Computational Linguistics in the Netherlands Journal 13 (2024) 7-29 Submitted 02/2024; Published 03/2024

Personality Style Recognition via Machine Learning: Identifying Anaclitic and Introjective Personality Styles from Patients' Speech

Semere Kiros Bitew* SEMEREKIROS.BITEW@UGENT.BE
Vincent Schelstraete*** VINCENT@TECHWOLF.BE
Klim Zaporojets* KLIM.ZAPOROJETS@UGENT.BE
Kimberly Van Nieuwenhove** KIMBERLY.VANNIEUWENHOVE@UGENT.BE
Reitske Meganck** REITSKE.MEGANCK@UGENT.BE
Chris Develder* CHRIS.DEVELDER@UGENT.BE

* *IDLab, Department of Information Technology, Ghent University – IMEC, Ghent, Belgium*

** *Department of Psycho-analysis and Clinical Consulting, Ghent University, Ghent, Belgium*

*** *TechWolf, Ghent, Belgium*

Abstract

In disentangling the heterogeneity observed in psychopathology, personality of the patients is considered crucial. While it has been demonstrated that personality traits are reflected in the language used by a patient, we hypothesize that this enables automatic inference of the personality type directly from speech utterances, potentially more accurately than through a traditional questionnaire-based approach explicitly designed for personality classification. To validate this hypothesis, we adopt natural language processing (NLP) and standard machine learning tools for classification. We test this on a dataset of recorded clinical diagnostic interviews (CDI) on a sample of 79 patients diagnosed with major depressive disorder (MDD) – a condition for which differentiated treatment based on personality styles has been advocated – and classified into anaclitic and introjective personality styles. We start by analyzing the interviews to see which linguistic features are associated with each style, in order to gain a better understanding of the styles. Then, we develop automatic classifiers based on (a) standardized questionnaire responses; (b) basic text features, i.e., TF-IDF scores of words and word sequences; (c) more advanced text features, using LIWC (linguistic inquiry and word count) and context-aware features using BERT (bidirectional encoder representations from transformers); (d) audio features. We find that automated classification with language-derived features (i.e., based on LIWC) significantly outperforms questionnaire-based classification models. Furthermore, the best performance is achieved by combining LIWC with the questionnaire features. This suggests that more work should be put into developing linguistically based automated techniques for characterizing personality, however questionnaires still to some extent complement such methods.[1]

1. Introduction

Personality is considered a crucial variable in explaining the heterogeneity observed in psychopathology (Blatt 2004). It has been demonstrated that the large group of people meeting the commonly used criteria for depression (e.g., DSM or ICD criteria) − which affects over 280 million people worldwide (according to the WHO's 2021 report),[2] − is not homogeneous. On the contrary, the heterogeneity of people covered by the general label of depression is widely recognized (Goldberg 2011).

From multiple perspectives, a two-polarity model has been put forward to distinguish between depressed patient groups on the basis of two personality dimensions, namely relatedness and self-definition (Blatt 1974, Beck 1983, Blatt 2004). Drawing on Freud's (Freud 1953) early distinction

1. Code will be available in `https://github.com/semerekiros/personapredict`
2. `https://www.who.int/news-room/fact-sheets/detail/depression`

between hysteria (libidinal, oral fixation) and obsessional neurosis (aggressive, anal fixation), Blatt (1974, 2004) distinguishes between an anaclitic and introjective personality style, respectively. The *anaclitic* personality style is focused on interpersonal relationships and dependency, whereas the *introjective* personality style is focused on self-definition, autonomy, and interpersonal distance. Beck (1983) proposed similar interpersonal characteristics, sociotropy and autonomy, to discern different personality styles in depression. Sociotropy indicates an intense need for intimate relationships, whereas autonomy is associated with a desire for independence and achievement. As such, sociotropy and autonomy correspond with the anaclitic and introjective personality dimensions, respectively. Similar to Blatt (1974), Beck (1983) assumes that individuals are more prone to react to certain stressors, dependent on their personality style. Anaclitic/sociotropic individuals are vulnerable for interpersonal stressors (e.g., breakups), whereas introjective/autonomous individuals are more sensitive to stressors on the level of achievements and self-definition (e.g., job loss). Consequently, a predominance of either personality dimension leads to the development of distinct symptoms when confronted with these specific stressors: different phenotypical manifestations of depression have been described from this perspective. The symptom-specificity hypothesis (Blatt 2004) assumes that anaclitic depression, typified by an overdetermination of dependency and a preoccupation with interpersonal consolidation, is associated with somatic symptoms and phobias. An introjective depression on the other hand, is dominantly focused on self-definition and interpersonal distance and involves symptoms on a more cognitive level, exemplified by self-criticism, perfectionism and pathological doubt. Noting these different manifestations, it has been argued that also treatment should be tailored according to patients' personality styles, with anaclitic patients benefiting more from structured, supportive therapy approaches, whilst patients with an introjective personality style would profit more from insight-focused treatments (Werbart et al. 2017). Thus, a first fundamental step supporting this would be effective determination of a patient's dominant personality style.

Strikingly, notwithstanding their solid theoretical underpinnings, after almost half a century, personality style assessment remains a challenge, which is embodied by research failing to yield consistent support for the symptom-specificity hypothesis (Coyne et al. 2004, Desmet 2007) and outcome studies, as of yet, not having convincingly demonstrated the benefits of matching personality styles to different treatment strategies (Meganck et al. 2017, Werbart et al. 2018). At the heart of this predicament probably lie the methodological difficulties associated with the assessment of personality styles (Desmet 2007). The determination of personality styles typically includes the administration of self-report questionnaires, such as the Depressive Experience Questionnaire (DEQ) (Blatt et al. 1976) and the Personal Style Inventory (PSI) (Robins and Luten 1991). However, research literature is inconsistent as to whether these questionnaires are apt (valid) to distinguish between the clinical manifestations (symptoms, interpersonal functioning) associated with an anaclitic and introjective personality style (Coyne et al. 2004, Desmet 2007). To make further progress, there is a need to explore alternative approaches to detect personality styles with greater precision. More recent approaches such as prototype matching (Werbart and Forsström 2014) are promising, since they combine scientific rigor with clinical expertise. Yet, they are time consuming because they rely on the expertise of several well-trained human raters. Natural Language Processing (NLP) and Machine Learning (ML) techniques, originating from the field of computer science, might alternatively provide ways to assess psychological features and phenotypes based on complex data without the interference of human raters (Dwyer et al. 2018). NLP focuses on automated processing of human language, learning and applying the underlying linguistics and semantics in computerized systems (Joseph et al. 2016). Today, NLP applications heavily rely on machine learning techniques. Machine learning (ML) is a subset of artificial intelligence that "automatically determines (i.e., learns) methods and parameters to reach an optimal solution to a problem rather than being programmed by a human a priori to deliver a fixed solution" (Dwyer et al. 2018). In other words, ML offers techniques to identify patterns, phenotypes or subgroups in a dataset (e.g., text, audio, self-report questionnaires, biological data) through learning from the data with minimal prior assumptions. There is over two decades' worth of research in using ML techniques for the psychological field, concerning the diagnosis, prognosis and treatment of diverse mental conditions (Dwyer et al. 2018), such as depression (Fu et al. 2019) and suicide contemplation (Bitew et al. 2019). Unfortunately, the use of mostly complex biological data (e.g., neuroimaging and MRI data) complicates the translation of ML solutions to practical and clinical implementations (Dwyer et al. 2018). In the context of psychotherapy research, however, what patients say about themselves and their symptoms is an essential, if not the most important source of information, which makes patients' narratives a particularly suitable source of data for NLP and ML. For the detection of depression in general, there are several studies using NLP on textual data (Al Hanai et al. 2018, Coppersmith et al. 2015, Özkanca et al. 2018, Tausczik and Pennebaker 2010). As an example, the Linguistic Inquiry and Word Count (LIWC), a text analysis program that counts the frequency of words and types of grammar typically used in certain contexts, Tausczik and Pennebaker (2010), found that depressed people use more negative words and terms related to death on social media, while also using the first person more frequently than healthy controls. Next to the detection of depression using textual information from social media (Coppersmith et al. 2015, Tausczik and Pennebaker 2010), depression has been successfully inferred using transcripts from screening interviews (Al Hanai et al. 2018) and therapy sessions (Özkanca et al. 2018). However, as Dwyer et al. (2018) note, it does not suffice to use ML to detect broad categories (e.g., depression yes/no), as treatment for depression in general shows relatively low efficacy (Driessen et al. 2015). Especially for the heterogeneous depressed population, it would be fruitful to use ML techniques to predict more specific features, such as personality style, which would allow more tailored clinical decision making.

