Proceedings of the RaPID-4 @LREC 2022, pages 31–40
Marseille, 25 June 2022

Classification of German Jungian Extraversion and Introversion Texts with Assessment of Changes during the COVID-19 Pandemic

Dirk Johannßen[1,2], **Chris Biemann**[1], **David Scheffer**[2]
[1]MIN Faculty, Dept. of Informatics, Universität Hamburg, [2]Dept. of Economics, Nordakademie
[1]22527 Hamburg, Germany [2]25337 Elmshorn, Germany
{biemann, johannssen}@informatik.uni-hamburg.de, {david.scheffer, dirk.johannssen}@nordakademie.de
http://lt.informatik.uni-hamburg.de/

Abstract

The corona pandemic and countermeasures such as social distancing and lockdowns have confronted individuals with new challenges for their mental health and well-being. It can be assumed that the Jungian psychology types of extraverts and introverts react differently to these challenges. We propose a Bi-LSTM model with an attention mechanism for classifying introversion and extraversion from German tweets, which is trained on hand-labeled data created by 335 participants. With this work, we provide this novel dataset for free use and validation. The proposed model achieves solid performance with $F_1 = .72$. Furthermore, we created a feature engineered logistic model tree (LMT) trained on hand-labeled tweets, to which the data is also made available with this work. With this second model, German tweets before and during the pandemic have been investigated. Extraverts display more positive emotions, whilst introverts show more insight and higher rates of anxiety. Even though such a model can not replace proper psychological diagnostics, it can help shed light on linguistic markers and to help understand introversion and extraversion better for a variety of applications and investigations.

Keywords: NLP, COVID-19, Implicit Motives, Introversion, Extraversion

1. Introduction

The first cases of individuals reportedly being infected with the SARS-CoV-2 or COVID-19 virus appeared in December of 2019. Ever since, a global pandemic of this highly infectious disease has emerged, which has been met with countermeasures. Those countermeasures include social distancing and temporary lockdowns (Balasa, 2020). Governments stand in the dichotomy of restricting social and public interactions as a measure of safety and risking the mental health of the people affected, as reports of declining mental well-being emerge (Hämmig, 2019).

Even though professional mental consultation and support do exist, it is difficult to identify and contact heavily impacted individuals (Lester and Howe, 2008). The direct approach would not be feasible, as it would tie up the capacities of mental health workers. Broad information campaigns might cause high costs and still not reach individuals in need. Lastly, affected people might not even be aware of their mental health risks and thus not reach out to available mental health consultations. Depression detection systems or even sentiment analyses of e.g. social media posts could potentially support mental health workers (Coppersmith et al., 2018). But those systems often rely on sufficient self-reports or on topics of mental health or loneliness being directly discussed, which require the individuals to already self-reflect and openly discuss their well-being, resp. the decline thereof (Zirikly et al., 2019).

Furthermore, the well-established safety net of e.g. educational facilities, whose staff could identify troubled individuals, can be unavailable due to the lockdown restrictions. Thus, it might be worthwhile to explore alternative and ideally automated approaches. Carl Gustav Jung researched psychological types (also known as psychological archetypes, (Jung, 1921)), and proposed two perceiving types – sensation and intuition – and two judging types – thinking and feeling. Furthermore, those types are moderated or influenced by the main attitude – extraversion and introversion.

Mental health detection often focuses on introverts due to their self-inflicted distancing and more frequent occurrence of signs of depression compared with extraverts. Recent empirical research on the effects of the pandemic confirms those findings (Wei, 2020). Other findings, however, contradict those results and report empirical findings of extraverts' suffering to be comparably worse (Wijngaards et al., 2020).

As with many psychometrics, manual assessment of psychology types can be costly (Johannßen et al., 2019). Furthermore, burdened individuals might not be reachable by broadly conducted surveys amongst a population. Thus, automation of those types with a focus on introverts and extraverts might reveal the additional potential for identifying individuals in need of support. Therefore, with this work, we aim to classify the Jungian psychological types of *extraversion* and *introversion* from German text and to apply such a model to utterances in 2019 compared with 2020 to investigate whether there are noteworthy well-being differences.

In this work, we will first discuss related work to automated psychometrics, depression detection, and some psychometrics in Section 2. Thereafter, the basics of the Jungian psychological types will be laid out in Section 3. The implicit personality test (IPT) utilized in this

work is described in Section 4, followed by the description of the dataset for training neural models and for identifying anxious individuals in Section 5. Section 6 discusses the methodology and approach. The results will be presented in Section 7 and will be discussed in Section 9. We conclude our findings in Section 10 and discuss future outlooks.

2. Related Work

The automated assessment of personality or personality traits is a rather recent application domain. Whilst earlier approaches relied more heavily on rule-based systems, themselves mostly divided into wordlist-based versus corpus-induced methods (Johannßen and Biemann, 2018), machine learning has become more widely utilized in recent years (Mehta et al., 2019). Accordingly, the MBTI and the five-factor model of personality (also called *Big Five*, (Goldberg, 1993)) have been (Angleitner, 1991) and are amongst the most widely utilized personality tests, both of which rely on the Jungian psychological typologies (see Section 3).

Jungian types have successfully been classified from natural language texts by employing a BERT model by Keh et al. (2019). For training their model, the authors scraped data from a self-reporting web forum. The resulting model was utilized for generating personality-induced natural language texts.

The effects of the COVID-19 pandemic have been researched extensively during its outbreak at the end of 2019. Johannßen & Biemann (2020) analyzed social unrest indicators on the application of the pandemic and found that an increase of an implicit motive *power* paired with a self-regulatory passive coping with fears were correlated with signs of crises.

