Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)

Neural classification with attention assessment of the implicit-association test OMT and prediction of subsequent academic success

Dirk Johannßen
MIN Faculty,
Dept. of Computer Science
Universität Hamburg
& Nordakademie

Chris Biemann
MIN Faculty,
Dept. of Computer Science
Universität Hamburg
22527 Hamburg, Germany

http://lt.informatik.uni-hamburg.de/
{biemann, johannssen}@informatik.uni-hamburg.de

Abstract

Operant motives are unconscious intrinsic desires that can be measured by implicit methods, such as the Operant Motive Test (OMT) employs. During the OMT, participants are asked to write freely associated texts to provided questions and images. Trained psychologists label these textual answers with one of four motives. The identified motives allow for psychologists to predict behavior, long-term development, and subsequent success. We use a long short-term memory neural network (LSTM) combined with an attention mechanism for classification of OMT textual answers and show state-of-the-art performance over previous work. When investigating tokens that have high associated attention weights with the Linguistic Inquiry and Word Count (LIWC) tool, we find a weak connection between LIWC categories and the OMT theory. Lastly, we automatically annotate and count motives per participant and correlate counts with academic grades, finding a weak correlation between certain motives and subsequent academic success.

1 Introduction

The goal of our research is to classify psychometric textual data. Furthermore, we aim to investigate algorithmic decision making and validate automatic annotation by predictions in accordance with the psychometric theory. To pursue this goal, we perform multi-label classification on the Operant Motive Test (OMT, Section 2) with four labels. During this OMT, participants textually answer questions on images such as displayed in Figure 1 to provided questions.

Recent advances in artificial neural network architectures have established mechanisms that allow researchers to, in a limited fashion, inspect reasons for algorithmic decisions. One of these mechanisms is called *attention* and was found by Young et al. (2018) to be among the most broadly investigated and adopted elements of deep neural machine learning. We want to investigate access to algorithmic decision making by employing this attention mechanism (Section 3).

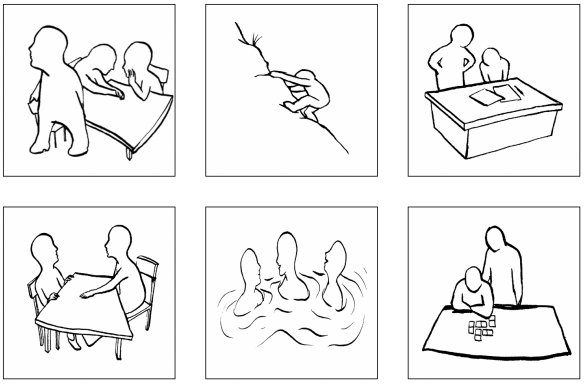

Figure 1: Some examples of images to be interpreted by participants utilized for the operant motive test (OMT). Exemplary answers given in Listing 1 correspond to the first picture. (Kuhl and Scheffer, 1999).

Lastly, the OMT theory states that some labeled motives allow for predictions of subsequent academic success, which we inspect by counting annotated labels and correlating these counts with participant's academic grades.

Even though there is a high demand for the automation of psychological textual data analysis (NLPsych), comparably little research has been performed on this interdisciplinary task (Johannßen and Biemann, 2018). Reasons for this circumstance include the lack of available labeled psychological text data, as Husseini Orabi et al. (2018) point out, and the mere difficulty of capturing psychological traits solemnly from texts, especially short texts. Since first, psychologists are skilled

workers for such a labeling task and secondly, said task is difficult, labeling such psychometric textual data is costly. Also, interpretability and transparency are crucial for gaining insights into the nature of some tasks including security, medicine, and psychology, which is often more valuable for researchers than reaching the highest classification performance scores (Zhang et al., 2018).

In this work, we focus on the following research questions: i) Do neural architectures outperform a previous non-neural machine learning approach and if so, which architectures perform how well? ii) Do the attention weights matter and reveal any insights into algorithmic decision making? iii) Is there a correlation between automatically predicted motives and subsequent academic success?

We describe the OMT in Section 2. Thereafter, we will discuss related work in Section 3. Section 4 describes the data basis of this work and its characteristics. Our research methodology will be described in Section 5. Results will be presented in Section 6. Finally, a conclusion will be drawn in Section 7.

2 Operant Motive Test

Implicit or operant motives are unconscious intrinsic desires, which can be measured by psychological implicit methods, which require participants to use introspection for the assessment of psychological attributes (Gawronski and De Houwer, 2014). During the testing procedure, participants are asked to write freely associated texts to provided questions and images. The OMT is such a test and emerged from the Thematic Apperception Test (TAT, (Murray, 1943)).

