0.771	0.0280	0.769	0.0219	-0.456	0.649
max_simple_18	0.764	0.0210	0.771	0.0280	-0.785	0.435
max_summary_10	0.768	0.0271	0.764	0.0210	1.31	0.194
max_summary_12	0.968	0.0060	0.768	0.0271	1.40	0.166
max_summary_17	0.966	0.0091	0.968	0.0060	1.13	0.263
mean_summary_2	0.971	0.0055	0.966	0.0091	-1.76	0.081
mean_summary_14	0.969	0.0096	0.971	0.0055	0.313	0.755
min_simple_3	0.975	0.0046	0.969	0.0096	-1.27	0.208
min_simple_8	0.973	0.0063	0.975	0.0046	-3.02	0.003*
min_simple_20	0.696	0.0094	0.973	0.0063	-2.27	0.025*
min_summary_9	0.700	0.0108	0.696	0.0094	-0.444	0.658
min_summary_17	0.930	0.0040	0.700	0.0108	-0.299	0.766
min_summary_20	0.930	0.0049	0.930	0.0040	-1.01	0.317
var_simple_5	0.489	0.0203	0.930	0.0049	-3.67	<0.001*
var_simple_9	0.494	0.0223	0.489	0.0203	-3.77	<0.001*
var_simple_10	0.361	0.0181	0.494	0.0223	-4.04	<0.001*
var_simple_14	0.375	0.0264	0.361	0.0181	-3.70	<0.001*
var_simple_17	0.313	0.0226	0.375	0.0264	-3.52	0.001*
var_simple_19	0.325	0.0311	0.313	0.0226	-3.55	0.001*
var_summary_5	0.786	0.0247	0.325	0.0311	-0.581	0.562

M stands for mean, SD for standard deviation. Scores for independent t-test where equal variances for patient group and control group are assumed are reported under t-score. Degrees of freedom is 98 for each entry. P-values smaller than 0.05 are highlighted.

Min_simple_8 is significantly lower for the patient group (M = 0.976, SD = 0.006) than for the control group (M = 0.975, SD = 0.005; t(98) = -3.02, p = 0.003), as is the case for *min_simple_20* ($M_{patients}$ = 0.696, $SD_{patients}$ = 0.009; $M_{controls}$ = 0.973, $SD_{controls}$= 0.006; t(98) = -2.27, p = 0.025). However, no significant differences were found for maximum or mean simple coherence measures. As *min_sample_8* and *min_sample_20* are measures for the lowest coherence for an individual's speech within a window of respectively 8 and 20 words, these findings suggest that speech from subjects with a diagnosis in the psychosis spectrum contains at least some utterances with lower coherence than can be found in normal speech.

4.5.3 – OpenSMILE

Table 4.12 contains the results of the independent t-tests for the selected openSMILE features.

Table 4.12

Results of independent t-tests on OpenSMILE data for patients and control group

Variable name	Patient group		Control group			
	M	SD	M	SD	T-score	P-value
mfcc_sma[9]_stddev	5.77	0.904	5.78	0.600	0.025	0.980
mfcc_sma[9]_kurtosis	3.43	0.373	3.44	0.351	0.218	0.828
mfcc_sma[12]_linregc1	6.65x10^{-5}	1.40x10^{-4}	1.95x10^{-5}	5.53x10^{-5}	2.12	0.037
pcm_fftMag_melspec_sma[6]_iqr1-3	1.14x10^{4}	1.54x10^{4}	1.64x10^{4}	1.84x10^{4}	1.48	0.142
pcm_fftMag_melspec_sma[15]_iqr1-3	1.42x10^{4}	2.57 x10^{4}	1.56x10^{4}	1.54x10^{4}	0.332	0.740
pcm_fftMag_melspec_sma[19]_qregc3	1.87x10^{4}	2.50 x10^{4}	3.23x10^{4}	6.87x10^{4}	1.32	0.191
pcm_fftMag_melspec_sma[24]_nzqmean	2.46x10^{9}	7.43x10^{9}	8.73x10^{10}	6.01x10^{11}	0.998	0.321
F0env_sma_qregc3	355	63.7	350	45.5	-0.444	0.658
pcm_fftMag_fband0-650_sma_variance	1.53x10^{-4}	5.23x10^{-4}	2.87x10^{-4}	8.44x10^{-4}	0.955	0.342
pcm_fftMag_spectralFlux_sma_amean	0.0149	0.0184	0.0359	0.117	1.25	0.213
pcm_fftMag_spectralCentroid_sma_amean	7.05x10^{3}	2.55x10^{3}	5.66x10^{3}	1.32x10^{3}	-3.41	0.001
mfcc_sma_de[4]_variance	2.12	0.532	2.54	0.630	3.60	0.001
pcm_fftMag_melspec_sma_de[4]_percentile95.0	6.03x10^{3}	6.74x10^{3}	9.03x10^{3}	8.23x10^{3}	1.99	0.049
pcm_fftMag_melspec_sma_de[17]_variance	1.04x10^{8}	3.39x10^{8}	1.07x10^{8}	2.47 x10^{8}	0.048	0.962
pcm_fftMag_spectralRollOff50.0_sma_de_peakMean	2.03x10^{3}	489	1.88x10^{3}	185	-1.93	0.057
mfcc_sma_de_de[2]_peakMean	0.631	0.058	0.678	0.0421	4.63	<0.001*
pcm_fftMag_melspec_sma_de_de[6]_qregerrA	4.69 x10^{7}	7.71x10^{7}	8.84x10^{7}	9.41 x10^{7}	2.41	0.018