Empirical research on the impacts of the COVID-19 pandemic on introverts and extraverts is somewhat contradictory. Whilst some recent works found extraverts to be more in danger of mental health degradation (Wijngaards et al., 2020; Gubler et al., 2020), other works come to the opposite conclusion (Wei, 2020).

3. Jungian psychological typologies

In "Psychological Types", Jung (1921) distinguished two main types, the Persona, and the Shadow. Whilst the Persona of a person is being shown to the environment and is individualistic, the Shadow remains disguised and is part of a collective unconsciousness. With this view, Jung differed from his tutor Freud to the extent that Freud assumed for the psyche to only be individual. Jung, on the other hand, assumed for humanity to share a collective unconsciousness, which manifests in the form of collectively shared psychological types, that determine our intrinsic desires.

Accordingly, there are two main types, namely the extraverts (e), and the introverts (i). A person either belongs to the former or the latter. Those two types moderate (i.e. influence) all other types, namely sensation (s) vs. intuition (n), thinking (t) vs. feeling (f), and judging (j) vs. perceiving (p).

Based on Jung's psychological types, many psychological tests, and psychometrics emerged thereafter, partly applying the theory directly or extending it. The modality and methodology of measuring types are versatile. Some employ direct questionnaires (e.g. the original Myers-Briggs Type-Indicator (MBTI), (Myers et al., 2000)), some employ visual assertions (e.g. the visual questionnaire or ViQ, (Scheffer and Manke, 2018)) and others analyze natural language (e.g the IPT, which will be described in Section 4).

Even though many of those testing procedures were not psychologically asserted in terms of reliability, stability and validity (e.g. the Big Five or MBTI), those psychological tests that are based on Jung's psychological types have nevertheless frequently been utilized for typing individuals, and were correlated with behavioral observations (Rammstedt et al., 2018).

4. Implicit personality test (IPT)

It is difficult to measure the psyche or personality directly (Fried and Flake, 2018). The research field of psychology has developed and researched different approaches for measuring manifestations of the underlying mental processes, all of which have advantages and shortcomings. E.g. psychoanalysis tries to assume cognitive mechanisms and past events in dialogues, whilst behaviorism strictly limits statements on empirical and reproducible observations (Mahoney, 1984). Both approaches require controlled environments, extensive manual labor, and time. Testing procedures try to determine personality traits with limited time and budget and thus oftentimes balance reliability (i.e. are results reproducible?), validity (i.e. do results correspond to other observations and measures?), and limited testing resources (Schultheiss and Brunstein, 2010, p. 76f).

Some personality testing procedures utilize questionnaires with high reliability. However, standardized surveys and direct questionnaires at times suffer from socio-expectation bias, i.e. participants rather worry about, what testing personnel might think about them, when answering a question in a certain way, rather than answering freely. This bias can occur if the intentions of questions can be guessed or are assumed (Bogner and Landrock, 2016).

Implicit or projective testing procedures overcome this shortcoming by providing participants with ambiguous and situational imagery and asking them to answer questions e.g. who the main character is and what that individual experiences and feels. Those projective methods reveal intrinsic desires. Since there is no socially accepted or wrong answer, the socio-expectation bias is said to be less severe. However, projective methods have been criticized for their reliability (Schultheiss and Brunstein, 2010, p. 119ff).

The IPT is such an implicit test and confronts participants with imagery such as displayed in Figure 1. Par-

ticipants chose the main person and answer questions about what is happening and how that person feels. Some of those answers, manually labeled with either i (introvert) or e (extravert) are displayed in Listing 1. The human annotators are psychologists and receive extensive training, which initially is wordlist centered but shifts to narrations over time[1]. The IPT is based on the MBTI and has mainly been utilized for business-oriented aptitude diagnostics.

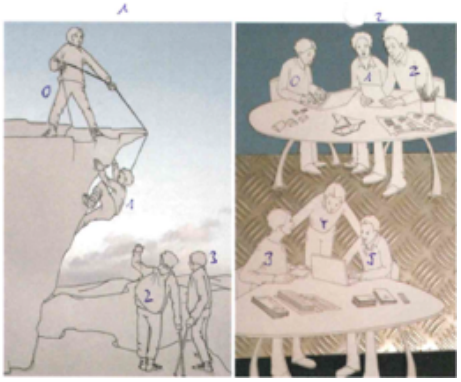

Figure 1: During the IPT, participants are presented with projective imagery, to which they answer questions such as who the main person might be and what that person is experiencing. Such projective or implicit tests are designed to reveal intrinsic desires (Schultheiss and Brunstein, 2010).

```
I   Sie sieht ihre Schüler. das die
    Schüler nach hause gehen. genervt
E   Erklärt jemandem etwas. Es
    richtig zu machen. Er kann es
    ——————— Translated from German ———————
I   she sees her students. That the
    students go home. Annoyed
E   Explains something to someone.
    To do it right. He can do it
```

Listing 1: Short examples of answers given during the IPT and corresponding manual labels

5. Data

Since manually asserting natural language texts on introversion or extraversion is costly and would not be scalable, we will first train a neural model (see Section 6) on the data described in this section. We collected German natural language textual data before and from the COVID-19 pandemic and apply said model to this data set. Furthermore, we train in-domain Twitter models.