Thus, several works were proposed that employ ML for automatic personality prediction using different personality frameworks (measures) such as Big Five (Digman 1990), MBTI (Myers and McCaulley 1988), and Catell's 16PF (Cattell and Mead 2008). The majority of research makes use of the Big Five personality measure for personality trait classification (Majumder et al. 2017, Schwartz et al. 2013, Mairesse et al. 2007, Mairesse and Walker 2006, Stachl et al. 2020) followed by MBTI (Yang et al. 2021, Gjurković and Šnajder 2018). The Big Five personality measure defines personality through the following five dimensions: extroversion, agreeableness, conscientiousness, neuroticism, and openness. Similar to depression detection systems, automatic personality detection ML models use a wide range of data sources; these sources include digital footprints in social media platforms (e.g., Reddit (Gjurković and Šnajder 2018), Twitter (Kalghatgi et al. 2015), Facebook (Hall and Caton 2017, Golbeck et al. 2011)), essays written in a controlled environment by volunteers (Tausczik and Pennebaker 2010, Pennebaker and King 1999), video and audio recordings of meetings (Carletta et al. 2005). The input modalities that these ML models use include text, audio, visual cues, and multimodal inputs. Many personality detection ML models that use *textual data* as input extract features using two approaches: closed-vocabulary and open-vocabulary. Closed-vocabulary approaches rely on a priori determined word categories, each represented by a fixed set of words (where some may belong to multiple categories). For example, LIWC, which is the most common closed-vocabulary method for personality detection from text, categorizes the words in a text into various psychologically relevant clusters like 'affective processes' (e.g., 'happy', 'cried', 'nervous') and 'social processes' (e.g., 'pal', 'buddy', 'cousin', 'woman'). The frequency counts of words for each cluster are used by ML models to predict the personality of the text author (Hall and Caton 2017). On the other hand, open-vocabulary methods rely on extracting a comprehensive collection of language features from textual input such as n-gram features, part of speech (POS) tags, usage of emoticons, length of posts of users from social media, and use of word embeddings (e.g., Word2vec, GloVe) to build ML models for personality detection (Schwartz et al. 2013, Plank and Hovy 2015). Apart from the actual words spoken, additional features can be extracted from the *audio signal*, e.g., loudness, spectrals, pitch, interruptions, acoustic features, etc. For instance, as the first stage in their deception detection pipeline, (Levitan et al. 2016) use prosodic, acoustic, and LIWC features to predict the personality traits of speakers engaging in a dialogue. They analyzed how each LIWC element contributed and discovered, among others, that the dimensions of "focus-future" and "drives" are the most helpful features for determining "extroversion" personality traits, whereas "time" and "work" are significant for determining "conscientiousness." They also show that incorporating the personality traits as an input improves their deception detection model. More recently, deep learning approaches (i.e., mainly based on the transformer model (Vaswani et al. 2017)) have been employed for personality prediction (Yang et al. 2021, Jain et al. 2022, Zhu et al. 2022).

Visual-based ML models for personality prediction mainly extract features of the body, especially the face. Facial features such the shape of nose and eyebrows, eye openness, and mouth have been used to infer personalities of users (Kamenskaya and Kukharev 2008). For example, Liu et al. (2016) trained an ML model to predict personality of people by analyzing their Twitter profile pictures and found users high in 'openness' prefer more aesthetic pictures, while users with strong 'agreeableness' and 'conscientiousness' display more positive emotions in their photos. Similarly, Cristani et al. (2013) studied the relationship between personality of users and the types of pictures they prefer (e.g., by examining the types of pictures they like). They found that pictures posted as "favourite" by users in Flickr[3] were correlated with their personality traits. For example, they found that extrovert individuals show a preference of pictures portraying people while introvert show the opposite preference.

Other works combine one or more of the above discussed modalities to predict personality traits (Kampman et al. 2018). For example, Kindroglu et al. (2017) combine features from audio and visual modalities to predict extraversion and leadership traits. As of yet, to the best of our knowledge, ML techniques have not been used to learn typical linguistic properties that assess personality styles at a specific level of behaviors (i.e., anaclitic vs. introjective) for clinical psychology purposes.

Therefore, the aim of our study is twofold, namely to (a) better understand anaclitic versus introjective primary personality traits through the **analysis** of natural language; and (b) investigate possible **ML solutions** for automatic personality prediction.

First, we explore important linguistic characteristics of anaclitic versus introjective patients by performing feature analysis on data from intake interviews (transcripts and audio) of depressed patients. Notwithstanding the explorative nature of feature analysis, there are certain expectations that can be drawn from a recent line of preliminary descriptive research into the different linguistic styles of anaclitic and introjective patients (Dávalos et al. 2017, Valdés and Krause 2015). Since anaclitic patients are highly concerned with the development of meaningful and satisfying interpersonal relationships (e.g., expressed through LIWC categories as 'social', 'family', 'affiliation'), we expect them to use more pronouns in the second and third person such as 'you', 'he', 'she', etc. Conversely, we expect introjective patients to frequently use the word 'I', since they are rather preoccupied with achieving a differentiated and consolidated identity (e.g., LIWC category 'achieve'). Furthermore, we expect anaclitic patients to use more emotional words (e.g., LIWC categories 'affect' and 'feel') and speak with an affective invested intonation, whereas more cognitive and causal words (e.g., LIWC categories 'CogProc', 'Insight', 'Cause') are expected in introjective patients' speech. In a nutshell, the first objective is to show how analyzing verbal behaviors coalesces into an interpretable and meaningful distinction in personality.

Second, we investigate whether personality classification into anaclitic or introjective can be automated via ML techniques, using basic and more advanced classifiers with diverse data as input (text, audio, questionnaires, and different combinations of these). The aim hereof is to (a) assess whether ML allows automatic personality style prediction as good (or better) than via the traditionally used self-report questionnaires that are crafted by experts, and (b) which data and selected ML method attains the better personality classification performance.

The remainder of the paper is organized as follows: Section 2 describes how the clinical diagnostic Interview (CDI) was conducted, as well as the data pre-processing, the feature extraction and the machine learning models we used. Section 3 presents the analytical findings followed by the predictive experimental results. Section 4 summarizes our conclusion and discussion.

3. `https://www.flickr.com/`

2. Materials and Methods

2.1 Participants

The dataset comprises the intake interviews of 79 Dutch-speaking patients enrolled in the Ghent Psychotherapy Study (GPS), a randomized controlled trial studying the differential efficacy of cognitive behavioral therapy and short-term psychodynamic psychotherapy for depressed patients with a predominant anaclitic or introjective personality style (Meganck et al. 2017). All patients are diagnosed with Major Depressive Disorder (MDD), assessed by the Hamilton Rating Scale for Depression (Hamilton 1967) and Structured Clinical Interview for DSM-IV-TR (First et al. 2002). Exclusion criteria were primary diagnosis of substance abuse/dependence, acute psychotic symptoms or suicidal ideation. The sample consists of 50 anaclitic patients (12 males (24%), 38 females (76%); age range 20-60 years (mean = 38.3, standard deviation = 12.3)) and 29 introjective patients (12 males (41.4%), 17 females (58.6%); age range 21-59 years (mean = 37.6, standard deviation = 11.0)).

2.2 Procedures and Measures

Before enrollment in the study, patients filled in a large test battery at home consisting of the *Depressive Experience Questionnaire* (DEQ) (Blatt et al. 1976) and the *Personal Style Inventory* (PSI) (Robins and Luten 1991) to assess personality characteristics; and the *Beck Depression Inventory-II (BDI-II)* (Beck et al. 1996), the *Depression Anxiety Stress Scale* (DASS) (Lovibond and Lovibond 1996), the *Symptom Checklist-90* (SCL-90) (Derogatis 1992), the *Inventory of Interpersonal Problems-32* (IIP-32) (Horowitz 2002), the *Self-rating Inventory for Posttraumatic Stress Disorder* (SIL) (Hovens et al. 2000), the *Outcome Questionnaire-45* (OQ-45) (Lambert et al. 2004), *Experiences in Close Relationships* (ECR) (Fraley et al. 2000) and ten emotions on a *Visual Analogue Scale* (VAS) (van Rijsbergen et al. 2014) to assess symptoms and complaints related to depression, anxiety, trauma, interpersonal problems, overall well-being and emotional experiences.

The Clinical Diagnostic Interview (CDI) (Westen and Muderrisoglu 2006) was administered and audio-taped by a member of the GPS research team upon the intake of the patient. The CDI is a semi-structured interview that questions both current clinical complaints and meaningful lifetime events. The interview guide was built using open questions to evoke rich narratives concerning inter- and intra-personal experiences. The raw audio files of these intake interviews were transcribed using preset standards, and the resulting textual transcripts saved as Word documents.

Personality style was assessed through an intensive prototype matching procedure (Werbart and Forsström 2014) conducted by three independent and trained GPS researchers (including the researcher who conducted the interview). The procedure consists of each researcher rating the CDI individually on a scale of 1 to 5 for both the anaclitic as well as the introjective personality style using prototype vignettes (Werbart and Forsström 2014). In a second step, these independent ratings were discussed to reach consensus. To assign a personality style to a participant, a score of at least 3 for one of the personality styles and a minimum of 2 points difference with the score on the other personality style were required. When no agreement could be reached, the difference in consensus scores between dependent and self-critical personality dimensions was less than 2 points, or there was no score of at least 3 on either dimension, patients were excluded from the trial.

2.3 Data Preprocessing

Our data analysis is based on features obtained from the CDI intake interviews, by processing both audio files and human produced transcripts thereof, as sketched in Fig. 1. The transcripts (i.e., Word documents) of the CDI were segmented into questions, answers, and background noises according to the text formatting. Anonymization was performed by first extracting capitalized words, thus obtaining a list that was manually filtered to retain a list of unique person and location names to anonymize (i.e., replace them with a special <*name*> or <*location*> tag). Using the *Natural*

Language Toolkit (NLTK) (Perkins 2010), the text was then further tokenized and standardized (e.g., lowercasing, mapping dialect words and numbers to standard language, applying tags for sounds and anonymizations) and eventually appended into a JSON-file. The audio (i.e., recorded in MP3-format) of the CDI, was converted to a mono-recording and stored as a WAV-file with a normalized volume, a sample rate of 16 kHz and a bit-depth of 16 bits per sample. Speaker segmentation and speaker clustering (Desplanques et al. 2017) was used to separate speech into interviewer, interviewee utterances and non-speech, using the Open Speech and Music Interpretation by Large Space Extraction (openSMILE).[4] openSMILE is a modular and flexible toolkit used for feature extraction from audio files. The resulting segments were stored as separate TextGrid files.

2.4 Feature extraction and Machine Learning Models

In this section, we discuss how the features extracted from the CDI intake interviews are further processed before they can be fed to machine learning models. Additionally, we discuss which machine learning techniques were used in order to create models that classify the personality styles of patients. Figure 1 shows the complete feature extraction pipeline. We built our machine learning models based on the feature categories we will discuss next.