pcm_fftMag_melspec_sma_de_de[12]_iqr1-2	120	245	161	190	0.939	0.350
pcm_fftMag_melspec_sma_de_de[14]_maxPos	2.73x10^{4}	2.03x10^{4}	2.70x10^{4}	1.97x10^{4}	-0.055	0.956
pcm_fftMag_spectralCentroid_sma_de_de_iqr2-3	94.6	29.9	95.3	19.9	0.137	0.891

M stands for mean, SD for standard deviation. Scores for independent t-test where equal variances for patient group and control group are assumed are reported under t-score. Degrees of freedom is 98 for each entry. P-values smaller than 0.05 are highlighted

Mfcc_sma[12]_linregc1 is higher for the patient group (M = 6.65x10^{-5}, SD = 1.40x10^{-4}) than for the control group (M = 1.95x10^{-5}, SD = 5.53x10^{-5}; t(98) = 2.12, p = 0.037), as is the case for *mfcc_sma_de_de[2]_peakMean* (M$_{\text{patient}}$ = 0.631, SD$_{\text{patiens}}$ = 0.058; M$_{\text{controls}}$ = 0.0421, SD$_{\text{controls}}$ = 4.63; t(98) = 4.63, p < 0.001). *Mfcc_sma_de[4]_variance* is lower for the patients group *(*M= 2.12, SD = 0.532) than for the control group (M = 2.54, SD = 0.630; t(98) = 3.60, p = 0.001). Mfcc is a measure for Mel-Frequency-Cepstral Coefficients and is a representation of the short-term power spectrum of a sound (Eyben, Wöllmer, & Schuller, 2010). In sound analysis, mfcc is often used to describe timbre, a feature that is observed trough the presence and absence of many different properties of sound and cannot be linked directly to one physical dimension (Cosi, De Goli, & Prandoni, 1994). These results suggest that the timbre of speech of psychosis patients differs from controls.

Pcm_fftMag_spectralCentroid_sma_amean is higher for the patient group (M = 7.05x10^{3}, SD = 2.55x10^{3}) than for the control group (M = 5.66x10^{3}, SD = -3.41; t(98) = -3.41, p = 0.001*), pcm_fftMag_melspec_sma_de[4]_percentile95.0* is lower for the patient group (M = 6.03x10^{3}, SD = 6.74x10^{3}) than for the control group (M = 9.03x10^{3}, SD = 8.23x10^{3}; t(98) = 1.99, p = 0.049), as is the case for *pcm_fftMag_melspec_sma_de_de[6]_qregerrA* (M$_{\text{oatients}}$ = 4.69 x10^{7}, SD$_{\text{patients}}$ = 7.71x10^{7}; M$_{\text{controls}}$ = 8.84x10^{7}, SD$_{\text{controls}}$ = 9.41 x10^{7}; t(98) = 2.41, p = 0.018). Pcm stands for pulse code modulation and is a conversion of speech waves into coded pulses (Eyben, Wöllmer, & Schuller, 2010).

Chapter 5

Discussion

We demonstrate the possibility of distinguishing psychotic patients from healthy controls using automatic speech-based techniques on different domains of speech. The findings show that by using syntactic, semantic or phonological information about a person’s speech, reasonably high accuracy scores for classifying psychosis can be achieved. Furthermore, all three domains of speech contain markers that can be used in the classification process. Our findings suggest that psychotic speech is syntactically less complex and is characterized by lower article density, a higher use of personal pronouns and interjections, lower coherence between clauses and deviations in the timbre and wave forms of the sound of speech.