Model training data

The German natural language textual data utilized for creating the model was collected by a company specialized in aptitude diagnostical testing[2] and is being made public for free use and validation[3]. 2,680 textual answers to provided projection imagery were given by 335 individuals. The population was drawn from the workforce with ages ranging from 18 to 65. Further demographic information was omitted under German data protection laws. The data has been split by separating participants into training (~90%, n=2,360), development, and held-out testing data sets (~5%, n=160 each). Since all 8 answers per participant remained in a data set without being shuffled and separated, we aim to increase the generalization of the model (i.e. rather training to learn the target label and not perform speaker identification). The distribution of answers labeled as extraversion is displayed in Table 1. The two labels are distributed unevenly with the vast majority being extraversion (67.4% of all labels with comparable distributions overall data sets). Answers consist of an average of 42 words and thus can be considered short texts. Each answer has been manually labeled with the four typology pairs. Compared to data sources like Twitter, the training data is rather clean without a lot of noise such as spelling mistakes, spam, or unusual characters. The Kohen's Cappa measure for annotator agreement on the task of extraversion and introversion IPT scores $K = .47$ – only *moderate agreement* (McHugh, 2012).

# extra	8	7	6	5	4	3	2	1	0
%	9.7	22.0	21.4	17.3	13.5	8.5	5.9	1.5	.3

Table 1: Distribution of answers labeled as extraversion in the training material. The upper row displays the counts of answers labeled as extraversion per participant (8 answers in total), the lower row displays the corresponding percentages. 67.4% of all instances were labeled with extraversion and 32.6% with introversion.

Experimental data

One goal of this work is to research transferability across different data domains, namely from the IPT to tweets. Before utilizing any model for validation purposes on tweets, we first need to measure transferability. For this validation data, we sampled 1,100 tweets from a corpus described hereafter, and had them manually labeled by experts on extraversion and introversion. The agreement scores $K = .68$ – which is a *strong agreement* (McHugh, 2012). The data is also made available[4].

Validation data

The experimental data was drawn from Twitter[5], a

[1]For a closely related testing procedure, please refer to Kuhl & Scheffer (1999)

[2]WafM Wirtschaftsakademie GmbH https://www.wafm.de/.

[3]https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/ipt-introextra-2022.html.

[4]https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/ipt-introextra-2022.html.

[5]Twitter https://www.twitter.com

micro-messaging service. The service offers an API for downloading 1% of the worldwide traffic of the social network (Gerlitz and Rieder, 2013). Since the goal of this research is to find new ways of identifying individuals in need during the COVID-19 pandemic, we crawled the Twitter API for the period from March to May 2019 and from March to May 2020. Linguistically, the samples are comparably similar (e.g. equal average lengths, equal part-of-speech (POS) tags, sentence lengths, etc).

The crawled instances were filtered by a German flag to only include posts from German individuals. Furthermore, we filtered non-German samples via language detection (Google translate python library[6].). Besides the texts themselves, the field *date time* was included, which functions both as an identifier hence the inclusion of milliseconds, and as an inclusion criterion for the experimental setup. In total, 10,000 instances were sampled, 5,000 per time period (2019, 2020). An answer from 2019 contains 19.77 words on average and 19.76 from 2020, which makes this a short-text classification task. Bias effects have to be assumed when comparing two different time periods. We aimed to reduce this bias by spreading the selection period over three months, hence selective topics like sports, weather, or cultural events should not overshadow the overreaching effects the pandemic might have.

6. Methodology

In this methodology section, we propose a two-stage approach to asserting domain transferability, describe two employed model architectures, and present the experimental setup.

Two-stage approach

Since there is a considerable difference in labeled data quality and availability between the training data from the IPT and the experimental validation data from Twitter, and since it can be assumed that domain transferability does not produce convincing results, we propose two consecutive experimental stages: i) first, we will train two models from previous experiments (Johannßen et al., 2019; Johannßen and Biemann, 2020) on the IPT data set and validate them on the Twitter dataset, and ii) secondly, we will train those models directly on the Twitter validation set. We critically evaluate transferability and validation applicability, as it is often aspired when performing NLP on psychological textual data (Stajner and Yenikent, 2021; Plank and Hovy, 2015a).

Bi-LSTM attention Model

Previous work on German natural language textual data with a focus on psychological measures have resulted in a viable model, which has reached state-of-the-art results on a shared task dataset and is being utilized for this work as well (Johannßen et al., 2019; Johannßen and Biemann, 2020).

The first model is displayed in Figure 2 and consists of a bi-directional long short-term memory (LSTM, (Hochreiter and Schmidhuber, 1997)) neural network, combined with an attention mechanism.

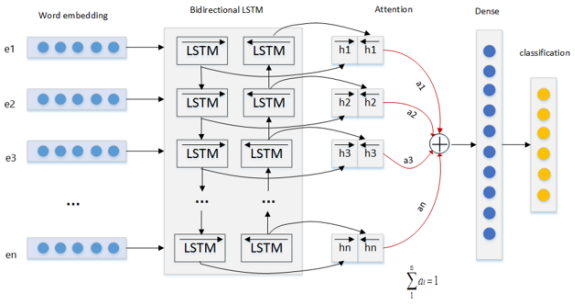

Figure 2: The employed model is a bi-directional long short-term memory neural network, combined with an attention mechanism (image by (Zhou and Wu, 2018)). This type of architecture allows for the model to observe the input from both sides, left and right. The attention supports algorithmic decisions made and at times allows for an analysis of more algorithmic important parts of an input or instance.

In addition to weight connections between each layer to its successor, LSTMs (a special type of Recurrent Neural Network (RNN)) also possess connections between units of the same layer. Furthermore, LSTMs possess a so-called forget gate, which can control which part of an unlimited memory to keep for decisions and which to *forget*. A bi-directional network combines both directions – forward and backward – of input and concatenates the impacts of a token in dependence of the previous and following context of this token. Lastly, the attention mechanism (Bahdanau et al., 2014) models the algorithmic importance of a network by multiplying hidden states with an alignment score to create a context vector, which then gets concatenated with a previous output.