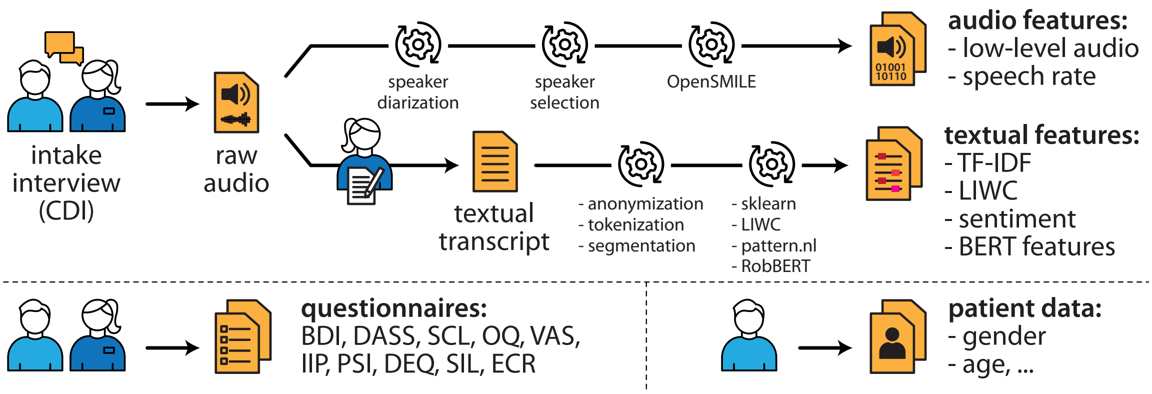


Figure 1: The input data and the feature extraction pipeline. The features extracted from the CDI intake interviews include: TF-IDF, LIWC, BERT features, and audio features. The questionnaire features are filled by the patients.

2.4.1 Questionnaire features

As briefly introduced in Section 2.2, patients filled a large battery of questionnaires before they were enrolled in the study. The joint outputs of all collected questionnaires yielded a feature vector with a total of 358 elements. This feature vector contains both categorical and numerical values. It is then fed to several classifiers for determining the personality trait of the the patients. For our questionnaire feature set, we trained a simple logistic regression (LR), a random forest (RF), and a CatBoost (CB) classifier. We use the LR as our baseline because it is one of the most basic and popular algorithms to solve binary classification problems. The second ML model, random forest (RF) is an ensemble of decision tree classifiers, in which it is ensured that each tree is an acceptably good classifier on its own and that the trees are uncorrelated. Finally, we use CatBoost as it is specifically designed for dealing with categorical variables and has proven to attain state-of-the-art results for classification tasks based on tabular data.

4. `https://www.audeering.com/research/opensmile/`

2.4.2 TF-IDF features

To process the textual transcripts, for each word in a document we calculated the TF-IDF scores as features (Ramos et al. 2003). An inverse document frequency (IDF) weight indicates the relevance of a certain word within a certain document, with lower weights being associated with words that are common in the corpus and thus have small discriminating power. This word IDF is then multiplied by the term frequency (TF) in a particular document to obtain the TF-IDF score of that word for the respective document. Calculating those TF-IDF scores for each word, we thus summarize a document as a vector of these per-word scores, indicating which words are important in the document and which are not. This procedure was applied to the interview transcriptions, converting each document to a high-dimensional sparse real-valued vector, with a value for each different word encountered. The same TF-IDF scores were also calculated for N-grams. An N-gram is a sequence of N consecutive words appearing in a document, thus allowing longer grammatical constructions to be captured as well. In sum, the set of features for the TF-IDF based classifiers consists of a vector for each document containing the TF-IDF score for that document of each N-gram (N = 1, 2, 3) in the corpus (i.e., 840,834 elements).

We used logistic regression (LR), random forest (RF) and CatBoost (CB) to classify the patients. Rather than considering all transcribed speech of the patient in the complete interview as a whole, we also experimented by considering "chunking" their utterances in the individual answers, treating each per-answer "chunk" as a separate document sample. We then applied majority voting on the per-chunk classification to predict the personality of a given patient.

2.4.3 Psychological features

Next, we turned to text processing techniques specifically tailored to an application with features known to be indicative of psychological characteristics and/or sentiment. First, we employed the 2015 Dutch version of the *Linguistic Inquiry Word Count-Lexicon* (LIWC), a computerized text analysis tool for counting words in 73 psychology-relevant word categories, established through psychometric studies on linguistics (Tausczik and Pennebaker 2010). The output of the LIWC tool is a vector that contains the absolute number of times a word from each category was used. We normalized these absolute count values to fractions of all uttered words in patient's answers. Mann-Whitney-Wilcoxon tests were performed to establish which categories are relevant discriminators between the two personality styles (Delacre et al. 2017) where the significance of each category was expressed as a p-value. The results will be discussed in Section 3.1.

Second, we used the sentiment analysis functionality from the *pattern.nl* library[5]. The sentiment-function takes a sentence or paragraph as input and returns a tuple of two values, subjectivity and polarity. The subjectivity has a value between 0.0 and 1.0, with low values when the text is objective (e.g., "The movie was about squirrels.") and higher values indicating subjective statements (e.g., "The movie was very good!"). The polarity indicates whether the sentiment expressed in a sentence (e.g., "I absolutely hated the movie", "it was quite good") is negative or positive, with values ranging between -1.0 and $+1.0$. The greater the absolute value, the stronger the sentiment. To extract sentiment features from transcripts, we first compute the polarity and subjectivity of each sentence in the transcript, then use the mean and standard deviation for both dimensions as the feature set. Again, Mann-Whitney-Wilcoxon tests were performed to examine whether the differences in terms of subjectivity and polarity between the two personality styles were significant and significance was expressed as a p-value. See Section 3.1 for details.

2.4.4 BERT features

Another feature set we extracted makes use of advanced deep neural models called transformers (Vaswani et al. 2017), which represent words depending on their semantic role in the context of the

5. `https://digiasset.org/html/pattern-nl.html`

text. Bidirectional Encoder Representations from transformers (BERT) (Devlin et al. 2018) is the most popular pre-trained language model that has been widely used to extract rich representations from textual data and proved to attain state-of-the-art results in several NLP tasks such as sentiment analysis, question answering, machine translation, etc. Moreover, several downstream tasks such as depression detection (Rodrigues Makiuchi et al. 2019), suicide risk assessment (Matero et al. 2019), and personality detection (Kazameini et al. 2020) have benefited from such context-aware representations. Since our data is in Dutch, we use RobBERT (Delobelle et al. 2020), a variant of BERT that is pretrained on Dutch language, which has achieved state-of-the-art performance for a wide range of tasks in Dutch. Typically, a BERT based model gets an input text (i.e., the textual transcripts) prepended by a special token called *CLS*, and produces output representations (i.e., a vector of 768 elements) for each of the words in the input and the CLS token. The CLS token is pretrained to represent the entire input sequence. Thus, the vector representation of CLS is used as the extracted feature for the input transcript. Another method is to combine the word representations excluding the CLS token, for example, taking the average or the maximum of all vector representations. In this paper, we experiment with both CLS representation and max-pooled representation.

Since the input to BERT models is limited to only 512 tokens, we segment our textual transcriptions into chunks of 512 tokens. For determining the class label of a transcript, we look at the predicted labels of its chunks and decide the final class label based on majority voting over the individual of the chunks labels. For this feature set, we built a simple logistic regression model.

2.4.5 Audio features

Since in depression studies, noticeable differences in speech production have been suggested as potential biomarkers for depression (Mundt et al. 2012), we also collected such features to investigate whether they are useful to detect personality differences. As it has also been established that there are language differences between males and females (Newman et al. 2008), we will also consider gender as a confounding variable. From the audio, features were extracted from the patients' speech using openSMILE, a modular and flexible toolkit used for feature extraction from audio files (Huang et al. 2018). We chose to use the extended version of the *Geneva Minimalistic Acoustic Parameter Set* (GeMAPS) as configuration, as it obsoletes feature selection (Eyben et al. 2015). This resulted in a feature vector with 88 low-level audio descriptors for each segment in which the patient was speaking. These descriptors include frequency-, energy- and amplitude-related parameters, as well as spectral and temporal parameters. These parameters were normalized to have zero mean and unit variance across all segments belonging to one patient. We also calculated the speech rate for the patient, using two different metrics. The first is the so-called *speaking rate*, calculated as the number of words spoken per second, where the time counted includes the hesitations between the words where no words are pronounced. The second is the *articulation rate*, expressed in characters per second, where silences or hesitations where no speech is produced are excluded from the timekeeping. Finally, we also calculated the fraction of time a patient hesitates while speaking and recorded it as the *hesitation fraction.* The audio feature set thus contains 91 features in total (i.e., 88 low-level audio features and 3 speech rate-related features). Again for this feature set, we experimented with logistic regression, random forest and CatBoost machine learning models.

3. Results

In this section, we present and discuss the analytical and experimental findings. Section 3.1 provides a thorough analysis of both textual and audio modalities to distinguish anaclitic from introjective personality traits. Section 3.2 reports the results of the predictive ML models that were built using the features introduced in Section 2.4.