Listing 1 displays a few of the training instances that correspond to the first picture of Figure 1, which displays some examples out of several. Those images show one or multiple persons often in unclear scenarios and situations. Applicants are asked to answer four questions: i) What is important for the persons in this situation and what is s/he doing? ii) What is the person feeling? iii) Why does the person feel this way? iv) How does the story end? The four answers are concatenated to a single string. On this string, it is possible to annotate one of the three motives a) Affiliation (German 'Anbindung', letter A), b) Achievement (German 'Leistung', letter L) and c) Power (German 'Macht', letter M). The very first observed motive applies to the whole string, which is the so-called primacy rule (Kuhl and Scheffer, 1999). Once participants express a motive, this motive is saturated. Therefore, the following motives ought to be ignored when analyzing the answers. If no motive can be identified, a zero will be annotated (the so-called zero rule).

```
A sie nimmt am Gespräch nicht teil und
  wendet sich ab. gelangweilt. es
  interessiert sie nicht, worüber die
  anderen beiden reden. schlecht.
M weicht ängstlich zuruück. unterlegen.
  wird zurechtgewiesen.
  Gelegenheit den Fehler zu korrigieren
 ------- Translation -----
A she does not take part in the con-
  versation and turns away. bored.
  She does not care what the other
  two are talking about. Bad.
M withdraws anxiously. Inferior.
  is rebuked. Opportunity to
  correct the mistake.
```

Listing 1: German text examples of OMT answers with A being Affiliation and M being the power motive. The texts correspond to the first picture of Figure 1. Translations into English provided by the authors.

Implicit motives allow for the prediction of clinically measured non-verbal interpersonal communication such as the amount of smiling, laughing or eye contact (McAdams et al., 1984) as well as the job performance (Lang et al., 2012). Scheffer (2004) was able to show a significant ($p < 0.02$) multiple regression correlation with a negative beta slope (hence the lower the German grade, the better with 1 being *very good* and 5 *having failed*) between the achievement motive and z-standardized average grades of students from different departments.

3 Related Work

Previous approaches to predicting psychological traits. So far, approaches to psychological traits identification from texts often examined the connection between language and mental diseases. Current research mostly focuses on e.g. the detection of dementia (Masrani et al., 2017), crises (Demasi et al., 2019), suicide risks (Matero et al., 2019), mental illnesses (Zomick et al., 2019) or anxiety (Shen and Rudzicz, 2017) by the use of some form of natural language processing.

Nonetheless, some findings focus on motivation, success or characteristics. Tomasello (2002) describes the psychology of language as the method of focusing on the way people express themselves

rather than to focus on what meaning is conveyed.

Linguistic Inquiry and Word Count (LIWC) is a tool developed by Pennebaker et al. (1999) for text analysis, that utilizes previously validated categories containing word lists for which the membership ratio of an input sequence is being asserted. Furthermore, the tool calculates statistical values e.g. the average word length, the average count of word per sentence or the frequency of words longer than 6 characters. LIWC can be considered to be a standard tool for the analysis of texts from the psychological domain due to its broad utilization among researchers (Johannßen and Biemann, 2018). The German version of LIWC has been developed by Wolf et al. (2008).

So-called closed-class words are by far more informative than open-class words in terms of psychological language research. Closed-class words are words that tend to not change over centuries, which can be e.g. pronouns, prepositions or adverbs. Open-class words, on the other hand, are words that are strongly influenced by the time being, such as historical events or names. Pennebaker et al. (2014) found a link between the usage of closed-class words and academic success. During the study, which used the LIWC tool on written essays of college applicants and connected these to subsequent academic success, the authors showed that the rate of closed-class words are significantly ($p < 0.01$) positively correlated to subsequent academic success, regardless of the chosen essay topic or sought major.

In (Johannßen et al., 2019) we engineered handcrafted features to train a logistic model tree (LMT, Landwehr et al. (2005)) for classifying the operant motives. An LMT is a decision tree, which performs logistic regressions at its leaves. The LMT model reached an F-score of 80.1. The perplexities of language models for each motive, closed-class words, and ratios (words per sentence ratio, type-token ratio) were the main features for classification decisions.

Deep learning. Since assessing psychological traits solemnly from language is a challenging task, many researchers circumvent this bottleneck by including further personal information e.g. from social media platforms (Souri et al., 2018). Husseini Orabi et al. (2018) adapted this approach when they employed convolutional neural networks (CNN, LeCun et al. (1998)) and recurrent neural networks (RNN) in combination with further information from social media as labels such as average age, gender or posting frequency to enhance the detection of mental disorders.

In order to detect crises, Kshirsagar et al. (2017) combined neural and non-neural techniques. The data was obtained from the anonymous emotional support network Koko[1], which is available through multiple messaging applications.

A long short-term memory neural network (LSTM, (Hochreiter and Schmidhuber, 1997)) is a type of RNN which, in turn, is a deep neural network architecture, that allows for the neural cells to access other cells of the same recurrent layer with a time delay and thus develop a so-called memory. An LSTM furthermore employs memory cells that allow storing information of an arbitrary time horizon. Forget and update gates allow for these cells to purposely omit information and control, how the memory is altered. LSTMs have successfully solved the issues of vanishing or exploding gradients present in general RNNs (Hochreiter, 1998) and have been utilized for classifying short texts.