The model is constructed with 5 layers (1 input, 3 hidden, 1 output) and contains 256 units in each hidden layer. Input tokens are represented by 300-dimensional fasttext embeddings, pre-trained on *Common Crawl*[7] and *Wikipedia*[8] (Grave et al., 2018). As optimizer we chose Adam (Kingma and Ba, 2017) and the loss was calculated via cross-entropy. Training parameters were set to a step-width of 1e-6, a dropout rate of .5, and mini-batch training of size 32 in 50 epochs.

Logistic Model Tree (LMT) Model

Since previous approaches (Johannßen et al., 2019) have shown strong results from trained logistic model trees on small datasets (LMT, Landwehr et al. (2005)), we trained an LMT, which is a decision tree with logistic regressions at its leaves, as a second model to be considered. We performed feature engineering but opted

[6] https://pypi.org/project/googletrans/

[7] Common Crawl, https://commoncrawl.org/.

[8] Wikipedia, https://www.wikipedia.org/.

for two different sets of hand-crafted features: one set of features for modeling the IPT and one set of features for modeling the same task on tweets directly.

IPT LMT: As described in our previous work (Johannßen et al., 2019), for firstly engineering the IPT features, the texts mostly were tokenized and processed per token. Engineered features were the type-token-ratio, the ratio of spelling mistakes, and frequencies between 3 and 10 appearances. Further features are LIWC and language model perplexities. The psychometric dictionary and software *language inquiry and word count* (LIWC) was developed by Pennebaker et al. (1999) and later transferred to German by Wolf et al. (2008). LIWC is a simple wordlist-based but well-established tool amongst psychologists and has been utilized for both, the private sector and research. When analyzing a text, LIWC increments categories (i.e. positive emotions, cognitive processes, or anxiety) based on matching dictionary terms per category, which have previously been psychologically validated (Wolf et al., 2008). E.g. the category *family* contains words such as sister, father, mother, mom, etc. The counts per category then get normalized over the length of the input. The results are percentages of words belonging to each category. The German LIWC allows for 96 categories to be assigned to each token, ranging from rather syntactic features such as personal pronouns to rather psychometric values such as familiarity, negativity, or fear. Part-of-speech (POS) tags were assigned to each token and thereafter counted and normalized to form a token ratio. We trained a POS tagger via the natural language toolkit (NLTK) on the TIGER corpus, assembled by Brants at al. (2004) and utilizing the STTS tagset, containing 54 individual POS tags.

We trained a bigram language model for each class and incorporated Good-Turing smoothing for calculating the perplexity. During training, we tuned parameters (e.g. which smoothing to use) via development set and tested the model with the held-out test set. The perplexity of a model q is: $2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 q(x_i)}$, with p being an unknown probability distribution, $x_1, x_2, \ldots x_N$ being the sequence (i.e. the sentence) drawn from p and q being the probability model.

Twitter model: Secondly, we engineered features for the same task on the labeled Twitter data directly. For the class extraversion, the most influential tasks reflected upon stimulus *from the outside*, such as many add symbols (@) and hashtags (#), plural forms, and plural pronouns. Furthermore, multiple exclamation marks (often used by German speakers to emphasize and *shout*), instances written in all caps, and emojis indicate extraversion in tweets. As for introversion, mostly the opposite features indicate the class: only few emojis, exclamation marks, hashtags, or add symbols. Singular forms and singular pronouns indicate introversion, as well as lowercased tokens (unusual in German, since common and proper nouns are spelled with an initial uppercase).

Pre-processing
Since additional features did not enhance the model's performance metrics in preliminary experiments, we decided against adding any (e.g. POS tags, spelling mistakes, or linguistic inquiry and word count (LIWC, (Pennebaker et al., 2007; Wolf et al., 2008)) category counts). We follow the pre-processing steps by Johannßen & Biemann (2020) by removing stop-words, numbers, emojis, or Twitter-typical special characters, as well as auto-correcting spelling mistakes. 1.000 remaining pre-processed tweets were drawn.

Experimental setup
As described in Section 1, there are contradictory empirical findings on whether introverts or extraverts are more mentally challenged during the pandemic. To investigate this contradiction, we collected data from 2019 and 2020, as described in Section 5. The proposed models (see Section 6) will be trained on the task of classifying extraverts and introverts by their use of natural textual language and will thereafter be utilized for classifying labels to the tweets from 2019 and 2020. Finally, we will divide extraverts and introverts of both years and investigate their linguistic tone and mood. This investigation will be performed by the use of LIWC. From those LIWC category word percentages, we will investigate, whether the tone of extraverts and introverts have significantly changed and in which way.

7. Results

Model benchmarks
Firstly, we performed benchmarks to confirm our model choices. The Benchmarks displayed in Table 2 have shown that the proposed Bi-LSTM model with attention mechanism achieves the best results on this classification task, even outperforming a BERT base model. It can be assumed that BERT base fails to capture the task due to little training data and diverging content meanings compared with everyday use of language (Ezen-Can, 2020).

Model	Accuracy	Precision	Recall	F1 Score
BERT base	.70	.49	.70	.58
CNN	.72	.70	.72	64
LMT + features	.66	.65	.66	.65
RNN	.66	.64	.66	.65
Self attention	.68	.71	.68	.69
LSTM	**.73**	.70	**.73**	.69
Bi-LSTM attn.	.71	**.73**	.71	**.72**

Table 2: Benchmark performances of different model architectures. The proposed Bi-LSTM model with attention mechanism achieves the highest F1 score. Whilst oftentimes BERT outperforms other architectures, the employed BERT base might fail to capture the signals due to diverging content meanings compared with everyday language use (Ezen-Can, 2020).