It is striking that deviancies in article use for schizophrenia patients were not found in the studies examining the relationship of syntax and schizophrenia that were included in our literature review. However, participants of these studies spoke either English (Çokal, et al., 2019; Covington, et al., 2005; Deutsch-Link, 2016; Kuperberg, 2010; Stanislawski, 2019) or Polish (Obrebska & Obrebski, 2007). These languages have different grammatical structures and rules than the Dutch language. Besides, not all studies examined the same domain of language; information was extracted from written essays (Deutsch-Link, 2016) and from speech generated during a sentence-picture matching task (Cokal, Zimmerer, Varley, Watson, & Hinzen, 2019) or during open ended interviews (Stanislawski, 2019). The underlying grammatical structure of written and spoken texts is different (O'Donnell, 1974).

From our literature review, we expected psychosis to also be characterized by talk about voices and sounds (Rezaii, Walker, & Wolff, 2019). However, the output generated by Word2Vec did not contain information of this sort. We expect classification to benefit from incorporation of content analysis. For future research, it is recommended to supplement the semantic analysis by performing latent content analysis.

Interpreting the results from the OpenSMILE data set proved to be challenging. Besides performance, interpretability of a classifier is very important. We would suggest training the classifiers on a subset of OpenSMILE features that are highly explicable, for example the low-level descriptors in combination with none or one functional applied to those.

Our results show that some subjects are misclassified by more classifiers than others. Due to time limitations, we have not been able to identify possible similarities for subjects for whom this is the case. We suggest researching the correlation between often misclassified subjects to Positive and Negative Syndrome Scale (PANSS) scores, which provide a rating for the symptoms of schizophrenia (Kay, Fiszbein, & Opler, 1987).

To our knowledge, this is the first study to examine syntactic, as well as semantic and phonologic features of speech as predictors of psychosis on a single data set. The use of 5 different classification algorithms on these domains of speech makes this research a comprehensive one. Taken together, our findings strongly suggest that speech features can serve as an objective classifier for psychosis. The ability to use spontaneous speech that is collected during an interview suggests that clinicians may be able to employ speech-based analyses to aid in the diagnostic process.

There were a number of limitations in this study. First of all, we have conducted internal cross-validation on several levels within the classification process. For the Word2Vec and OpenSMILE data set, a Random Forest classifier was used for feature selection. Performance scores are thus based on the classification of data points that at some point have already been involved in the training of the model. To guarantee generalizability, performance of the final classification model should be evaluated using a test set that consists of never-before seen data. In the case of a larger sample size, a subset of the data set can be held back for this purpose. In future research, a larger sample size is also desirable in order to investigate more of the features that the automated speech analysis tools make available. The size of our data set led us to select only a small subset of features from the available variables to be used in the classification process, with the result that several markers for psychosis were most likely discarded. Performance of the classification models benefits from incorporation of more variables into the design (Jain & Chandrasekaran, 1982). Also, variance can be expected to decrease given a larger training set (Brain & Webb, 1999).

There are numerous methods for feature selection. Which method is chosen greatly impacts the performance of a classification model. For this study, the RF classifier was used to create a subset of 20 features for the Word2Vec data and 19 features for the OpenSMILE data. It is possible that other methods, and consequently other data subsets, would have resulted in better performance of the classifiers. For future research, we recommend implementing the filter or wrapper methods for feature selection. In filter methods, features are selected on the basis of their scores in various statistical tests for their outcome variable. Wrapper methods try to use a subset of features and train a model using that subset. Based on previous models, features are chosen to be added or removed to the subset (Chandrashekar & Sahin, 2014).

Binary classification was used to predict psychosis. A data point was either classified as psychotic or healthy. However, this view of psychotic symptoms either being present or completely absent does not correspond to the disease profile. Symptom severity varies within the patient group, which leads to a large variance. Besides, some psychotic symptoms can to some extent also be present in healthy controls. As a result of these factors, speech of a psychotic patient that shows less severe symptoms could be more similar to that of a participant in the control group than that of another patient. Thus, the perfect dichotomy of participants in psychotic of healthy is not expected to perfectly represent the heterogeneous distribution. We expect that the classification of psychosis benefits from incorporating symptom severity into the classes of the model, or datasets with a clearer division, even if these would be less generalizable across populations.

We are hesitant to state with certainty that we developed a model that only predicts psychosis. As stated before, many psychiatric illnesses are characterized by disturbances in thought and language. Some symptoms of psychotic speech are also markers for other mental disorders. For instance, both PTSD and depression are characterized by slower, more monotonous speech (Marmar et al, 2019; Alpert, Pouget & Silva, 2001). While healthy controls were screened for former or current mental illnesses, the presence of symptoms of depression within the patient group was not ruled out; indeed, depression is one of the components of the PANSS symptom score (Leucht, et al., 2005). For the purpose of creating a device or procedure that can be used to guide diagnosis of mental diseases, future research should focus on incorporating various mental disorders in the classification process, thus using features of speech to give a probability measure for various mental diagnoses based on a speech sample. The use of binary classification could be the reason that performance of the classifiers did not improve significantly by combining syntactic, semantic and phonological features. It is possible that the division of the participants into the two classes prevents better classification, because the underlying division is not binary.