Lai et al. (2015) designed a recurrent convolutional neural network (RCNN) for text classification with promising results. An RCNN is an RNN with a max-pooling layer as its output. The main advantages of an RCNN in comparison with RNNs is the enhanced selection of targets or regions to have an impact on algorithmic decision making.

Young et al. (2018) found attention mechanisms as part of decoder-encoder-architectures to be amongst these recent advancements in their survey. Accordingly, attention mechanisms allow for decoders to assess their memory by referring back to their input sequence, which can enhance the network's performance. The idea of employing attention to a sequence-to-sequence (Seq2Seq) encoder-decoder system originated from Bahadanau et al. (2015).

With a sequence of annotations h_i being $(h_1, \ldots, h_{(T_x)})$, a context vector c_i represents the weighted sum of the annotations via:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \tag{1}$$

The weights α_{ij} are computed as:

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T_x} exp(e_{ik})} \tag{2}$$

[1]https://itskoko.com/

whilst $e_{ij} = a(s_{i-1}, h_j)$, with $a(\ldots)$ being a score function describing how well two words are aligned.

In other words, the system encodes an input sequence (this could be e.g. a certain language or a whole text to be summarized) into a context vector. This context vector together with hidden states functions as input for the attention mechanism, which computes attention weights and passes this context vector together with the attention weights on to the output layer. This process is illustrated in Figure 2.

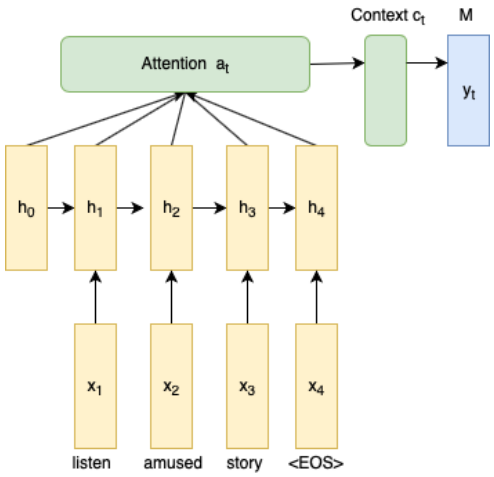

Figure 2: Illustration of the LSTM with attention mechanism. The LSTM receives hidden states and attention weights as inputs in order to output a corresponding context vector, which thereafter gets fed to a softmax output layer.

Attention mechanisms were successfully employed for various tasks. Gupta et al. (2018) utilized a CNN on group images for learning the global representation of the image and employed an attention mechanism for merging faces in order to learn local representations of only the faces, thus leading to a network capable of detecting emotions from entire groups of people. For this, the authors employed a Seq2Seq system with attention mechanism (the additional attention mechanism was proposed by Vaswani et al. (2017)). Images received automated descriptions by using a CNN encoder, an attention layer, and an LSTM decoder by Xu et al. (2015). Furthermore, the authors were able to project the attention weights onto the images, visualizing the gaze of the network. Speech has been analyzed for detecting emotions utilizing an attention mechanism by Ramet et al. (2018).

On textual data, attention mechanisms have enhanced the performance of classification and comprehension tasks. Hermann et al. (2015) advanced automated reading comprehension and question answering for texts with minimal prior knowledge. So-called self-attention was the enabler of semantic role labeling (SRL) for Tan et al. (2018). Self-attention is a special case of an attention mechanism, that only requires a single sequence to compute its representation. Vinyals et al. (2015) showed that a Seq2Seq model with attention mechanism could enhance syntactic constituency parsing to state-of-the-art performance.

A small subset of this data was annotated by utilizing attention over words. The authors were able to find the explanation of depressions from texts with a performance as well as human annotators had, which the authors refer to as *gold explanation*.

On the contrary, recent studies have questioned the interpretability of attention weights and suggested not to equate attention with explanation (Jain and Wallace, 2019). The authors found that if attention weights contribute to algorithmic decision making, the shuffling of these weights should significantly worsen results.

4 Data

The available data set has been collected and hand-labeled by researchers of the University of Trier. More than 14,600 volunteers participated in answering the OMT questions described in Section 2 to 15 provided images such as displayed in Figure 1. These participants produced 220,859 unique answers. Each answer was labeled by psychologists, which were trained with the OMT manual by Kuhl and Scheffer (1999). After pre-processing and cleaning the data, 209,716 text instances remain. The test and development set both constitute 10% of the available data, which is 20,960 instances each. The amount of motives in the available data is unbalanced with power (M) being by far the most frequent with 59%, achievement (L) constituting 19% of the data, affiliation (A) 17% and zero 5% (shown in Table 2 and in Figure 3). The pairwise annotator intraclass correlation was r = .85 on the Winter scale (Winter, 1994).