IPT model performances and Twitter validation
The confusion matrix of the IPT Bi-LSTM is displayed

in Table 4. The current state-of-the-art (SOTA) approach for classifying English introversion and extraversion by Plank & Hovy (2015b) scores $F_1 = .72$. Even though those scores are not comparable due to the differing languages and datasets, the proposed model nonetheless achieves comparable results with $F_1 = .72$ on the task with German textual data. The performance of the IPT LMT model is slightly worse than the performance of the Bi-LSTM attention model with $F_1 = .69$ with perplexity (and thus introversion/extraversion bigram language models) being the discriminating feature on its root node.

Model	**Bi-LSTM att.**	**LMT**
Precision	.736	.693
Recall	.7125	.685
F-Measure	.7203	.689

Table 3: Bi-LSTM attention model and LMT model performance measures of precision, recall, and the F-measure for the task of classifying the Jungian psychology types of extraversion and introversion. The model was trained on the IPT.

		Predicted		
		Extra	**Intro**	Σ
Actual	**Extra**	83	29	112
	Intro	17	31	48
	Σ	100	60	160

Table 4: The confusion matrix of the Bi-LSTM attention model on the IPT classification task test set.

Despite the proposed Bi-LSTM model scoring well on the held-out test IPT dataset, it does not validate well on the experimental Twitter dataset. When utilizing this model on a held-out test set ($n = 160$) of the 1,000 hand-labeled tweets and measuring its performance, the model scores $F_1 = .5$, indicating uninformed decisions based on chance. The same can be observed for the proposed IPT LMT model, which scores an even worse $F_1 = .3$, rendering it unapplicable for cross-domain tweet classification.

In-domain Twitter model and validation

The proposed Bi-LSTM model with attention mechanism fails to capture the aspects of introversion and extraversion from the small Twitter dataset. The model scores a mere $F_1 = .4$ on the Twitter held-out test set and thus is not applicable for being utilized for any further predictions.

In contrast to the Bi-LSTM model, the feature engineered and in-domain trained LMT twitter model achieves good results on the held-out Twitter test set with $F_1 = .69$. The LMT model's confusion matrix is displayed in Table 5, showing that the model performs sufficiently well on both classes and especially introversion, which seems to be harder to model in general (Stajner and Yenikent, 2021). Influential features include the POS tags KOUI, PPOSAT, VAPP, and pronouns, as well as LIWC categories Other, Past, School, and Physical. Lastly, frequencies of exclamation marks, hashtags, emojis, and add tags.

From those results, we can conclude that the out-of-domain transferability between IPT models and tweets does not validate. The Bi-LSTM model performs well on the IPT but fails when being trained directly on the Twitter dataset. The LMT IPT model performs slightly worse. When training a feature-engineered LMT directly on tweets, it performs sufficiently. Hereafter, we will only discuss the IPT Bi-LSTM and Twitter LMT. Additionally, we will utilize the Twitter LMT for further validation studies on the Covid-19 validation dataset described in Section 5.

		Predicted		
		Extra	**Intro**	Σ
Actual	**Extra**	37	21	58
	Intro	13	37	50
	Σ	50	58	108

Table 5: The confusion matrix of the LMT model on the Twitter data test set.

Error Analysis

The employed attention mechanism at least partially allows for the investigation of the algorithmic importance of single input tokens for the IPT Bi-LSTM classification task at hand. As Kain & Wallace (2019) point out, the distribution of attention weight mass does not necessarily correspond to the underlying theories of the task at hand. However, in earlier work, we have explored the attention weights of the proposed model in more depth and found them to be in line with implicit test theory (Johannßen and Biemann, 2019). With the limitations and the possibility of some explainability in mind, we present the attention weight mass during the training phase in Table 6. Those tokens with higher mass indeed appear to correspond with the psychological theory of introversion and extraversion. In those examples, calmness is rather associated with introversion and togetherness rather than extraversion.

verwenden use	erschaffen create	ruhe calm	arbeit work	vertieft being absorbed	intro
gemeinsam together	ideen ideas	nachbar neighbour	vertrauen trust	gedicht poem	extra

Table 6: Visualization of the attention weight mass per German token with corresponding translations during the training phase. Pre-processing steps were applied, e.g. stop-words removal (thus the choppy utterances). The tokens that received the highest mass do correspond with the psychological theory of extroversion vs. introversion (in this example calmness for introverts vs. togetherness for extraverts).

The errors made by the IPT Bi-LSTM attention model are displayed in Table 7. Very short and uncontextualized answers were more often mistaken by the model

and classified incorrectly. Furthermore, instances that require broader world knowledge (e.g. holding a rope being equivalent to team mountaineering) were misclassified.

Label	Text	Pred.
E	King kills; kills; drill in his hand	I
E	Hears his volleagues; to understand everything	I
I	Persons climbing; secures rope; in focus; reaction	E
I	sees landscape; holds rope; feels responsible	E

Table 7: Errors made by the Bi-LSTM attention model. Apparently, short answers and those that require broader world knowledge were difficult to model. The labels read E for Extraversion and I for Introversion.

The LMT Twitter model made similar mistakes as the IPT Bi-LSTM model, which indicates, that despite the data sources being different (IPT vs. tweets), there are overreaching linguistic challenges when attempting to model the task of classifying Jungian introversion and extraversion. Once again, short and noisy instances are prone to being misclassified, as well as those instances, which require world knowledge. This is in line with the findings from Stajner et al. (2021) on why the MBTI (including introversion and extraversion) is difficult to model.