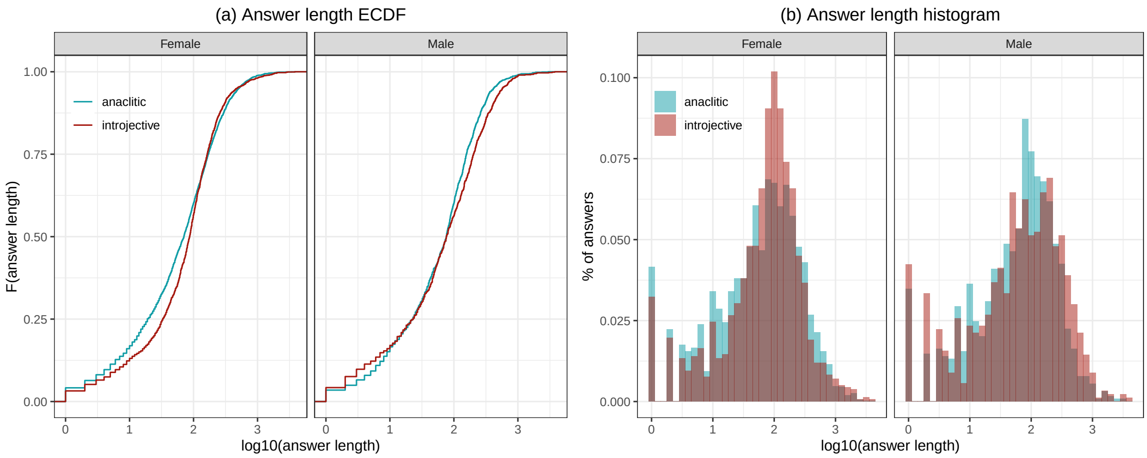


Figure 2: Distribution of the answer length (in number of tokens without punctuation) of the patients' aggregated answers in the CDI transcripts: (a) empirical cumulative distribution function (ECDF), (b) histograms

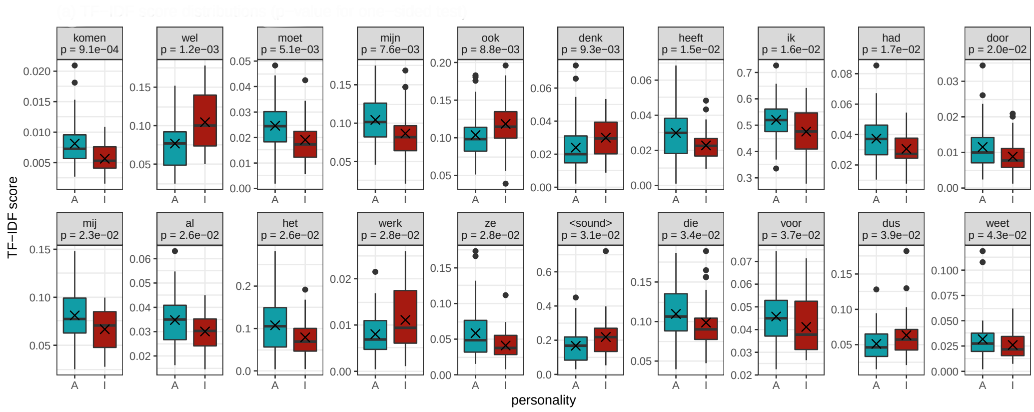
3.1 Analysis

3.1.1 Text Analysis

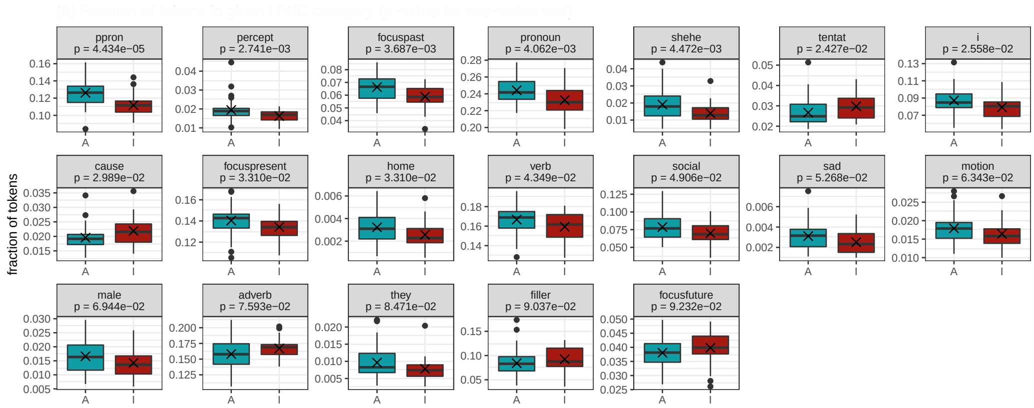
- **Answer Length** The average transcript length of the CDIs is 15,226 words (disregarding punctuation), with a standard deviation of 5,208 words. The patient's answers constitute the bulk thereof, i.e., 13,556 words (standard deviation of 5,131 words). This standard deviation is fairly large, which can be explained by the rather limited size of the dataset. The distribution of the document length is shown in Fig. 2 for the complete transcription set and for the different personality styles separately. This figure shows that the length of answers does not seem to significantly differ between personality styles or sexes.

- **TF-IDF Analysis**: Before considering specific word categories in LIWC, we analyzed the pure vocabulary usage by means of individual words' TF-IDF scores. In Fig. 3a, we plot the distributions of TF-IDF scores in the anaclitic and introjective patient populations for words that (i) occur in all transcripts; and (ii) occur significantly in one of the personality categories, according to a Mann-Whitney-Wilcoxon test at $p < 0.05$. Even though these features consider documents as bags of words, and thus ignore word order or the context in which they occur, we still observe some meaningful insights. We note the higher frequency of personal pronouns — which are notable features in distinguishing personality styles (Pennebaker 2011) — in anaclitics (see "ik (I)", "mij (me)", "mijn (my)", "ze (she/they)"). For introjectives, we further note more prevalent "denk (think)", "wel (well)", "ook (also)" which one would expect to occur more in statements reflecting cognitive processes, nuanced reasonings, etc. and thus is in accordance with our a priori expectations. We also remark that introjectives utter more non-word sounds (the "<sound>" marker), vocalized as "uh", "pff" and the like, which could be seen as hesitations that reflect an ongoing cognitive process while talking.

- **LIWC Analysis**: Because of the variation in interview length (see earlier), instead of absolute word counts, we opted for relative usage of LIWC word categories (i.e., the fraction of words used belonging to a given LIWC category). Analysis of these statistics is summarized in Table 1, which reports the p-values of a Mann-Whitney-Wilcoxon test with null hypothesis that the feature distribution in anaclitic and introjective patient populations is the same. Note that we only consider word categories that are used at least once by all patients, and at least 10 times



(a) TF-IDF score distributions



(b) Fraction of tokens in given LIWC category

Figure 3: Lexical features that are potentially useful for distinguishing anaclitic (A) versus inrojective (I) personality styles: (a) TF-IDF score distribution for the most discriminating words, (b) LIWC category frequency (measured as fraction of text tokens). The stated p-values result from a Mann-Whitney-Wilcoxon with null hypothesis that the feature distributions are the same for anaclitics and introjectives. Note: $\times$ represents the mean value.

per interview on average over all patients. We find that personal pronouns ('ppron', 'shehe', 'I', 'they') as well as pronouns ('pronoun') in general, comprise a larger fraction of uttered words for anaclitics than for introjectives. Further, also the categories 'home' and 'sad' are significantly more prominent in anaclitic than introjective patients (at $p < 0.05$). In line with our expectations, words belonging to the cognitive process sub-categories 'cause' (causation, e.g., 'because', 'effect') and 'tentat' (tentative, e.g., 'maybe', 'perhaps') are significantly (at $p < 0.05$) more common in introjective patients' speech compared to anaclitic patients and, vice versa, words belonging to the category 'social' are more frequently used by anaclitic patients. Another interesting finding is that anaclitic patients seem to be preoccupied with the present and past (categories 'focuspresent' and 'focuspast'), while introjective patients seem more invested in the future (category 'focusfuture', only marginally significant at $p < 0.10$). Figure 3b shows the distributions of these LIWC category usage frequencies in the form of box plots. This confirms that mainly (personal) pronouns, as well as 'sad' and 'focuspast' are discriminative word categories to distinguish anaclitics from introjectives.

Table 1: LIWC categories whose relative frequency among the words used by introjective versus anaclitic patients is significantly different, at least at $p < 0.10$.

Rank	Category	p-value	Dominant personality style
1	ppron	4.43E–05	anaclitic
2	percept	2.74E–03	anaclitic
3	focuspast	3.69E–03	anaclitic
4	pronoun	4.06E–03	anaclitic
5	hear	4.19E–03	anaclitic
6	shehe	4.47E–03	anaclitic
7	tentat	2.43E–02	introjective
8	i	2.56E–02	anaclitic
9	cause	2.99E–02	introjective
10	focuspresent	3.31E–02	anaclitic
11	home	3.31E–02	anaclitic
12	body	4.24E–02	anaclitic
13	verb	4.35E–02	anaclitic
14	social	4.91E–02	anaclitic
15	sad	5.27E–02	anaclitic
16	motion	6.34E–02	anaclitic
17	male	6.94E–02	anaclitic
18	adverb	7.59E–02	introjective
19	they	8.47E–02	anaclitic
20	filler	9.04E–02	introjective
21	focusfuture	9.23E–02	introjective
22	death	9.63E–02	introjective

- **Sentiment Analysis**: we analyzed the sentiment of the answers given by the patients during their interview with therapists. First, we discarded short answers (i.e., containing <10 sentences because the sentiment of a patient is more confidently inferred from longer replies). Then, as introduced in Section 2.4.3, we calculated the polarity and subjectivity of patients' answers by taking the average of the polarity and subjectivity of all individual answers. Studying the distributions of the patients' answer polarity and subjectivity did not reveal any statistically significant conclusion.

  Because the answer sentiment analysis did not yield conclusive findings, we opted to use an alternative approach by comparing question-answers pairs (Özkanca et al. 2018). From these pairs, we wanted to investigate two dimensions: (1) whether therapists are more likely to ask positive or negative questions to patients with anaclitic or introjective personality styles, and (2) whether anaclitic and introjective patients respond more positively or negatively to these questions.

  To examine the first dimension, we look at the polarity scores of the questions and answers. Since polarity scores range between -1 and 1, we divided them into three equal parts to reflect positive, neutral and negative polarity. A question/answer was regarded positive if the average polarity over its sentences was larger than 0.3, negative if it was less than -0.3, and neutral otherwise. Figure 4 suggests that interviewers/therapists are not biased towards a certain personality style in terms of differentiating the proportions of positive and negative questions.