5 Methodology

Our methodology can be divided into two parts: the first is a natural language processing (NLP) task, which addresses research questions i) and ii)

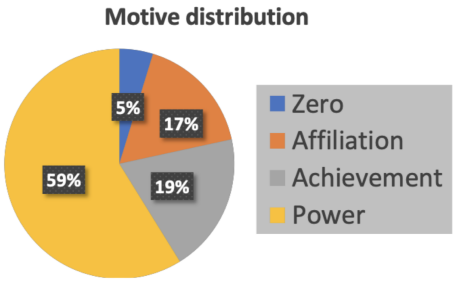

Figure 3: Graphical representation of the unevenly distributed motive labels amongst the data set.

and the second task answers research question iii) by counting classified motives per participant and correlating this count to academic grades.

In order to test whether an LSTM with an attention mechanism succeeds in outperforming the former best model for classifying the OMT, we employ the approach by Xu et al. (2015) on an already existing code basis for multiple text classifiers, which is utilized for further benchmarks as well.[2]

As for the word representations, we employed pre-trained fastText word embeddings for German (Bojanowski et al., 2017), provided by the developers.[3] In contrast to Word2Vec word embeddings by Mikolov et al. (2013), fastText has the capability of representing tokens not included in the embedded words on the basis of character n-grams. The OMT data (described in Section 4) is noisy, has many spelling mistakes and would probably not sufficiently be represented by word-based embeddings.

5.1 Benchmarking systems

To our knowledge, psychometrics closely related to the TAT have not been classified with neural methods yet. The only classification on the OMT has been performed by utilizing an LMT model in our previous work (2019), which we compare to our neural approach. In order to put different architectures into perspective and to explore the relationship of our proposed LSTM system with attention mechanism, we performed multiple benchmarking experiments on the task of automatically assigning the four classes of operant motives described in Section 2 and thus aim to answer the second research question of how well other neural approaches perform in comparison.

For this, we employed the following neural architectures, as reviewed in Section 3: LSTM, CNN, RNN, RCNN, Bi-LSTM with self-attention, LSTM with attention and Seq2One (a Seq2Seq variant with only one label as output) with attention. Since neural approaches are non-deterministic (Lai et al., 2015), we trained each model three times and averaged the F-scores for a stable assessment of results.

Three modifications of the LSTM with attention mechanism are employed: Firstly, we shuffled the attention weights before they got applied to the hidden states. Secondly, we reversed the direction of the input sequence to honor the OMT primacy rule. If this rule is followed and processing order has an influence, processing from right-to-left and classifying on the entire representation could improve results since the most influential signal (the first motive in the text) is accumulated last into the representation. Thirdly, we add comparable hand-crafted features as a fully connected input to the final classification softmax layer (e.g. part-of-speech (POS) tags, LIWC categories or the perplexities of trained language models per target motive), following Johannßen et al. (2019) to investigate in how far neural feature induction subsumes these features.

5.2 Psychometric predictions

After benchmarking, we utilize the most promising system for predictions in accordance with the OMT theory. 103 participating students answered the questions to 15 images, resulting in 1,545 answer sequences. Further, the data collection includes the grade of their bachelor's thesis, which was completed a few years after the OMT was taken. We employ the experimental design of our previous work (Johannßen et al., 2019) to ensure a fair comparison. For this, we predict the motives of each of the 15 answers given per participant, count the appearances per motive and correlate these to the bachelor's thesis grade.

5.3 Model training

All parameters of the models were tuned on a development set. Different fixed input sizes were considered for every architecture: Firstly we considered a fixed input length of 81 since the longest answer contains 81 words. Secondly, the average answer contains 20 words, which we considered as fixed input size in order to take the primacy rule (Section 2) into account. Shorter answers than

[2]https://github.com/prakashpandey9/Text-Classification-Pytorch/tree/master/

[3]Facebook's AI Research, https://fasttext.cc

the fixed input length receive the padding token (<pad>), longer ones were truncated. Human annotators are asked to ignore the rest of a sequence after a very first motive could be identified. Terms not observed in the training vocabulary were replaced by an out-of-vocab (OOV) token. Dropouts of 0.3, 0.5 and 0.8 were evaluated, whereas 0.5 has shown to perform best for the RNNs and has also been suggested by Hinton et al. (2012). The number of iterations was set to 3,600 in 32 batches and two epochs. The models received word embedded fastText inputs with 100 and 300 dimensions, of which the 300-dimensional embeddings reached better results, and had two hidden layers with 256 cells each. Learning rates were set to 0.0001, 0.001 and 0.01 for each model, with 0.001 performing best. All results are displayed in Table 1 and were achieved with these unified best-performing parameters.

As for the LSTM with attention mechanism, which has shown to perform best, the model converged quickly to a loss of approx. 0.4 and oscillates thereafter.