8. Twitter LMT Model & LIWC categories

The most precise method of identifying individuals in need of support would either be self-reports or medical diagnoses made by trained physicians. Both information are sparse and those individuals with the most severe threat of mental suffering oftentimes do not self-report their struggling or visit facilities. With limited information, we aim to determine whether classifications of introversion and extraversion differentiate the observed tweets not only into those two psychological types, but also into groups that are challenged by the pandemic at different levels.

As described in Section 6, we utilize the psychological dictionary tool LIWC. Table 8 displays those results. Six LIWC categories were investigated that correspond to mental health and the social background (Pennebaker et al., 2007). Those are *inhibition positive feeling*, *insight*, *anxiety*, *sad*, *sex* and *eat*.

Table 8 is divided into three table paragraphs. The first displays tweets classified as introversion from 2019 compared with 2020. The second table paragraph displays tweets classified as extraversion, and the third table paragraph compares the whole instance data set without this introversion/extraversion differentiation in order to provide a comparison point (whether those changes are specific for either of the two psychological types or are present in the entire data set).

Even though we investigated the changes from 2019 compared with 2020 a confounding analysis showed differences in LIWC categories between extraversion and introversion in multiple categories, including those in Table 8, indicating an unrecognized explanatory variable.

		Inhibition	Positive feeling	Insight	Anxiety	Sad	Sex	Eat
Introversion	'19	.27	.20	1.35	.12	.34	.33	.13
	'20	.31	.21	1.71	.20	.28	.25	.09
	Δ	.04	.24	.36	.08	-.06	-.09	-.04
	%	12.4	3.7	22.1	40.3	-21.8	-35.0	-56.7
Extraversion	'19	.29	.21	1.54	.13	.31	.24	.13
	'20	.27	.27	1.57	.12	.37	.35	.17
	Δ	-.02	.06	-.03	-.01	-.07	.10	.04
	%	-7.1	24.2	3.0	-9.8	18.9	30.1	26.7
Control	'19	.28	.21	1.42	.12	.33	.30	.13
	'20	.29	.23	1.65	.16	.32	.29	.12
	Δ	.01	.04	.23	.04	.01	-.01	.01
	%	5.2	13.3	14.9	26.2	-2.7	-3.3	-8.6

Table 8: The first table paragraph displays psychological LIWC categories per instance with noticeable fluctuations from 2019 compared with 2020, which were classified as introversion. The displayed LIWC values represent the percentages of words of an instance (i.e. an answer) belonging to a category. In each case, the first row displays the LIWC category counts in 2019, the second in 2020, the third displays the absolute differences (Δ), and the fourth row displays the relative percentage difference. The second table paragraph displays the corresponding LIWC categories for extraversion predictions. A control investigation is displayed in the third and last table paragraph, where all instances from 2019 are compared with 2020 as a point of comparison of the change magnitudes.

Table 8 shows some fundamental differences between the groups of tweets classified as introverted and extraverted. Accordingly, *inhibition* declined for rose by 12%, whilst having increased by 7% for extraverts. While *positive feelings* barely changed for introverts, they increased by 24% for extraverted. Insight was greatly increased for introverts (+ 22%). The big difference occurs for anxiety, which sharply increased by 40% for individuals classified as introverts, whilst having declined roughly 10% for extraverted instances.

Noteworthy, *sad* did increase for extraverts (+19%), whilst having decreased for introverts (-22%). The category includes utterances such as crying, grief, or sadness. Instance examinations showed that instances high in *sadness* mostly read 'i miss you' or missing someone or something.

The social factors of *sex* and *eat* (being physical closeness and topics such as restaurants, dining, etc.) further differentiate those two groups by having decreased for introverts (-35% and -57%), whilst being increased in its frequency for instances classified as extraversion (+30% and +27%).

Needless to say, neither the attention weights, the binary classifications, nor the LIWC psychological categories can assert the individual's state of mind for certain. Nonetheless, they can serve as indicators. Following, we will discuss those findings, put them into relation to the pandemic, and will discuss the current research on

this topic from Section 2 with regard to those findings.

9. Discussion

As shown in Section 7, the proposed IPT Bi-LSTM model reaches comparably strong performances on the binary classification task between introversion and extraversion. The attention weights during training as displayed in Table 6 appear to be aligned with the theory of Jungian psychology types. For tweets, an in-domain LMT was trained.

The results in Table 8 add novel findings to the current discussion. Whilst introverts expressed fewer optimistic utterances, those worries did not increase for extraverts. Rather than that, negative emotions rose sharply for introverts, which can be interpreted as clear signs of worry. Anxiety generally increased but slightly more for introverts. Noteworthy, sadness increased for extraverts. But as single instance observations reveal, instances high in *sadness* mostly miss persons or e.g. restaurants. This direction of energy towards the outside suits extraversion and would explain this rather negative emotion being increased for extraverts. The last two observed LIWC categories with remarkable changes from 2019 compared with 2020 are of social relevance (*sex* and *eat*). Firstly, utterances associated with physical closeness are less frequent for introverts, whilst being by far more frequent for extraverts. Utterances associated with dining, eating, or visiting restaurants decreased for introverts, whilst being increased for extraverts. This, again, suits the understanding of Jungian extraversion (see Section 3).

Extraversion has been interpreted as sensitivity to positive affect and optimism, introversion, on the other hand, as lacking sensitivity to positive affect and pessimism (Watson and Clark, 1997; Watson and Tellegen, 1985). Positive affect (i.e. extraversion) is crucial in times of crisis to see the broader picture, cope with depressive thoughts and ruminations, and stay action-oriented. Introverts, which lack this disposition to experience positive affect tend to be "state-oriented" and even depressed, especially in times of crisis (Kuhl and Kazén, 1999). This could explain the higher frequencies of negative emotions in the tweets.