Keeping these limitations in mind, we believe that our findings demonstrate the usability of automated speech-based techniques in predicting psychosis. We have shown that all examined domains of speech contain markers for psychosis. This study contributes to our understanding of the exact ways that speech is deviant in schizophrenia patients. Our results can be used to guide feature selection in similar studies in the future.

References

Alberto, P., Arndis, S., Vibeke, B., & Riccardo, F. (2019). Voice patterns in schizophrenia: A systematic review and Bayesian meta-analysis. *Unpublished manuscript*.

Alpert, M., Pouget, E., & Silva, R. R. (2001). Reflections of depression in acoustic measures of the patient's speech. *Journal of affective disorders* , 59-69.

Bar, K. Z. (2019). Semantic Characteristics of Schizophrenic Speech.

Boersma, P., & Weenink, D. J. (2013). Praat: doing phonetics by computer. *Amsterdam: Insitute of Phonetic Sciences of the University of Amsterdam*.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT'2010*, 177-186.

Brain, D., & Webb, F. I. (1999). On the effect of data set size on bias and variance in classification learning. *Proceedings of the Fourth Australian Knowledge Acquisition Workshop*, 117-128.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, 16-28.

Cohen, A. S., & Elvevag, B. (2014). Automated Computerized Analysis of Speech in Psychiatric Disorders. *Current opinion in Psychiatry*, 203.

Cohen, A. S., Kim, Y., & Najolia, G. M. (2013, May 1). Psychiatric Symptom versus Neurocognitive Correlates of Diminished Expressivity in Schizophrenia and Mood Disorders. *Schizophrenia research*, 249-253.

Çokal, D., Zimmerer, V., Varley, R., & Watson, S. H. (2019). Comprehension of Embedded Clauses in Schizophrenia With and Without Formal Thought Disorder. *The Journal of Nervous and Mental Disease*, 384-392.

Cokal, D., Zimmerer, V., Varley, R., Watson, S., & Hinzen, W. (2019). Comprehension of Embedded Clauses in Schizophrenia with and Without Formal Thought Disorder. *The Journal of Nervous and Mental Disease*, 384-392.

Cosi, P., De Goli, G., & Prandoni, P. (1994). Timbre Characterization with Mel-Cepstrum and Neural Nets. *ICMC*.

Covington, M. A., He, C., Brown, C., Naçi, L., McClain, J. T., Fjordbak, B. S., . . . Brown, J. (2005). Schizophrenia and the Structure of Language: The Linguist's View. *Schizophrenia Research*, 85-98.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 10-49.

de Boer, J. N., Voppel, A. E., Begemann, M. J., Schnack, H. G., Wijnen, F., & Sommer, I. E. (2018). Clinical use of semantic space models in psychiatry and neurology: a systematic review and meta-analysis. *Neuroschience & Biobehavioral Reviews*, 85-92.

Deutsch-Link, S. (2016). Language In Schizophrenia: What We Can Learn From Quantitative Text Analysis.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. *Proceedings of the 9th ACM International Conference on Multimedia.* Firenze, Italy.

Fatouros-Bergman, H. C. (2014, July 30). Meta-analysis of cognitive performance in drug-naïve patients with schizophrenia. pp. 156-162.

Fineberg, S. K., Deutsch-Link, S., Ichnose, M., McGuinness, T., Bessette, A. J., Chunch, C. K., & Corlett, P. R. (2015). Word Use in First-Person Accounts of Schizophrenia. *The British Journal of Psychiatry*, 32-38.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data. *Bioinformatics*, 906-914.

Gnana, D. A., Balamurugan, S. A., & Leavline, E. J. (2016). Literature review on feature selection methods for high-dimensional data. *International Journal of Computer Applications*, 8887.

Hoffman, R. E., & Sledge, W. (1988). An Analysis of Grammatical Deviance Occuring in Spontaneous Schizophrenic Speech.

Insel, T. C. (2010). Research Domain Criteria (RDoC) Toward a New Classification Framework for Research on Mental Disorders.

Jain, A. K., & Chandrasekaran, B. (1982). 39 Dimensionality and sample size considerations in pattern recognition practice. *Handbook of statistics*, 835-855.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.* Springer.

Jurafsky, D., & Martin, J. H. (2009). Speech and Language Processing. In *Speech and Language Processing* (p. 419). New Jersey: Pearson International Edition.

Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, 261-276.

Kuperberg, G. R. (2010). Language in Schizophrenia Part 1: An Introduction. *Language and Linguistics Compass*, 576-589.