5.4 Attention weights assessment

As shown by Vaswani et al. (2017), the attention mechanism (described in Section 3) has broadly been believed to contribute to explainable artificial intelligence by shedding light on algorithmic decision making. Many authors have followed the initial idea and e.g. applied heat maps according to attention weights for input sequences and investigated algorithmic decision making. Other studies find contrary evidence that attention weights do not necessarily reflect true meaning (Jain and Wallace, 2019). Even though we are aware of these controversies and limitations, we follow the critic's suggestion to investigate whether attention weights make a difference in the performance of a system. For this, we measure on which index the most attention weight mass is accumulated. We hypothesized that this might often be the last token since attention weights usually traverse a sequence *in search* (metaphorically speaking) for suiting candidates and mostly does not find any of such, applying the most of the available attention weight to the last possible candidate – the last token. We will further collect sequences that do not show this behavior and thus have the largest attention weight mass assigned to other tokens than the last one. These tokens will be evaluated with the LIWC tool. We would expect the motives to be reflected in the LIWC categories if they meant anything at all. We automatically assembled all classified instances, whose highest attention weight did not assemble on the very last token, exceeded 0.3 and was classified correctly.

6 Results

6.1 Model performance

Table 1 shows classification performance of the different approaches on the test set. We were able to improve over our previous classifier (Johannßen et al., 2019). Even though neural approaches often perform better than earlier machine learning (Zhang et al., 2018), only the results of the best-performing model, the LSTM with an attention mechanism, outperforms the feature-engineered LMT classification model by an F-score of 81.55 (the LMT scored 81.10 and thus only slightly worse) with a fixed input size of 20 tokens. The same model with the fixed size of the longest answer of 81 tokens performed worse with an F-score of 80.71 (not shown in Table 1). The other approaches, also with a fixed input size of 20 tokens, performed worse, mostly around a 79 F-score except for the CNN. Including 129 hand-crafted features, reversing the reading direction and shuffling attention weights did not improve the results, thus indicating that firstly, attention matters, secondly, the direction of classification is not as important and thirdly, the LSTM attention model learns the features (POS, LIWC categories, perplexity) incidentally. The confusion matrix of the best-performing model is displayed in Table 2. The same LSTM with attention mechanism enriched by similar hand-crafted features does not improve results further, indicating that the information from these features is subsumed by the induced representations. The inversion of the input sequence resulted in lower scores, indicating that either the model cannot make use of seeing earlier tokens later to account for the primacy rule, or that the primacy rule has not been followed consequently during annotation. Shuffling of the attention weights worsens the results, indicating that these weights matter for the classification task.

6.2 Assessment of the attention weights

Table 1 shows that the LSTM with attention mechanism scored significantly lower when its attention weights were shuffled compared to the one with

Model	∅ Accuracy	∅ Precision	∅ Recall	∅ F-score	F σ
CNN	63.26	59.34	63.62	61.41	2.36
RNN	68.73	73.10	68.73	70.85	1.59
LSTM	77.84	78.05	77.84	77.92	0.65
Sequence to One (Seq2One) with attention	77.34	76.81	77.43	77.12	1.53
LSTM Attn with shuffled attention weights	79.03	78.05	79.03	78.54	0.13
RCNN	79.70	79.35	79.81	79.58	0.77
Bi-LSTM with self-attention	81.16	80.35	81.16	80.75	0.31
LSTM Attn with 129 addit. handcrafted features	80.85	79.86	80.86	80.35	1.23
LSTM Attn with a reversed direction	80.87	80.05	80.87	80.46	0.99
LSTM with an attention mechanism (LSTM Attn)	**81.94**	**81.15**	**81.96**	**81.55**	0.09
LMT with 129 handcrafted features (baseline)	81.56	80.90	81.60	81.10	0.00

Table 1: Performance comparison between the LMT and neural systems. All models classified with a fixed input size of 20 tokens. The only system overcoming the strong baseline of the feature-based LMT is an LSTM with attention mechanism. This system was also tested in reversed direction, with shuffled attention weights and with 129 additional handcrafted features, all of which performed worse than the best model. We averaged all scores (∅) from three trained models each, and provide the standard deviation across runs (σ).

		Predicted				
		0	**A**	**L**	**M**	**Σ**
		5%	**17%**	**19%**	**59%**	**100%**
Actual	**0**	283	102	150	478	1,013
	A	29	2,739	112	646	3,526
	L	90	91	3,079	872	4,132
	M	126	657	404	11,102	12,289
	Σ	528	3,589	3,745	13,098	20,960

Table 2: The relative motive amounts and confusion matrix of the best performing system (LSTM Attn).

properly trained attention and assigned weights. Jain and Wallace (2019) stated that this case had occurred only rarely in their experiments, but that if this circumstance holds true, they would assume that attention weights could be considered for interpretation and explanation.