All of those characteristics are unfavorable during lockdowns or other inclined types of isolations and social distancing. Those findings are supported by current empirical research, such as conducted by Wei (2020), who also found introverts to be rather inclined to suffer during the pandemic.

10. Conclusion & Outlook

The Corona or COVID-19 pandemic can be described as an event of a century. Many governments have resorted to measurements of social distancing or lockdowns. Even though those measurements save lives and help to fight this menacing disease, it also burdens individuals. The aim of this work to build an NLP binary classifier of the Jungian psychology types of introverts and extraverts and investigate whether they react differently to those methods has been reached with comparably strong results. Even though the model showed strong results on the held-out test set, the Bi-LSTM model was not applicable for out-of-domain data from Twitter. Therefore, we crafted a second model on hand-labeled tweets. All data was made public.

Experiments on Twitter data from 2019 compared with 2020 differentiated by introverts and extraverts revealed that the mental suffering of introverts during the pandemic is comparably more severe, adding novel findings to the current and contradictory debate. Introverts show a higher frequency of utterances associated with isolation, showed less optimism, spoke less about social interactions, and showed more frequent anxiety utterances. Meanwhile, extraverts showed less frequent utterances of isolation and more frequent friendships. With our approach, we offer an approach to identify individuals, that show elevated signs of worry. With those findings, those individuals could be supported by mental health services. Furthermore, it underlines the necessity as a society to look out for those individuals, that have become especially retracted or express themselves with isolating language.

A future outlook, some indicators such as the confounding analysis, some already infrequent LIWC counting measures, and the rather weak introversion classification capabilities of the model should be taken into account for further critical analyzations. The findings in this paper should be viewed critically and examined with complementary experiments. Furthermore, we aim to deepen those findings and provide systems for automated personality detections, which then could help society to better overall mental health.

11. Ethical Consideration

Even though this research is intended to foster psychological diagnostic research and mental health, such work poses the problem of an ethical dilemma between risks and promises (Johannßen et al., 2020). NLPsych systems can be misused (dual use (Williams-Jones et al., 2014)), misunderstood (Luhmann system theory (Görke and Scholl, 2006)), and will contain severe biases, which are hard to detect due to data protection laws (Diehl et al., 2015).

The proposed classification approach can neither replace clinical examinations nor should it be used for anything else than the performed validation study: mass observations with in-domain data for research purposes and without the intention of diagnosing individuals. This, however, is not what this work intends to provide. Rather, we aimed to support psychologists with additional and evaluation objectivity tools and shed validating light on the effects of the pandemic. We believe this work to add insights into human well-being during the COVID-19 pandemic and hope to foster research for increased mental health, which is a result of a wide range of research findings.

12. References

Angleitner, A. (1991). Personality psychology: trends and developments. *European journal of personality*, 5(3).

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473 [cs, stat]*, 1409.

Balasa, A. (2020). COVID – 19 on Lockdown, Social Distancing and Flattening the Curve – A Review. *European Journal of Business and Management Research*, 5.

Bogner, K. and Landrock, U. (2016). *Response Biases in Standardised Surveys (Version 2.0)*. GESIS - Leibniz-Institut für Sozialwissenschaften.

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2004). The TIGER treebank. *Journal of Language and Computation*, 2:597–620.

Coppersmith, G., Leary, R., Crutchley, P., and Fine, A. (2018). Natural Language Processing of Social Media as Screening for Suicide Risk. *Biomedical Informatics Insights*, 10.

Diehl, C., Hunkler, C., and Kristen, C. (2015). *Ethnische Ungleichheiten im Bildungsverlauf: Mechanismen, Befunde, Debatten*. Springer VS.

Ezen-Can, A. (2020). A Comparison of LSTM and BERT for Small Corpus. *CoRR*, abs/2009.05451. arXiv: 2009.05451.

Fried, E. I. and Flake, J. K. (2018). Measurement Matters. *APS Observer*, 31(3).

Gerlitz, C. and Rieder, B. (2013). Mining One Percent of Twitter: Collections, Baselines, Sampling. *M/C Journal*, 16(2)(620).

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *The American Psychologist*, 48(1):26–34.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3483–3487, Miyazaki, Japan.

Gubler, D. A., Makowski, L. M., Troche, S. J., and Schlegel, K. (2020). Loneliness and Well-Being During the Covid-19 Pandemic: Associations with Personality and Emotion Regulation. *Journal of Happiness Studies*, 22(5):2323–2342.

Görke, A. and Scholl, A. (2006). Niklas Luhmann'S theory of social systems and journalism research. *Journalism Studies*, 7:644–655.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-term Memory. *Neural computation*, 9:1735–1780.

Hämmig, O. (2019). Health risks associated with social isolation in general and in young, middle and old age. *PLoS ONE*, 14(7).

Jain, S. and Wallace, B. C. (2019). Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, MN, USA.

Johannßen, D. and Biemann, C. (2020). Social media unrest prediction during the COVID-19 pandemic: Neural implicit motive pattern recognition as psychometric signs of severe crises. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 74–86, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Johannßen, D., Biemann, C., and Scheffer, D. (2019). Reviving a psychometric measure: Classification and prediction of the operant motive test. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 121–125, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Johannßen, D. and Biemann, C. (2018). Between the Lines: Machine Learning for Prediction of Psychological Traits - A Survey. In *Proceedings of the International Cross-Domain Conference*, pages 192–211, Hamburg, Germany.

Johannßen, D. and Biemann, C. (2019). Neural classification with attention assessment of the implicit-association test OMT and prediction of subsequent academic success. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019, Erlangen, Germany, October 9-11, 2019*.