Lavesson, N., & Davidsson, P. (2006). Quantifying the impact of learning algorithm parameter tuning. *AAAI*, 395-400.

Leucht, S., Kane, J. M., Kissling, W., Hamann, J., Etschel, E., & Engel, R. R. (2005). What does the PANSS mean? *Schizophrenia research*, 231-238.

M. J. Kas, B. P. (2019). A Quantitative Approach to Neuropsychiatry: the Why and the How.

Maat, H. P., Kraf, R., & Dekker, N. (2017). *Handleiding T-Scan.* Utrecht, Nijmegen, Tilburg.

MacWhinney, B. (2014). The CHILDES project: Tools for analyzing talk, volume II: The Database. *Psychology Press*.

MacWhinney, B., & Wagner, J. (2014). Transcribing, searching and data sharing: the CLAN software and the TalkBank data repository. *Gesprachsforschung*, 154-173.

Magaud, E., Kebir, O., Gut, A., Willard, D., Chauchot, F., Ollie, J., & Krebs, M. (2010). Altered semantic but not phonological verbal fluency in young help-seeking individuals with ultra high risk of psychosis. *Schizophrenia research*, 53-58.

Marmar, C. R., Brown, A. D., Qian, M., Laska, E., Siegel, C., Li, M., . . . Vergyri, D. (2019). Speech-based markers for posttraumatic stress disorder in US veterans. *Depression and Anxiety*.

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*.

Millon, T., & Davis, R. O. (1996). *Disorders of personality: DSM-IV and beyond.* John Wiley & Sons.

National Collaborating Centre for Mental Health. (2014). Psychosis and schizophrenia in adults. *Psychosis and Schizophrenia in Adults: Treatment and Management: Updated Edition 2014.* .

Obrebska, M., & Obrebski, T. (2007). Lexical and Grammatical Analysis of Schizophrenic Patients' Language: A Preliminary Report. *Psychology of Language and Communication*.

O'Donnell, R. C. (1974). Syntactic differences between speech and writing. *American Speech*, 102-110.

Oostdijk, N. H. (2000). Het corpus gesproken Nederlands.

Pal, M. (2005). Random Forest Classifier for Remote Sensing Classification. *International Journal of Remote Sensing*, 217-222.

Polikar, R. (2012). Ensemble learning. *Ensemble machine learning*, 1-34.

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Rezaii, N., Walker, E., & Wolff, P. (2019). A machine learning approach to predicting psychosis using semantic density and latent content analysis. *NPJ schizophrenia*.

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 41-46.

Sheenan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P. J., Weiller, E., & Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The journal of clinical psychiatry*.

Stanislawski, E. B. (2019, April). Analyzing Negative Symptoms and Language in Youths at Risk for Psychosis Using Automated Language Analysis. *Schizophrenia Bulletin*.

Vehtari, A., Gelman, A., & Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*.

Wijnhoven, R. G., & de With, P. H. (2010). Fast training of object detection using stochastic gradient descent. *20th International Conference on Pattern Recognition*, 424-427.

Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent Developments in OpenSMILE, the Munich Open-Source Multimedia Feature Extractor. *ACM international conference on Multimedia* (pp. 835 - 838). Barcelona: ACM New York.

Maat, H. P., Kraf, R., & Dekker, N. (2017). *Handleiding T-Scan.* Utrecht, Nijmegen, Tilburg.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in neural information processing systems*, 3111-3119.

Rish, I. (2001). An emperical study of the naive Bayes classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 41- 46.

Appendices

This thesis comes with python scripts for implementation of the machine learning models, that are written by myself for this research. These scripts are not included here because of their size.

Appendix 4.A.A

Accuracy (acc), precision (prec) and recall (rec) scores for each fold in the leave-two-out cross-validation process of all classifiers on the T-Scan data set.