Next, to understand the second dimension, we studied how patients' responses were affected by the polarity of the questions posed to them by the interviewers. To achieve this, we first used the

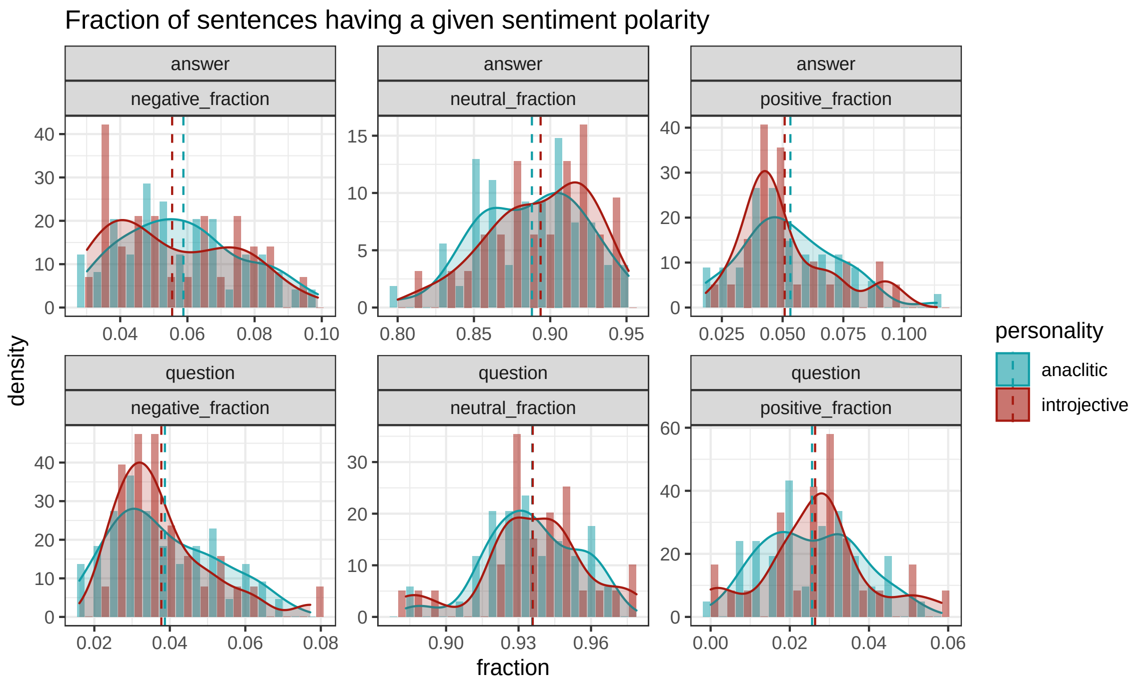


Figure 4: Answer and question polarity distributions. Polarity is determined positive for values > 0.3, neutral in the range [-0.3, +0.3], and negative if < -0.3

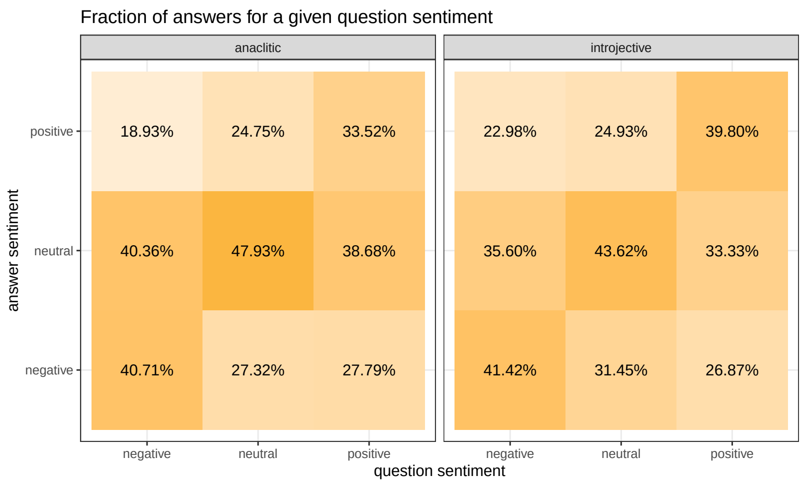


Figure 5: Question-answer pair sentiment polarity combinations: per question sentiment, we calculate the fraction of corresponding answers for each of the positive/neutral/negative sentiment categories. The sentiment category of a question/answer is assigned according to the most polarized sentence therein

most polarizing sentence in the question/answer as our approximation of its sentiment (negative for an extreme polarity < -0.3 , positive for > 0.3, else neutral). Then, for a given question, we calculated the the fraction of answers that fall in each polarity category for both personality styles as depicted in the heat map in Fig. 5. Anaclitic patients seem to stick more to giving neutral answers, while introjectives seem more inclined to follow the question's polarity.

3.1.2 Audio Analysis

We first analyzed the potential difference in audio feature measurements between anaclitic and introjective patients. For the 88 raw audio features, we calculated ANOVA F-scores to determine their respective importance. A high F-score indicates that a feature explains a higher fraction of the total variance and thus is more informative. Using this approach, the 10 most relevant features are mean F3-frequency, mean F2-frequency, mean alpha ratio, 20^{th} percentile loudness, mean harmonic difference H1-A3, standard deviation local shimmer (dB), mean local shimmer (dB), mean pitch, mean Hammarberg index (voiced) and mean loudness, respectively. (Memon 2020) describes the acoustic meanings of these features. For example, the mean F2 and F3 frequencies indicate a resonant voice while the mean harmonic difference H1-A4 correlates to a creaky voice. However, we did not find indicative features that are meaningfully associated with one of the personality styles, i.e., whether any of the features has different group means between introjective versus anaclitic patients.

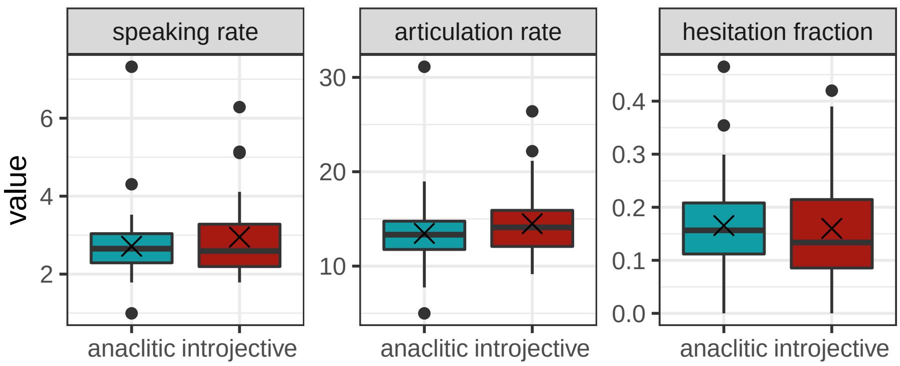


Figure 6: High-level audio feature measure distribution in anaclitic and introjective patient populations. Note. $\times$ represents the mean value

Additionally, for the more high-level features related to speech rate and hesitations, (a) speaking rate, (b) articulation rate, and (c) hesitation fraction, we performed Mann-Whitney-Wilcoxon tests. Yet, again we found no statistically significant differences in those measures between anaclitic and introjective patient populations, as can also be observed from the largely overlapping box-plots in Fig. 6.

3.2 Machine Learning for Automatic Personality Classification

In this section, we provide the results of the several classification ML models we built to predict personality traits of patients. We use F1-score and Cohen's kappa as evaluation metrics to measure the performance of our models. F1-score can be defined in terms of true/false positives, and true/false negatives. For a given class (i.e., personality style), a true positive (TP) is a positively classified sample (i.e., having the given personality style) that was actually positive and a true negative example (TN) can be defined similarly. A false positive (FP) is a positively classified patient that in reality is negative for the given class, while a false negative is the counterpart of this. As such, the precision (P) of a classifier is the fraction of positive classifications that was actually positive, as shown in equation Eq. (1). The recall (R) of a classifier is the fraction of positive training examples that is actually identified as being such (see equation Eq. (2)). Finally, the F1-score is calculated as the geometric mean of precision P and recall R as shown in equation Eq. (3).

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{3}$$

On the other hand, *Kappa* is a measure originally designed as a measure of agreement between two judges, based on the accuracy but corrected for chance agreement. In a classification setting, it is used to evaluate the agreement between the actual and the predicted classes by the model. We use it because it implicitly accounts for class imbalance. Cohen's Kappa is calculated with the following formula:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{4}$$

where P_o is the overall accuracy of the model (i.e., calculated as $(TP+TN)/(TP+TN+FN+FP)$) and P_e is the expected agreement between the actual classes & predicted classes. For a binary task like ours, P_e is the sum of P_{e1}, the probability of the predictions agreeing with actual values of class 1 (i.e., anaclitic) by chance, and P_{e2}, the probability of the predictions agreeing with the actual values of class 2 (i.e., introjective) by chance. Assuming that the actual class values and model predictions are independent, these probabilities (i.e., P_{e1} and P_{e2}) are calculated by multiplying the proportion of the actual class and the proportion of the predicted class. P_e is calculated as follows:

$$P_e = P_{e1} + P_{e2} = P_{e1,actual} \cdot P_{e1,predicted} + P_{e2,actual} \cdot P_{e2,predicted} \tag{5}$$

To evaluate our proposed classifiers' performance, we used stratified 5-fold cross-validation, where the dataset of 79 patients was split into 5 data segments that (approximately) have the same ratio of anaclitic and introjective patients (63% resp. 37%). We used the same splits for each individual classifier, training a classifier model using 4 data segments as training data and evaluating it on the remaining unseen data segment. To report our results, we calculate each metric as an average over the 5-fold cross-validation runs. We repeat this split into 5 folds randomly for 100 times, which results in 100 scores for each metric. Our final results are averages over these iterations and the corresponding standard deviation, as shown in Table 2. It is important to note that the comparison between different models was done using a parallel analysis over the 100 runs. We do a simple t-test between two sets of 100 scores (i.e., coming from two models) to check if one model reliably outperforms the other.