Johannßen, D. and Biemann, C. (2020). Social Media Unrest Prediction during the COVID-19 Pandemic: Neural Implicit Motive Pattern Recognition as Psychometric Signs of Severe Crises. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 74–86, Barcelona, Spain (Online).

Johannßen, D., Biemann, C., and Scheffer, D. (2019). Reviving a psychometric measure: Classification of the Operant Motive Test. In *Proceedings of the Sixth Annual Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, pages 121–125, Minneapolis, MN, USA.

Johannßen, D., Biemann, C., and Scheffer, D. (2020). Ethical considerations of the GermEval20 Task 1. IQ assessment with natural language processing: Forbidden research or gain of knowledge? In *Proceedings of the GermEval 2020 Task 1 Workshop in conjunction with the 5th SwissText & 16th KONVENS Joint Conference 2020*, pages 30–44, Zurich, Switzerland (online).

Jung, C. G. (1921). *Psychologische Typen.* Zürich, Rascher.

Keh, S. S. and Cheng, I.-T. (2019). Myers-Briggs Personality Classification and Personality-Specific Language Generation Using Pre-trained Language Models. *ArXiv*, abs/1907.06333.

Kingma, D. P. and Ba, J. (2017). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*.

Kuhl, J. and Kazén, M. (1999). Volitional facilitation

of difficult intentions: Joint activation of intention memory and positive affect removes Stroop interference. *Journal of Experimental Psychology: General*, 128(3):382–399.

Kuhl, J. and Scheffer, D. (1999). *Der operante Multi-Motiv-Test (OMT): Manual [The operant multimotive-test (OMT): Manual]*. Impart, Osnabrück, Germany: University of Osnabrück.

Landwehr, N., Andrew Hall, M., and Frank, E. (2005). Logistic Model Trees. *Machine Learning*, 59(1):161–205.

Lester, H. and Howe, A. (2008). Depression in primary care: three key challenges. *Postgraduate Medical Journal*, 84(996):545–548.

Mahoney, M. J. (1984). Psychoanalysis and Behaviorism. In *Psychoanalytic Therapy and Behavior Therapy: Is Integration Possible?*, pages 303–325. Springer US, Boston, MA.

McHugh, M. (2012). Interrater reliability: The kappa statistic. *Biochemia medica: časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82.

Mehta, Y., Majumder, N., Gelbukh, A., and Cambria, E. (2019). Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, 53(4):2313–2339.

Myers, I. B., Kirby, L. K., and Myers, K. D. (2000). *Introduction to Type: A Guide to Understanding Your Results on the Myers-Briggs Type Indicator*. Oxford Psychologists Press.

Pennebaker, J., Francis, M. E., and John Booth, R. (1999). Linguistic inquiry and word count (LIWC). *Software manual*.

Pennebaker, J., Chung, C., Ireland, M., Gonzales, A., and Booth, R. (2007). The Development and Psychometric Properties of LIWC2007. *Software manual. http://liwc.wpengine.com*.

Plank, B. and Hovy, D. (2015a). Personality traits on Twitter—or—How to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98, Lisboa, Portugal, September. Association for Computational Linguistics.

Plank, B. and Hovy, D. (2015b). Personality Traits on Twitter—or—How to Get 1,500 Personality Tests in a Week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98, Lisboa, Portugal.

Rammstedt, B., Lechner, C. M., and Danner, D. (2018). Relationships between Personality and Cognitive Ability: A Facet-Level Analysis. *Journal of Intelligence*, 6(2):28.

Scheffer, D. and Manke, B. (2018). The significance of implicit personality systems and implicit testing: Perspectives from PSI theory. In *Why people do the things they do: Building on Julius Kuhl's contributions to the psychology of motivation and volition*, pages 281–300. Hogrefe Publishing, Boston, MA, US.

Schultheiss, O. and Brunstein, J. (2010). *Implicit Motives*. Oxford University Press, 1 edition.

Stajner, S. and Yenikent, S. (2021). Why Is MBTI Personality Detection from Texts a Difficult Task? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3580–3589, Online.

Watson, D. and Clark, L. A. (1997). Chapter 29 - Extraversion and Its Positive Emotional Core. In *Handbook of Personality Psychology*, pages 767–793. Academic Press, San Diego, CA, USA.

Watson, D. and Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin*, 98(2):219–235.

Wei, M. (2020). Social Distancing and Lockdown – An Introvert's Paradise? An Empirical Investigation on the Association Between Introversion and the Psychological Impact of COVID19-Related Circumstantial Changes. *Frontiers in Psychology*, 11.

Wijngaards, I., Sisouw de Zilwa, S. C. M., and Burger, M. J. (2020). Extraversion Moderates the Relationship Between the Stringency of COVID-19 Protective Measures and Depressive Symptoms. *Frontiers in Psychology*, 11.

Williams-Jones, B., Olivier, C., and Smith, E. (2014). Governing 'Dual-Use' Research in Canada: A Policy Review. *Science and Public Policy*, 41:76–93.

Wolf, M., Horn, A. B., Mehl, M. R., Haug, S., Pennebaker, J. W., and Kordy, H. (2008). Computergestützte quantitative Textanalyse: Äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count. *Diagnostica*, 54(2):85–98.

Zhou, Q. and Wu, H. (2018). NLP at IEST 2018: BiLSTM-Attention and LSTM-Attention via Soft Voting in Emotion Classification. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 189–194, Brussels, Belgium.

Zirikly, A., Resnik, P., Uzuner, Ö., and Hollingshead, K. (2019). CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, MN, USA.