	Logistic Regression			Naive Bayes			Random Forest			Stochastic Gradient Descent			Support Vector Machine		
Fold	Acc	Prec	Rec	Acc	Prec	Rec	Acc	Prec	Rec	Acc	Prec	Rec	Acc	Prec	Rec
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	0.5	0.5	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	0.5	0.5	1	0.5	0.5	1	0.5	0.5	1	1	1	1	0.5	0.5	1
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0.5	1	0.5	0	0
7	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0
8	1	1	1	1	1	1	0.5	0	0	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0
11	0.5	0.5	1	0	0	0	0.5	0	0	0.5	0.5	1	0.5	0.5	1
12	0.5	0.5	1	0.5	0.5	1	0.5	0.5	1	0.5	0.5	1	0.5	0.5	1
13	1	1	1	1	1	1	0.5	0.5	1	1	1	1	1	1	1
14	0.5	0.5	1	0	0	0	1	1	1	0	0	0	0	0	0
15	1	1	1	0.5	0	0	0.5	0	0	0.5	0	0	1	1	1
16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
17	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0
18	0.5	0.5	1	0.5	0.5	1	1	1	1	1	1	1	0.5	0.5	1
19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
21	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0.5	0	0	0.5	0	0	0.5	0.5	1	0.5	0	0
23	0.5	0	0	1	1	1	1	1	1	0.5	0	0	0.5	0	0
24	1	1	1	0.5	0	0	0.5	0	0	1	1	1	1	1	1
25	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0
26	1	1	1	1	1	1	0.5	0.5	1	1	1	1	1	1	1
27	1	1	1	0.5	0.5	1	1	1	1	0.5	0.5	1	1	1	1
28	1	1	1	0.5	0	0	0.5	0	0	1	1	1	1	1	1
29	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
30	0.5	0	0	0	0	0	0.5	0.5	1	1	1	1	0.5	0	0
31	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
32	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0
33	0.5	0.5	1	0.5	0.5	1	1	1	1	0.5	0.5	1	0.5	0.5	1
34	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
35	1	1	1	0.5	0	0	0.5	0	0	0.5	0	0	1	1	1
36	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
37	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

38	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0
39	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
40	0.5	0	0	1	1	1	1	1	1	0.5	0	0	1	1	1
41	0.5	0	0	1	1	1	1	1	1	1	1	1	0.5	0	0
42	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0
43	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0
44	1	1	1	1	1	1	0.5	0.5	1	1	1	1	1	1	1
45	1	1	1	1	1	1	0.5	0.5	1	1	1	1	1	1	1
46	1	1	1	1	1	1	1	1	1	1	1	1	0.5	0	0
47	0.5	0	0	0.5	0	0	1	1	1	1	1	1	0.5	0	0
48	0.5	0.5	1	0.5	0.5	1	1	1	1	1	1	1	1	1	1
49	0.5	0	0	0.5	0	0	1	1	1	0.5	0	0	0.5	0	0
50	0.5	0.5	1	0.5	0.5	1	1	1	1	0.5	0.5	1	0.5	0.5	1

Appendix 4.A.B

Accuracy (acc), precision (prec) and recall (rec) scores for each fold in the leave-two-out cross-validation process of all classifiers on the Word2Vec data set.

	Logistic Regression			Naive Bayes			Random Forest			Stochastic Gradient Descent			Support Vector Machine		
Fold	acc	prec	rec	acc	prec	rec	acc	prec	rec	acc	prec	rec	acc	prec	rec
1	0.5	0	0	0.5	0	0	1	1	1	0.5	0.5	1	0.5	0	0
2	0.5	0.5	1	0.5	0.5	1	0.5	0.5	1	0.5	0.5	1	0.5	0.5	1
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	1	1	1	0.5	0	0	0.5	0	0	1	1	1	1	1	1
5	0.5	0	0	1	1	1	1	1	1	0.5	0	0	1	1	1
6	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	0.5	0	0	1	1	1	1	1	1	0.5	0	0	0.5	0	0
9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	0.5	0.5	1	0	0	0	0.5	0	0	0	0	0	0	0	0
11	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0.5	1	0.5	0	0
12	1	1	1	1	1	1	1	1	1	0.5	0	0	1	1	1
13	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	0.5	0.5	1	0	0	0	0	0	0	0.5	0.5	1	0.5	0.5	1
15	0.5	0	0	0.5	0	0	0.5	0	0	1	1	1	0.5	0	0
16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
17	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0
18	0.5	0	0	1	1	1	1	1	1	1	1	1	1	1	1
19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	1	1	1	1	1	1	1	1	1	0.5	0	0	1	1	1
21	0.5	0	0	1	1	1	0.5	0	0	0.5	0	0	0.5	0	0
22	1	1	1	1	1	1	1	1	1	0.5	0	0	1	1	1
23	1	1	1	0.5	0.5	1	0.5	0.5	1	1	1	1	1	1	1
24	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0
25	0	0	0	0.5	0	0	0.5	0	0	0	0	0	0.5	0	0
26	1	1	1	1	1	1	0.5	0.5	1	1	1	1	1	1	1
27	0.5	0.5	1	1	1	1	1	1	1	0.5	0	0	1	1	1
28	0	0	0	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0
29	0.5	0.5	1	1	1	1	0.5	0.5	1	1	1	1	0.5	0.5	1
30	1	1	1	1	1	1	1	1	1	0.5	0	0	1	1	1
31	1	1	1	1	1	1	1	1	1	0.5	0	0	1	1	1
32	0.5	0	0	1	1	1	1	1	1	0.5	0.5	1	0.5	0	0
33	1	1	1	0.5	0	0	0.5	0	0	1	1	1	1	1	1
34	0.5	0.5	1	1	1	1	1	1	1	0.5	0.5	1	1	1	1
35	0.5	0	0	0.5	0	0	1	1	1	0.5	0	0	0.5	0	0
36	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
37	0.5	0.5	1	0	0	0	0.5	0	0	0.5	0.5	1	0.5	0.5	1
38	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0