Additionally, since there is a class imbalance in our dataset (i.e., 50 anaclitic and 29 introjective patients), we experiment with oversampling techniques to balance the dataset and examine the effect on the results. Particularly, we use synthetic minority oversampling technique (SMOTE) (Chawla et al. 2002) to oversample from the introjective minority class so that the class distribution becomes 1:1. The approach generates new instances by interpolating positive instances (i.e., represented by vectors in the feature space) in the minority class that are close to each other.

The model evaluation for each feature category is shown in Table 2. For the *questionnaire features*, the Logistic regression model achieves the best performance with a macro F1 score of 0.588 and Kappa 0.194. Similarly, for the basic *TF-IDF features,* the LR model outperforms the other models with macro F1 score of 0.578 and kappa score of 0.194. The difference in performance between the best questionnaire model and best TF-IDF model is found to be not statistically significant. Moreover, the LR model based on advanced *BERT features*, which take into account the contextual information, outperforms the best model based on the TF-IDF features. We found that this improvement is statistically significant. The t-test with the null hypothesis that the means of the two models are equal at 0.05 level of significance is strongly rejected ($p < 0.001$). Another interesting finding is the max-pooled BERT feature representation yields better classification performance when compared with the the CLS representation confirming conclusions of previous works such as (Reimers and Gurevych 2019, Kim et al. 2021).

The best classifier performance using only *one feature category* is achieved using the more advanced features, designed for psychological analysis, i.e., LIWC features — the Random Forest classifier employing these features significantly outperforms all the other classifiers with F1 score of 0.896 and kappa score of 0.801. We observe that adding sentiment features and/or patient gender (i.e., the psychological features) does not benefit classification performance (i.e., F1 score drops to 0.884 and kappa drops to 0.778). However, combining the LIWC features with the questionnaire features produces the best classification performance. The Catboost model that takes the concatenation of the LIWC and questionnaire features as an input achieves 0.93 of F1 score and 0.867 of kappa κ score. The improvement is statistically significant with $p = 0.001$. Table 3 in Appendix A lists the significance analyses that compare the various models.

Looking at the *audio-based* classification models, we find that the random forest model yields the best results with 0.621 F1 score and 0.268 kappa score. Moreover, this model reliably outperforms the best TF-IDF model. Analyzing our models' means using the t-test results in rejecting the null hypothesis that they are equal with $p < 0.001$.

Finally, the oversampling technique (SMOTE) we used helps achieve higher classification performance as can be seen in Table 2. All the best-performing models in each feature category are built using the balanced data. We run a t-test on the best models in each feature category to determine the statistical significance of the stated improvements following dataset balancing. The SMOTE models' performance improvement is statistically significant as reported in Table 4 in Appendix A.

Table 2: Classification performance of 5-fold cross-validation results over 100 runs reported as mean ± standard deviation. The F1 scores are macro averages of the two classes. The best scores for each feature category are underlined, and the best scores over all categories are indicated in **boldface**.

Features	Models	imbalanced data		balanced data with SMOTE	
		F1	Kappa	F1	Kappa
Questionnaire	LR	0.586 ± 0.101	0.188 ± 0.194	<u>0.588 ± 0.097</u>	<u>0.194 ± 0.187</u>
	RF	0.498 ± 0.113	0.115 ± 0.160	0.554 ± 0.118	0.182 ± 0.188
	CB	0.453 ± 0.128	0.086 ± 0.171	0.573 ± 0.116	0.194 ± 0.204
TF-IDF	LR	0.570 ± 0.138	0.185 ± 0.228	<u>0.578 ± 0.130</u>	<u>0.195 ± 0.210</u>
	RF	0.503 ± 0.095	0.114 ± 0.158	0.524 ± 0.113	0.128 ± 0.192
	CB	0.526 ± 0.112	0.137 ± 0.188	0.543 ± 0.121	0.144 ± 0.209
LIWC	LR	0.822 ± 0.129	0.662 ± 0.229	0.802 ± 0.129	0.625 ± 0.230
	RF	0.882 ± 0.148	0.779 ± 0.268	<u>0.896 ± 0.127</u>	<u>0.801 ± 0.238</u>
	CB	0.838 ± 0.221	0.717 ± 0.363	0.873 ± 0.153	0.763 ± 0.273
Psychological	LR	0.851 ± 0.091	0.709 ± 0.175	0.845 ± 0.112	0.701 ± 0.20
	RF	0.874 ± 0.161	0.766 ± 0.295	<u>0.884 ± 0.142</u>	<u>0.778 ± 0.271</u>
	CB	0.824 ± 0.221	0.689 ± 0.366	0.868 ± 0.172	0.758 ± 0.306
RobBERT-CLS	LR	0.575 ± 0.080	0.182 ± 0.149	0.619 ± 0.090	0.248 ± 0.178
RobBERT-max	LR	0.606 ± 0.081	0.257 ± 0.127	<u>0.627 ± 0.110</u>	<u>0.282 ± 0.202</u>
Audio	LR	0.388 ± 0.006	0.076 ± 0.253	0.270 ± 0.014	0.082 ± 0.196
	RF	0.529 ± 0.143	0.169 ± 0.217	<u>0.621 ± 0.129</u>	<u>0.268 ± 0.248</u>
	CB	0.556 ± 0.149	0.156 ± 0.288	0.577 ± 0.099	0.173 ± 0.190
LIWC + Questionnaire	LR	0.572 ± 0.081	0.159 ± 0.151	0.591 ± 0.094	0.199 ± 0.180
	RF	0.822 ± 0.175	0.674 ± 0.300	0.865 ± 0.138	0.749 ± 0.247
	CB	0.812 ± 0.163	0.653 ± 0.286	<u>**0.930 ± 0.108**</u>	<u>**0.867 ± 0.206**</u>

4. Discussion and Conclusion

The *first aim* of this study was to gain a better understanding of anaclitic versus introjective personality traits through the analysis of natural language. We investigated whether using NLP tools to process natural utterances from patients (as recorded in interviews with a therapist) could uncover typical language features for anaclitic and introjective personality styles. We discovered several language characteristics common to one or the other personality style, such as word use, grammar structures, and language polarity. In line with our expectations, we found that anaclitic patients typically use words related to the interpersonal sphere (personal pronouns, 'social' category words). In addition, we found that the category 'home' was more common for anaclitic patients. This might be linked to the dominance of interpersonal themes, as the boundary between an anaclitic person and the outer world is weaker than for an introjective person. As a result, they tend to talk more about their home situation. For example: *"I moved to the family home and my parents moved to the house next door."* This example also helps explain why anaclitic patients — contrary to our a priori expectation — use the first person more frequently than introjective patients. In focusing on their relationships with others, they will use more personal pronouns than introjective patients, for talking about others (*'shehe'*) in relation to themselves (*'I'*). Interestingly, we also found that anaclitic patients tend to be more preoccupied with the present and the past (see LIWC categories *'focuspresent'*, *'focuspast'*), while introjective patients seem to care more about the future (LIWC category *'focusfuture'*). With a linguistic pun, this corresponds with hysterics' emphasis on history (hysteria) (Verhaeghe 2013), framing their complaints and symptoms in the context of their life story, much more than an obsessional or introjective patient would, the latter being more likely to isolate these different domains. Minimally, this finding suggest that a descriptive approach to personality styles must be supplemented with theory to understand the phenomenology of the different personality styles and comprehend the clinical usefulness of discriminating between them.

From the sentiment analysis, we note that the polarity of the questions asked by the interviewer is not differently distributed between the two personality styles. This suggests that polarized sentiment is not evoked by a different attitude of the therapist based on the patient's personality. We also note that introjective patients seem to have answers that are more aligned in sentiment with the interviewer, compared to anaclitic subjects who stick to giving more neutral answers (i.e., the fraction of answers corresponding in sentiment to that of the question asked is slightly higher for introjectives). Our findings within this regard are preliminary at best, but nonetheless suggest potential avenues for further research into turn taking in therapy.

The *second goal* of this research was to see whether personality classification into anaclitic or introjective personality styles can be automated well using machine learning techniques. We conclude from our findings that the advanced psychological features based on LIWC provide the most discriminating features. The best model, which achieves the maximum classification performance, is a random forest utilizing these features. When comparing these results to the ones obtained using only questionnaire features, the improvement is significant (increasing from 0.588 to 0.896 and 0.194 to 0.801 for F1 and kappa scores, respectively). More importantly, we find that combining these two feature categories further improves the classification performance. A CatBoost model trained on this combined feature set achieves 0.93 F1 score and 0.867 kappa score. See Table 3 in Appendix A for the t-test significance tests.

Classification of personality styles based on of self-report surveys have been criticized in the past (Coyne et al. 2004, Desmet 2007) due to concerns about the validity of the questionnaires introduced in Section 2.2. The findings of our study confirm that the questionnaires' discriminatory power is questionable, as demonstrated by the poor performance of machine learning models built with these as the only features when compared to (textual) data extracted directly from patient interviews. However, it is worth noting that questionnaires complement advanced textual characteristics, as seen by the rise in F1 score and Kappa score (i.e., LIWC-based model vs. LIWC+Questionnaire model) of an additional 3 and 6 percentage points, respectively. Our results suggest that we should go back to the basics when addressing the assessment issue, meaning that patients' narratives, what people say when they first come to consult a health care professional, is most indicative for the personality style of the patient. Our results can aid researchers and clinicians in their attempt to determine the personality style of their patients. Despite the good performance of our proposed automatic personality classification, we believe the classifying features that stem from our study should only be used to assist a human psychologist in confirming or challenging his/her thoughts. They should not replace the psychologist as a decision taker.