We can observe that on average, 79.85% of the available attention weight mass was assigned to the very last token of each instance. It appears that the mechanism considered one token at a time from left to right and determines whether attention weight mass should be assigned to the token in question. If this is not the case, the attention weight mass is being kept and the successor token is considered. When the mechanism reaches the end of the sequence, it assigns whatever attention weight mass is left to the very last token. The second and third index with the highest following attention weight masses are the second last and third last tokens respectively. According to the OMT theory, the last tokens of a sequence, in general, should not provide the main information for encoding the whole sequence due to the primacy rule, this high attention weight mass on the last token indicates, that for the majority of classified instances, the attention weights do not serve as a widely applicable means to interpret the reasons for classification decisions in this setup.

Besides these last tokens, we aimed to investigate the mechanism further and compare these non-concluding tokens to all tokens by automatically assembling instances and attention weights.

Table 3 compares the four most prominent psychologically validated LIWC category memberships in percent per motive of all tokes versus non-final tokens with high attention weight masses. Most of the LIWC category names appear to be representative for the wordlists that they consist of. E.g. *positive emotion* consists of e.g. *love, nice and sweet*.

According to the OMT theory, people with a strong achievement motive desire intrinsic excellence. They tend to analyze problems thoroughly and focus on tasks. This description is reflected by *cognitive mechanism* that is almost twice as present for high attention mass tokens as it is for all tokens (27.39% compared to 14.11%). The categories *occuptation* (e.g. observe, conduct, advancing) with 24.66% and *achieve* – already with the same name as the OMT motive – with 23.28% are high in presence as well. Compared to rather low *social,*

	High attention weight mass			All tokens		
	LIWC	per cent	words	LIWC	per cent	words
Achievment	cognitive mechanism	27.39	intense concentrated motivated capabilities	social	15.17	- - - -
	occupation	24.66		cognitive mechanism	14.11	
	achieve	23.28		other references	11.44	
	insight	10.96		affect	10.49	
Affiliation	affect	12.12	important secure partner interested	social	19.76	- - - -
	positive emotion	12.12		other references	12.04	
	humans	9.09		affect	10.31	
	social	9.09		cognitive mechanism	9.48	
Power	affect	33.95	can feels dominant humiliated	social	18.99	- - - -
	cognitive mechanism	28.91		cognitive mechanism	11.46	
	positive emotion	24.93		other references	11.25	
	insight	20.16		affect	9.91	

Table 3: LIWC analysis of tokens that received the most attention weight mass on the left with all tokens on the right separated by predicted labels (left) versus manually annotated labels (right).

affect and other references, the OMT theory for the achievement motive appears to be better represented by tokens with high attention. Single words include *intense, concentrated, motivated and capabilities*.

Similarly, the LIWC categories for the affiliation motive are *affect, positive emotion, humans and social* for the left columns and apparently reflect the description of a desire to solve problems cooperatively, whilst avoiding conflicts. However, scores for LIWC categories are rather low at 12.12% and 9.09%. The social LIWC category is strongly present on the right column for all tokens with 19.76%, as well as *affect* with 12.04%. The other two LIWC categories of the right columns *other references* and *cognitive mechanism* do not appear to align well with the affiliation motive.

Even though the desire to influence and alter one's surrounding and fellow beings, the power motive can be identified by positive expressions as well as rather harsh ones. All LIWC categories of these columns on the left appear to align with the power motive, which are *affect (33.95%), cognitive mechanism (28.91%), positive emotion (24.93%) and insight (20.16%)*. The corresponding LIWC categories for all tokens on the right columns correspond with the exception of *other references* but are comparably weaker.

This comparison shows that tokens with high attention mass per motive correspond to the OMT theory e.g. occupation and insight for achievement, whilst all tokens do show some correspondence (e.g. social and affiliation), but in general, do not align well with the OMT theory. Interestingly, when removing the tokens (besides the last ones) that received the most attention weight mass and re-evaluating the answers with the LIWC tool to test the counterhypothesis that high-attention tokens do not reflect the classes, the categories shift to ones that do not correspond to the OMT theory.

gelangweilt *bored*	weil *because*	sie *she*	jeden *every*	tag *day*	0
geborgen *protected*	weil *because*	die *the*	andere *other*	person *person*	A
gefordert *challenged*	will *wants*	das *the*	ziel *goal*	erreichen *to reach*	L
zu *to*	maßregeln *disciplin*	dominant *dominant*	die *the*	andere *other*	M

Table 4: Heatmap according to the attention weights displayed on four example snippets of OMT answers in German with their glossed translations and targets (A for affiliation, M for power and L for achievement).

Examples are given in Table 4, which displays some tokens highlighted, according to the token's attention weight masses. These examples do not reflect the whole data basis but illustrate a possible aid for understanding the task at hand and might help develop tool support for this task or related psychometrics.