39	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
40	0	0	0	0.5	0.5	1	0	0	0	0.5	0	0	0	0	0
41	1	1	1	0.5	0	0	0	0	0	0.5	0	0	0.5	0	0
42	0.5	0.5	1	0.5	0	0	1	1	1	0.5	0	0	0.5	0	0
43	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0
44	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
45	1	1	1	0.5	0	0	0.5	0	0	1	1	1	1	1	1
46	1	1	1	1	1	1	1	1	1	0.5	0.5	1	1	1	1
47	1	1	1	0.5	0	0	1	1	1	0.5	0.5	1	1	1	1
48	1	1	1	0.5	0.5	1	0.5	0.5	1	1	1	1	1	1	1
49	0.5	0.5	1	0.5	0	0	0.5	0.5	1	0.5	0.5	1	0.5	0.5	1
50	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Appendix 4.A.C

Accuracy (acc), precision (prec) and recall (rec) scores for each fold in the leave-two-out cross-validation process of all classifiers on the OpenSMILE data set.

	Logistic Regression			Naive Bayes			Random Forest			Stochastic Gradient Descent			Support Vector Machine		
Fold	Acc	Prec	Rec	Acc	Prec	Rec	Acc	Prec	Rec	Acc	Prec	Rec	Acc	Prec	Rec
1	0.5	0.5	1	0.5	0.5	1	1	1	1	0.5	0.5	1	0.5	0.5	1
2	1	1	1	0.5	0.5	1	0.5	0.5	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	0.5	0.5	1	1	1	1	1	1	1	1	1	1	0.5	0.5	1
7	1	1	1	1	1	1	1	1	1	0.5	0.5	1	1	1	1
8	1	1	1	0.5	0.5	1	1	1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	0.5	0	0	1	1	1	0.5	0	0	0.5	0	0	0.5	0	0
11	0	0	0	0.5	0.5	1	0	0	0	0	0	0	0	0	0
12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
13	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	1	1	1	0	0	0	1	1	1	0.5	0	0	0.5	0	0
15	1	1	1	1	1	1	0.5	0	0	0.5	0	0	0.5	0	0
16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
17	1	1	1	1	1	1	1	1	1	0.5	0	0	1	1	1
18	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
19	0.5	0	0	1	1	1	0.5	0	0	0.5	0.5	1	0.5	0	0
20	1	1	1	1	1	1	0.5	0	0	1	1	1	1	1	1
21	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0
22	0.5	0	0	1	1	1	0.5	0	0	0.5	0	0	1	1	1
23	1	1	1	0.5	0	0	0.5	0	0	0.5	0	0	1	1	1
24	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1
25	0	0	0	0.5	0	0	1	1	1	0.5	0	0	0.5	0	0
26	1	1	1	1	1	1	1	1	1	0.5	0.5	1	1	1	1
27	0.5	0.5	1	0.5	0.5	1	0.5	0.5	1	0.5	0.5	1	0.5	0.5	1
28	0.5	0	0	0.5	0.5	1	0	0	0	0.5	0	0	0.5	0	0
29	0.5	0	0	1	1	1	1	1	1	0.5	0.5	1	0.5	0	0
30	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0.5	1	0.5	0	0
31	1	1	1	1	1	1	1	1	1	0.5	0.5	1	1	1	1
32	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0	0.5	0	0
33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
34	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1
35	0.5	0.5	1	0.5	0.5	1	1	1	1	0.5	0.5	1	0.5	0.5	1
36	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
37	0.5	0	0	0.5	0.5	1	0.5	0	0	1	1	1	0.5	0	0
38	0.5	0	0	0.5	0.5	1	1	1	1	0.5	0.5	1	0.5	0	0

39	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
40	0.5	0	0	0.5	0.5	1	0.5	0.5	1	0.5	0	0	0.5	0	0
41	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
42	0.5	0	0	0	0	0	0.5	0	0	0.5	0.5	1	0.5	0	0
43	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
44	1	1	1	0.5	0.5	1	1	1	1	1	1	1	1	1	1
45	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
46	0.5	0.5	1	0.5	0.5	1	1	1	1	1	1	1	0.5	0.5	1
47	0.5	0.5	1	0.5	0.5	1	0.5	0.5	1	0.5	0.5	1	0.5	0.5	1
48	1	1	1	0.5	0.5	1	1	1	1	1	1	1	1	1	1
49	0.5	0.5	1	0.5	0.5	1	0.5	0.5	1	0.5	0.5	1	1	1	1
50	1	1	1	0.5	0.5	1	0.5	0.5	1	1	1	1	1	1	1

Appendix 4.B
Predicted labels for each combination of classifier and data set, modus and percentage of correctly classified labels for each data point.