A limitation of this study is the relative small sample size (N = 79 patients). Even though the main goals of this study were achieved, a larger sample size might have led to more robust and significant findings. New data augmentation techniques that use large language models such as GPT4 could be employed to obtain additional and more realistic data points. Further improvements can thus possibly be established using larger samples. Further, the classifiers in this study were based on only a single intake interview per patient. Thus, we cannot confidently claim that we can generalize to intake conversations with structures different from the adopted CDI. Further research should aim to replicate our findings using alternative interviews or natural therapy intake sessions. As we focused specifically on a depressed population, further research could also aim to investigate a more heterogeneous (clinical) sample. Moreover, it would be interesting to investigate associations between personality styles and the process of therapy from diverse therapy orientations in an attempt to discover patterns that have remained obscure in traditional outcome-process research.

References

Al Hanai, Tuka, Mohammad M Ghassemi, and James R Glass (2018), Detecting depression with audio/text sequence modeling of interviews., *Interspeech*, pp. 1716–1720.

Beck, Aaron T (1983), Cognitive therapy of depression: New perspectives, *Treatment of depression: Old controversies and new approaches*, Raven Press.

Beck, Aaron T, Robert A Steer, and Gregory Brown (1996), Beck depression inventory–ii, *Psychological assessment.*

Bitew, Semere Kiros, Giannis Bekoulis, Johannes Deleu, Lucas Sterckx, Klim Zaporojets, Thomas Demeester, and Chris Develder (2019), Predicting suicide risk from online postings in Reddit the UGent-IDlab submission to the CLPsych 2019 Shared Task A, *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pp. 158–161.

Blatt, Sidney J (1974), Levels of object representation in anaclitic and introjective depression, *The psychoanalytic study of the child* **29** (1), pp. 107–157, Taylor & Francis.

Blatt, Sidney J (2004), *Experiences of depression: Theoretical, clinical, and research perspectives.*, American Psychological Association.

Blatt, Sidney J, Joseph P D'Afflitti, and Donald M Quinlan (1976), Experiences of depression in normal young adults., *Journal of Abnormal psychology* **85** (4), pp. 383, American Psychological Association.

Carletta, Jean, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. (2005), The AMI meeting corpus: A pre-announcement, *International workshop on machine learning for multimodal interaction*, Springer, pp. 28–39.

Cattell, Heather EP and Alan D Mead (2008), The sixteen personality factor questionnaire (16pf), *The SAGE handbook of personality theory and assessment* **2**, pp. 135–159, SAGE Publications Ltd. London.

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer (2002), SMOTE: synthetic minority over-sampling technique, *Journal of artificial intelligence research* **16**, pp. 321–357.

Coppersmith, Glen, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell (2015), CLPsych 2015 shared task: Depression and PTSD on Twitter, *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pp. 31–39.

Coyne, James C, Richard Thompson, and Valerie Whiffen (2004), Is the promissory note of personality as vulnerability to depression in default? reply to zuroff, mongrain, and santor (2004)., American Psychological Association.

Dávalos, Gonzalo Orellana, Luz Cantizano Rioseco, and Nelson Valdés Sánchez (2017), How do introjective and anaclitic patients speak? microanalysis of the words used by depressive patients during the psychotherapeutic process/¿ cómo hablan las pacientes introyectivas y anaclíticas? microanálisis de laspalabras utilizadas por pacientes depresivas durante el proceso psicoterapéutico?, *Revista Argentina de Clínica Psicológica* **26** (III), pp. 283, FUNDACIÓN AIGLÉ.

Delacre, Marie, Daniël Lakens, and Christophe Leys (2017), Why psychologists should by default use welch's t-test instead of student's t-test, *International Review of Social Psychology*, Ubiquity Press.

Delobelle, Pieter, Thomas Winters, and Bettina Berendt (2020), Robbert: a dutch roberta-based language model, *arXiv preprint arXiv:2001.06286.*

Derogatis, Leonard R (1992), Scl-90-r: Administration, scoring & procedures manual-ii for the (revised) version and other instruments of the psychopathology rating scale series., *Clinical Psychometric Research.* pp. 1–16.

Desmet, Mattias (2007), *Histerical and obsessive-compulsive depression: a psychometric study*, PhD thesis, Ghent University.

Desplanques, Brecht, Kris Demuynck, and Jean-Pierre Martens (2017), Adaptive speaker diarization of broadcast news based on factor analysis, *Computer Speech & Language* **46**, pp. 72–93, Elsevier.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018), Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805.*

Digman, John M (1990), Personality structure: Emergence of the five-factor model, *Annual review of psychology* **41** (1), pp. 417–440, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA.

Driessen, Ellen, Lisa M Hegelmaier, Allan A Abbass, Jacques P Barber, Jack JM Dekker, Henricus L Van, Elise P Jansma, and Pim Cuijpers (2015), The efficacy of short-term psychodynamic psychotherapy for depression: A meta-analysis update, *Clinical psychology review* **42**, pp. 1–15, Elsevier.

Dwyer, Dominic B, Peter Falkai, and Nikolaos Koutsouleris (2018), Machine learning approaches for clinical psychology and psychiatry, *Annual review of clinical psychology* **14**, pp. 91–118, Annual Reviews.

Eyben, Florian, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. (2015), The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing, *IEEE transactions on affective computing* **7** (2), pp. 190–202, IEEE.

First, Michael B, Robert L Spitzer, Miriam Gibbon, Janet BW Williams, et al. (2002), Structured clinical interview for dsm-iv-tr axis i disorders, research version, patient edition, *Technical report*, SCID-I/P New York, NY, USA:.

Fraley, R Chris, Niels G Waller, and Kelly A Brennan (2000), An item response theory analysis of self-report measures of adult attachment., *Journal of personality and social psychology* **78** (2), pp. 350, American Psychological Association.

Freud, Sigmund (1953), Three essays on the theory of sexuality (1905), *The standard edition of the complete psychological works of Sigmund Freud, volume VII (1901-1905): A case of hysteria, three essays on sexuality and other works*, pp. 123–246.

Fu, Cynthia HY, Yong Fan, and Christos Davatzikos (2019), Addressing heterogeneity (and homogeneity) in treatment mechanisms in depression and the potential to develop diagnostic and predictive biomarkers, *NeuroImage: Clinical* **24**, pp. 101997, Elsevier.

Gjurković, Matej and Jan Šnajder (2018), Reddit: A gold mine for personality prediction, *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pp. 87–97.

Golbeck, Jennifer, Cristina Robles, and Karen Turner (2011), Predicting personality with social media, *CHI'11 extended abstracts on human factors in computing systems*, pp. 253–262.

Goldberg, David (2011), The heterogeneity of "major depression", *World Psychiatry* **10** (3), pp. 226, World Psychiatric Association.

Hall, Margeret and Simon Caton (2017), Am i who i say i am? unobtrusive self-representation and personality recognition on facebook, *PloS one* **12** (9), pp. e0184417, Public Library of Science San Francisco, CA USA.

Hamilton, MAX (1967), Development of a rating scale for primary depressive illness, *British journal of social and clinical psychology* **6** (4), pp. 278–296, Wiley Online Library.

Horowitz, Leonard M (2002), *IIP: Inventory of interpersonal problems: Manual*, Psykologiförlaget.

Hovens, Johannes Eduard Joseph Marie, I Bramsen, and Hendrik Mels van der Ploeg (2000), *Handleiding bij de zelfinventarisatielijst posttraumatische stressstoornis*, Swets Test Publishers (STP).

Huang, Zhaocheng, Julien Epps, Dale Joachim, and Michael Chen (2018), Depression detection from short utterances via diverse smartphones in natural environmental conditions., *INTERSPEECH*, pp. 3393–3397.

Jain, Dipika, Akshi Kumar, and Rohit Beniwal (2022), Personality bert: A transformer-based model for personality detection from textual data, *Proceedings of International Conference on Computing and Communication Networks: ICCCN 2021*, Springer, pp. 515–522.

Joseph, Sethunya R, Hlomani Hlomani, Keletso Letsholo, Freeson Kaniwa, and Kutlwano Sedimo (2016), Natural language processing: A review, *International Journal of Research in Engineering and Applied Sciences* **6** (3), pp. 207–210.

Kalghatgi, Mayuri Pundlik, Manjula Ramannavar, and Nandini S Sidnal (2015), A neural network approach to personality prediction based on the big-five model, *International Journal of Innovative Research in Advanced Engineering (IJIRAE)* **2** (8), pp. 56–63.

Kamenskaya, Ekaterina and Georgy Kukharev (2008), Recognition of psychological characteristics from face, *Metody Informatyki Stosowanej* **1** (1), pp. 59–73, Citeseer.

Kampman, Onno, Elham J Barezi, Dario Bertero, and Pascale Fung (2018), Investigating audio, visual, and text fusion methods for end-to-end automatic personality prediction, *arXiv preprint arXiv:1805.00705*.

Kazameini, Amirmohammad, Samin Fatehi, Yash Mehta, Sauleh Eetemadi, and Erik Cambria (2020), Personality trait detection using bagged svm over bert word embedding ensembles, *arXiv preprint arXiv:2010.01309*.

Kim, Taeuk, Kang Min Yoo, and Sang-goo Lee (2021), Self-guided contrastive learning for BERT sentence representations, *arXiv preprint arXiv:2106.07345*.

Lambert, Michael J, Ann T Gregersen, and Gary M Burlingame (2004), The outcome questionnaire-45., Lawrence Erlbaum Associates Publishers.

Levitan, Sarah Ita, Yocheved Levitan, Guozhen An, Michelle Levine, Rivka Levitan, Andrew Rosenberg, and Julia Hirschberg (2016), Identifying individual differences in gender, ethnicity, and personality from dialogue for deception detection, *Proceedings of the second workshop on computational approaches to deception detection*, pp. 40–44.