6.3 Correlation with bachelor's thesis grades

As described in Section 5, in order to analyze the predictive power of motives, we count predicted motives and correlate these counts to academic grades. While we previously found a weak correlation of $r = -0.2$ between power motive counts and the bachelor's thesis grade, the experiment in this work revealed a a correlation of $r = -0.25$ between the bachelor's thesis grade and the achievement motive in this work, i.e. the higher the achievement motive count, the better the German grade value (1 equals *good*, 5 equals *having failed*). The power motive is positively correlated with a small $r = 0.14$, i.e. the higher the power motive count, the worse the German grade. Figure 4 shows scatter plot displaying the counts of the power and achievement motives and the achieved bachelor's thesis grade.

This discrepancy of both model's predictions is anomalous. If both models performed comparably well on the same type of data, both mod-

els should reveal comparable correlations between counted motives and grades. The investigation of each model's motive predictions per student shows that the LSTM with attention mechanism often assigns the power motive but never zero, whilst the LMT model assigns zero on 17.76% of all cases, indicating that the LMT model often did not predict any motive. Thus, even though the models behave comparably well on test data of the same origin as the training data, they differ in their algorithmic decision making on data from a different origin.

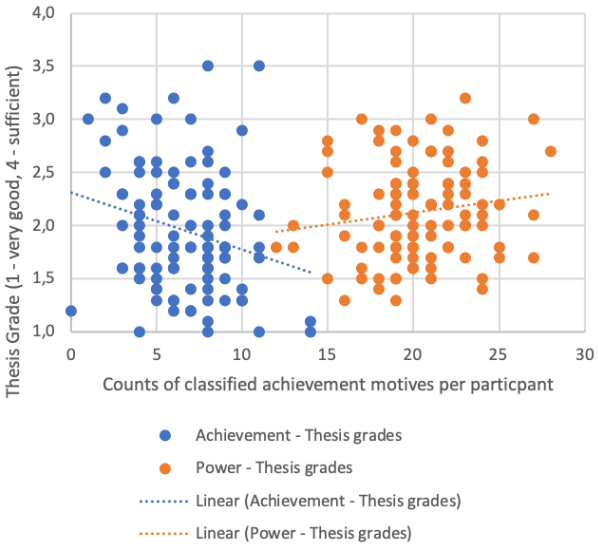

Figure 4: After predicting motives, the four motives per participants were counted. The power motive has the highest frequency. By counting predicted motives and correlating them to academic grades, a weak correlation of $r = -0.25$ could be observed between the achievement motive (blue dots) and the bachelor's thesis grade (in Germany, the best grade is 1, reading: the higher the achievement motive count, the better the grade). In contrast, the plots shows that the higher the power motive counts (orange dots), the worse the grade with $r = 0.14$.

7 Conclusion and outlook

We were able to outperform prior classification of the OMT by employing an LSTM with an attention mechanism achieving an F-score of 81.55 and thus can positively answer research question i), asking whether our proposed model could outperform our former approach. Other architectures such as the RNN, LSTM, Bi-LSTM or the RCNN mostly reached an F-score of approx. 79. Attention weights only matter in thus far that the shuffling of these weights worsens the results, asked by research question ii). The attention weight mass mostly accumulates on the very last token and thus does not allow for insights in the general case. For these cases where the attention weight mass was distributed among other tokens than the last one of a sequence, an analysis with the LIWC tool showed conformity of LIWC categories with the corresponding operant motives compared to these of all words. This indicates an overlap between the memberships per word of both linguistic assessments. This behavior of the highest attention mass on last tokens could be canceled out by employing a Bi-LSTM with attention mechanism and concatenating the attention weights of both systems, which we consider for future experiments. When removing these tokens and re-evaluating the sequence with the LIWC tool, the results shift, which has to be investigated further. Research question iii) questioned a correlation between identified motives and subsequent academic success as prior research has shown. This correlation could slightly be outperformed with $r = -0.25$ between the counted achievement motives and bachelor's thesis grade, which is a weak correlation much different to former predictions of the LMT model that assigned zeros more often than the LSTM model with attention mechanism. Since zero marks indecisiveness, it can be assumed that the LMT model does not generalize as well as the LSTM – though this assumption would have to be further examined by e.g. having trained psychologists assess the outputs of both models. Furthermore, direct predictions from language to grades could be investigated, hence losing information at the intermediate step of automatically annotated motives.

Nonetheless, further validation is appropriate due to recent debates upon attention weights as indicators of interpretation. One approach for validation would be to provide trained psychologists for labeling the OMT with tokens that received comparably much attention weight mass and with tokens that did not to measure how many cases would have been identified by said psychologists. Furthermore, we aim to provide annotators with a tool with attention-based highlighting for possibly saving time and expenses during the labeling process. Further numerical improvements could result from using contextualized embeddings, e.g. Bidirectional Encoder Representations from Transformers (BERT, Devlin et al. (2019)).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR*, San Diego, CA, USA.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Orianna Demasi, Marti A. Hearst, and Benjamin Recht. 2019. Towards augmenting crisis counselor training by improving message retrieval. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 1–11, Minneapolis, MN, USA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, MN, USA.