		T-Scan					Word2Vec					OpenSmile					**M**	**%**
fold	point	LR	NB	RF	SGD	SVM	LR	NB	RF	SGD	SVM	LR	NB	RF	SGD	SVM		
1	1	1	1	1	1	1	0	0	1	1	0	1	1	1	1	1	1	0.800
	0	0	0	0	0	0	0	0	0	1	0	1	1	0	1	1	0	0.667
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1.000
	0	0	0	0	1	0	1	1	1	1	1	0	1	1	0	0	1	0.467
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1.000
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.000
4	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	0.867
	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0.733
5	1	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	0.867
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.000
6	1	0	0	0	1	0	0	0	0	0	0	1	1	1	1	1	0	0.400
	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0.800
7	1	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0.667
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0.933
8	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	0.800
	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0.867
9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1.000
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.000
10	1	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0.200
	0	1	1	1	0	1	1	1	0	1	0	0	0	0	0	0	0	0.533
11	1	1	0	0	1	1	0	0	0	1	0	0	1	0	0	0	0	0.333
	0	1	1	0	1	1	0	0	0	1	0	1	1	1	1	1	1	0.333
12	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0.933
	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0.667
13	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1.000
	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.933
14	1	1	0	1	0	0	1	0	0	1	1	1	0	1	0	0	0	0.467
	0	1	1	0	1	1	1	1	1	1	1	0	1	0	0	0	1	0.333
15	1	1	0	0	0	1	0	0	0	1	0	1	1	0	0	0	0	0.333
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.000
16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1.000
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.000
17	1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0.267
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.000
18	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0.933
	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0.800
19	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	0.800
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0.933
20	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0.867
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.000
21	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0.067
	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0.733
22	1	0	0	0	1	0	1	1	1	0	1	0	1	0	0	1	0	0.467
	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0.867
23	1	0	1	1	0	0	1	1	1	1	1	1	0	0	0	1	1	0.600
	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0.867
24	1	1	0	0	1	1	0	0	0	0	0	1	0	1	1	1	0	0.467
	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0.933
25	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.067
	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0.800
26	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1.000
	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0.800
27	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0.933
	0	0	1	0	1	0	1	0	0	0	0	1	1	1	1	1	1	0.467

28	1	1	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0.267
	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0.800
29	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	0.867
	0	0	0	0	0	0	1	0	1	0	1	0	0	0	1	0	0	0.733
30	1	0	0	1	1	0	1	1	1	0	1	0	0	0	1	0	0	0.467
	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0.800
31	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0.933
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0.933
32	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0.200
	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0.933
33	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	0.867
	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0.733
34	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0.933
	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0.800
35	1	1	0	0	0	1	0	0	1	0	0	1	1	1	1	1	1	0.533
	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	0.733
36	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1.000
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.000
37	1	1	1	1	1	1	1	0	0	1	1	1	1	0	1	0	1	0.733
	0	0	0	0	0	0	1	1	0	1	1	1	1	0	0	0	0	0.600
38	1	0	0	0	0	0	0	1	0	0	1	1	1	1	1	0	0	0.400
	0	0	0	0	0	0	0	1	0	0	1	1	1	0	1	0	0	0.667
39	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1.000
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.000
40	1	0	1	1	0	1	0	1	0	0	1	0	1	1	0	0	0	0.467
	0	0	0	0	0	0	1	1	1	0	0	0	1	1	0	0	0	0.667
41	1	0	1	1	1	0	1	0	0	0	0	1	1	1	1	1	1	0.600
	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0.933
42	1	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0.200
	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0.800
43	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0.333
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.000
44	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1.000
	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0.867
45	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	0.867
	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.933
46	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0.933
	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	0.733
47	1	0	0	1	1	0	1	0	1	1	1	1	1	1	1	1	1	0.733
	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	0	0.600
48	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1.000
	0	1	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0.667
49	1	0	0	1	0	0	1	0	1	1	1	1	1	1	1	1	1	0.667
	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	0	1	0.467
50	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1.000
	0	1	1	0	1	1	0	0	0	0	0	0	1	1	0	0	0	0.600

LR: Logistic Regression; NB: Naïve Bayes; RF: Random Forest; SGD: Stochastic Gradient Descent; SVM: Support Vector Machine; M: modus; %: percentage of classifiers that correctly labeled a data point. Green labels are correctly classified, red labels are misclassified. This table shows predicted labels for the classifiers trained with optimal parameter settings.