Lovibond, Sydney H and Peter F Lovibond (1996), *Manual for the depression anxiety stress scales*, Psychology Foundation of Australia.

Mairesse, François and Marilyn Walker (2006), Automatic recognition of personality in conversation, *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pp. 85–88.

Mairesse, François, Marilyn A Walker, Matthias R Mehl, and Roger K Moore (2007), Using linguistic cues for the automatic recognition of personality in conversation and text, *Journal of artificial intelligence research* **30**, pp. 457–500.

Majumder, Navonil, Soujanya Poria, Alexander Gelbukh, and Erik Cambria (2017), Deep learning-based document modeling for personality detection from text, *IEEE Intelligent Systems* **32** (2), pp. 74–79, IEEE.

Matero, Matthew, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammadzaman Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H Andrew Schwartz (2019), Suicide risk assessment with multi-level dual-context language and bert, *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pp. 39–44.

Meganck, Reitske, Mattias Desmet, Claudi Bockting, Ruth Inslegers, Femke Truijens, Melissa De Smet, Rosa De Geest, Kimberly Van Nieuwenhove, Vicky Hennissen, Goedele Hermans, et al. (2017), The ghent psychotherapy study (gps) on the differential efficacy of supportive-expressive and cognitive behavioral interventions in dependent and self-critical depressive patients: study protocol for a randomized controlled trial, *Trials* **18** (1), pp. 1–11, Springer.

Memon, Shahan Ali (2020), Acoustic correlates of the voice qualifiers: A survey, *arXiv preprint arXiv:2010.15869*.

Mundt, James C, Adam P Vogel, Douglas E Feltner, and William R Lenderking (2012), Vocal acoustic biomarkers of depression severity and treatment response, *Biological psychiatry* **72** (7), pp. 580–587, Elsevier.

Myers, Isabel Briggs and Mary H McCaulley (1988), *Myers-Briggs type indicator: MBTI.*, Consulting Psychologists Press Palo Alto.

Newman, Matthew L, Carla J Groom, Lori D Handelman, and James W Pennebaker (2008), Gender differences in language use: An analysis of 14,000 text samples, *Discourse processes* **45** (3), pp. 211–236, Taylor & Francis.

Özkanca, Yasin Serdar, Cenk Demiroğlu, Asli Besirli, and Selime Celik (2018), Multi-lingual depression-level assessment from conversational speech using acoustic and text features, *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association.

Pennebaker, James W (2011), The secret life of pronouns, *New Scientist* **211** (2828), pp. 42–45, Elsevier.

Pennebaker, James W and Laura A King (1999), Linguistic styles: language use as an individual difference., *Journal of personality and social psychology* **77** (6), pp. 1296, American Psychological Association.

Perkins, Jacob (2010), *Python text processing with NLTK 2.0 cookbook*, PACKT publishing.

Plank, Barbara and Dirk Hovy (2015), Personality traits on twitter—or—how to get 1,500 personality tests in a week, *Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pp. 92–98.

Ramos, Juan et al. (2003), Using tf-idf to determine word relevance in document queries, *Proceedings of the first instructional conference on machine learning*, Vol. 242, Citeseer, pp. 29–48.

Reimers, Nils and Iryna Gurevych (2019), Sentence-bert: Sentence embeddings using siamese bert-networks, *arXiv preprint arXiv:1908.10084*.

Robins, Clive J and Alice G Luten (1991), Sociotropy and autonomy: differential patterns of clinical presentation in unipolar depression., *Journal of abnormal psychology* **100** (1), pp. 74, American Psychological Association.

Rodrigues Makiuchi, Mariana, Tifani Warnita, Kuniaki Uto, and Koichi Shinoda (2019), Multimodal fusion of bert-cnn and gated cnn representations for depression detection, *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pp. 55–63.

Schwartz, H Andrew, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. (2013), Personality, gender, and age in the language of social media: The open-vocabulary approach, *PloS one* **8** (9), pp. e73791, Public Library of Science San Francisco, USA.

Stachl, Clemens, Florian Pargent, Sven Hilbert, Gabriella M Harari, Ramona Schoedel, Sumer Vaid, Samuel D Gosling, and Markus Bühner (2020), Personality research and assessment in the era of machine learning, *European Journal of Personality* **34** (5), pp. 613–631, SAGE Publications Sage UK: London, England.

Tausczik, Yla R and James W Pennebaker (2010), The psychological meaning of words: Liwc and computerized text analysis methods, *Journal of language and social psychology* **29** (1), pp. 24–54, Sage Publications Sage CA: Los Angeles, CA.

Valdés, Nelson and Mariane Krause (2015), Verbal expressions used by anaclitic and introjective patients with depressive symptomatology: Analysis of change and stuck episodes within therapeutic sessions, *Clínica y Salud* **26** (2), pp. 103–119, Elsevier.

van Rijsbergen, Gerard D, Huibert Burger, Steven D Hollon, Hermien J Elgersma, Gemma D Kok, Jack Dekker, Peter J de Jong, and Claudi LH Bockting (2014), How do you feel? detection of recurrent major depressive disorder using a single-item screening tool, *Psychiatry research* **220** (1-2), pp. 287–293, Elsevier.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017), Attention is all you need, *Advances in neural information processing systems*.

Verhaeghe, Paul (2013), *Does the woman exist?: from Freud's hysteric to Lacan's feminine*, Other Press, LLC.

Werbart, Andrzej and David Forsström (2014), Changes in anaclitic–introjective personality dimensions, outcomes and psychoanalytic technique: A multi-case study, *Psychoanalytic Psychotherapy* **28** (4), pp. 397–410, Taylor & Francis.

Werbart, Andrzej, Mikael Hägertz, and Nadja Borg Ölander (2018), Matching patient and therapist anaclitic–introjective personality configurations matters for psychotherapy outcomes, *Journal of Contemporary Psychotherapy* **48** (4), pp. 241–251, Springer.

Werbart, Andrzej, Siri Aldén, and Anders Diedrichs (2017), Changes in the anaclitic-introjective personality configurations following psychoanalytic psychotherapy with young adults, *Research in Psychotherapy: Psychopathology, Process, and Outcome*, PAGEPress.

Westen, Drew and Serra Muderrisoglu (2006), Clinical assessment of pathological personality traits, *American Journal of Psychiatry* **163** (7), pp. 1285–1287, Am Psychiatric Assoc.

Yang, Feifan, Xiaojun Quan, Yunyi Yang, and Jianxing Yu (2021), Multi-document transformer for personality detection, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, pp. 14221–14229.

Zhu, Yangfu, Linmei Hu, Xinkai Ge, Wanrong Peng, and Bin Wu (2022), Contrastive graph transformer network for personality detection, *Proc. IJCAI Conf.*

5. Appendices

Appendix A. Statistical Significance test details

This section provides the statistical significance tests we run to check whether one model reliably outperforms another. Table 3 shows the p-value after running a t-test to compare the best models in each feature category. First, we run each model 100 times, which thus yields 100 kappa scores. We then use a t-test to test the null hypothesis that the means of the two models are equal at a 0.05 level of significance. The only comparison where we found no significant difference is between the questionnaire and TF-IDF-based models with p=0.847. We found that the difference was significant for each of the other comparisons.

Similarly, Table 4 shows the p-values for statistical significance tests using the t-test. The aim is to examine whether the performance increase gained by balancing the dataset is significant. Again, we test the null hypothesis that the means (in terms of kappa scores) of balanced- vs. unbalanced-based models are equal at a 0.05 level of significance. As shown in the table, the null hypothesis was rejected for all the comparisons, indicating that improvements of using SMOTE are significant.

Table 3: The comparison of best-performing models (in terms of kappa score k) from each category using the t-test. Each model is run 100 times yielding 100 kappa scores. We then use a t-test to compare the different models. Null hypothesis: the means of the two models are equal.

	Question-naire κ=0.194	TF-IDF κ=0.194	LIWC κ=0.801	Psycho-logical κ=0.778	BERT-CLS κ=0.248	BERT-max κ=0.282	Audio κ=0.268	LIWC+Question-naire κ=0.867
Questionnaire (κ=0.194)	-	p=0.847	p <0.001	p <0.001	p <0.001	p <0.001	p <0.001	p <0.001
TF-IDF (κ=0.194)	-	-	p <0.001	p <0.001	p <0.001	p <0.001	p <0.001	p <0.001
LIWC (κ=0.801)	-	-	-	p <0.001	p <0.001	p <0.001	p <0.001	p <0.001
Psychological (κ=0.778)	-	-	-	-	p <0.001	p <0.001	p <0.001	p <0.001
BERT-CLS (κ=0.248)	-	-	-	-	-	p <0.001	p <0.001	p <0.001
BERT-max (κ=0.282)	-	-	-	-	-	-	p <0.001	p <0.001
Audio (κ=0.268)	-	-	-	-	-	-	-	p <0.001

Table 4: The statistical significance test results for comparing models based on a balanced vs. unbalanced dataset. For each feature category, we suse the best performing model type (i.e., RF, LR, CB) and compare a model that uses SMOTE (in terms of kappa score) with the counterparts that uses the unbalanced dataset. Each model is run 100 times yielding 100 kappa scores. We then use a t-test to compare the different models. Null hypothesis: the means of the SMOTE vs. unbalanced dataset-based models are equal.

Features	Dataset		t-test
	Unbalanced *Kappa* κ	Balanced (SMOTE) *Kappa* κ	
Questionnaire	0.188	0.194	p <0.001
TF-IDF	0.185	0.194	p <0.001
LIWC	0.779	0.801	p <0.001
Psychological	0.766	0.778	p <0.001
BERT-CLS	0.182	0.248	p <0.001
BERT-max	0.257	0.282	p <0.001
Audio	0.169	0.268	p <0.001
LIWC+Ques'ire	0.653	0.867	p <0.001