Bertram Gawronski and Jan De Houwer. 2014. Implicit measures in social and personality psychology. *Handbook of research methods in social and personality psychology*, 2:283–310.

Aarush Gupta, Dakshit Agrawal, Hardik Chauhan, Jose Dolz, and Marco Pedersoli. 2018. An attention model for group-level emotion recognition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, pages 611–615, New York, NY, USA.

Karl M. Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 1693–1701, Cambridge, MA, USA.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. *Neural computation*, 9(8):1735–1780.

Sepp Hochreiter. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of Twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97, New Orleans, LA, USA.

Sarthak Jain and Byron C. Wallace. 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.

Dirk Johannßen and Chris Biemann. 2018. Between the Lines: Machine Learning for Prediction of Psychological Traits - A Survey. In *International Cross-Domain Conference, CD-MAKE*, pages 192–211. Hamburg, Germany.

Dirk Johannßen, Chris Biemann, and David Scheffer. 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.

Rohan Kshirsagar, Robert Morris, and Samuel Bowman. 2017. Detecting and explaining crisis. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 66–73, Vancouver, BC, Canada.

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

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*, pages 2267–2273, Austin, TX, USA.

Niels Landwehr, Mark A. Hall, and Eibe Frank. 2005. Logistic Model Trees. *Machine Learning*, 59(1):161–205.

Jonas W. B. Lang, Ingo Zettler, Christian Ewen, and Ute R. Hülsheger. 2012. Implicit motives, explicit traits, and task and contextual performance at work. *The Journal of Applied Psychology*, 97(6):1201–1217.

Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Vaden Masrani, Gabriel Murray, Thalia Field, and Giuseppe Carenini. 2017. Detecting dementia through retrospective analysis of routine blog posts by bloggers with dementia. In *Proceedings of*

the 16th Workshop on Biomedical Natural Language Processing, pages 232–237, Vancouver, BC, Canada.

Matthew Matero, Akash Idnani, Youngseo Son, Sal Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath C. Guntuku, and H. Andrew Schwartz. 2019. Suicide risk assessment with multi-level dual-context language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44, Minneapolis, MN, USA.

Dan P. McAdams, R. Jeffrey Jackson, and Carol Kirshnit. 1984. Looking, laughing, and smiling in dyads as a function of intimacy motivation and reciprocity. *Journal of Personality*, 52(3):261–273.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Lake Tahoe, NV, USA.

Henry A. Murray. 1943. *Thematic Apperception Test*. Harvard University Press.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 1999. Linguistic inquiry and word count (LIWC). *Software manual. http://liwc.wpengine.com*.

James W. Pennebaker, Cindy K. Chung, Joey Frazee, Gary M. Lavergne, and David I. Beaver. 2014. When small words foretell academic success: The case of college admissions essays. *PLOS ONE*, 9(12):e115844.

Gaetan Ramet, Philip N. Garner, Michael Baeriswyl, and Alexandros Lazaridis. 2018. Context-aware attention mechanism for speech emotion recognition. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 126–131, Athens, Greece.

David Scheffer. 2004. *Implizite Motive: Entwicklung, Struktur und Messung [Implicit Motives: Development, Structure and Measurement]*. Hogrefe Verlag, Göttingen, Germany, 1st edition.

Judy H. Shen and Frank Rudzicz. 2017. Detecting anxiety on Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 58–65, Vancouver, BC, Canada.

Alireza Souri, Shafigheh Hosseinpour, and Amir M. Rahmani. 2018. Personality classification based on profiles of social networks' users and the five-factor model of personality. *Human-centric Computing and Information Sciences*, 8(1):24.

Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018. Deep semantic role labeling with self-attention. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 4929–4936, New Orleans, LA, USA.

Michael Tomasello. 2002. *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*. Psychology Press, 2nd edition.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Long Beach, CA, USA.

Oriol Vinyals, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey E. Hinton. 2015. Grammar as a foreign language. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 2773–2781, Cambridge, MA, USA.

David Winter. 1994. *Manual for scoring motive imagery in running text*. Dept. of Psychology, University of Michigan (unpublished).

Markus Wolf, Andrea B. Horn, Matthias R. Mehl, Severin Haug, James W. Pennebaker, and Hans Kordy. 2008. Computergestützte quantitative Textanalyse: Äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count. *Diagnostica*, 54(2):85–98.

Kelvin Xu, Jimmy L. Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2048–2057, Lille, France.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. *IEEE Computational Intelligence Magazine*, 13:55–75.

Xinyang Zhang, Ningfei Wang, Shouling Ji, Hua Shen, and Ting Wang. 2018. Interpretable deep learning under fire. *CoRR*, abs/1812.00891.

Jonathan Zomick, Sarah I. Levitan, and Mark Serper. 2019. Linguistic analysis of schizophrenia in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 74–83, Minneapolis, MN, USA.