



# Applying Attention-Based Models for Detecting Cognitive Processes and Mental Health Conditions

Esaú Villatoro-Tello<sup>1,2</sup> · Shantipriya Parida<sup>2</sup> · Sajit Kumar<sup>3</sup> · Petr Motlicek<sup>2</sup>

Received: 7 October 2020 / Accepted: 23 June 2021 / Published online: 17 July 2021  
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

## Abstract

According to the psychological literature, implicit motives allow for the characterization of behavior, subsequent success, and long-term development. Contrary to personality traits, implicit motives are often deemed to be rather stable personality characteristics. Normally, implicit motives are obtained by Operant Motives, unconscious intrinsic desires measured by the Operant Motive Test (OMT). The OMT test requires participants to write freely descriptions associated with a set of provided images and questions. In this work, we explore different recent machine learning techniques and various text representation techniques for facing the problem of the OMT classification task. We focused on advanced language representations (e.g. BERT, XLM, and DistilBERT) and deep Supervised Autoencoders for solving the OMT task. We performed an exhaustive analysis and compared their performance against fully connected neural networks and traditional support vector classifiers. Our comparative study highlights the importance of BERT which outperforms the traditional machine learning techniques by a relative improvement of 7.9%. In addition, we performed an analysis of how the BERT attention mechanism is being modified. Our findings indicate that the writing style features acquire higher importance at the moment of accurately identifying the different OMT categories. This is the first time that a study to determine the performance of different transformer-based architectures in the OMT task is performed. Similarly, our work propose, for the first time, using deep supervised autoencoders in the OMT classification task. Our experiments demonstrate that transformer-based methods exhibit the best empirical results, obtaining a relative improvement of 7.9% over the competitive baseline suggested as part of the GermEval 2020 challenge. Additionally, we show that features associated with the writing style are more important than content-based words. Some of these findings show strong connections to previously reported behavioral research on the implicit psychometrics theory.

**Keywords** Operant motive test · Psycholinguistics · BERT · Supervised autoencoder · Deep learning · Natural language processing

## Introduction

The idea that language use reveals information about personality has long circulated in the social and medical sciences. Previous research has demonstrated that the way people use words convey a great deal of information about themselves and their mental health conditions [1–4], including academic success [5]; however, much of the previous research has focused on the analysis of self-reports or essays. In contrast, implicit motives, which are indicators used by professional psychologists during the aptitude diagnosis, are not readily accessible to the conscious mind and, therefore, not detected using self-reports of personal needs, or through essay writing [6]. Instead, they are primarily assessed using indirect measures that rely on projective techniques that instruct

✉ Esaú Villatoro-Tello  
evillatoro@cua.uam.mx; esau.villatoro@idiap.ch

Shantipriya Parida  
shantipriya.parida@idiap.ch

Sajit Kumar  
kumar.sajit.sk@gmail.com

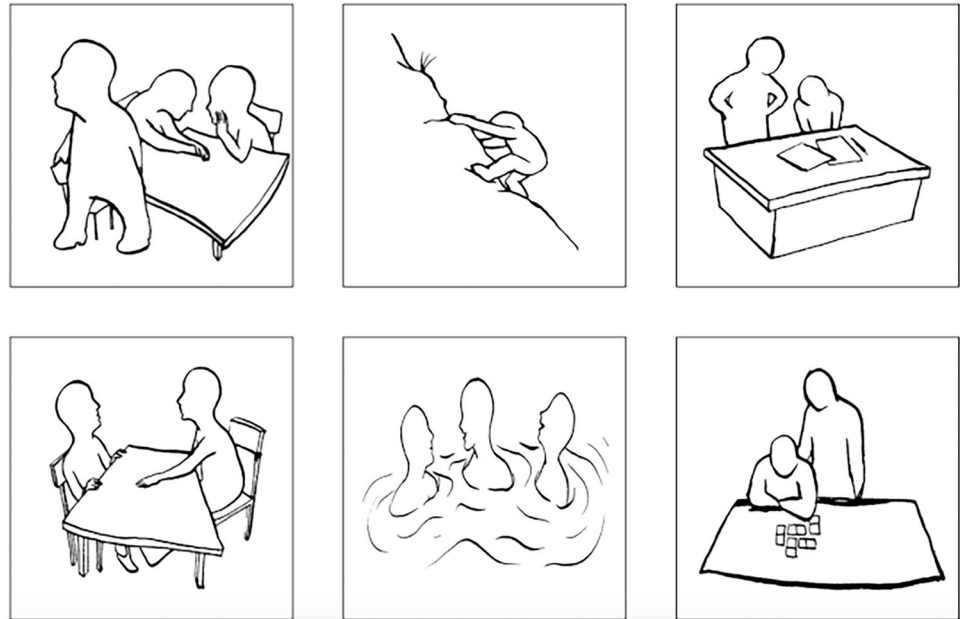
Petr Motlicek  
petr.motlicek@idiap.ch

<sup>1</sup> Universidad Autónoma Metropolitana, Unidad Cuajimalpa, Mexico City, Mexico

<sup>2</sup> Idiap Research Institute, Rue Marconi 19, 1920, Martigny, Switzerland

<sup>3</sup> GreatLearning, Bangalore, Karnataka, India

**Fig. 1** Sample images that are shown to subjects during the OMT test. Credits of the image correspond to organizers of the GermEval 2020 shared task<sup>2</sup>



individuals to produce imaginative stories based on ambiguous pictures stimuli that depict people in different situations (Examples of these stimuli pictures are shown in Fig. 1). These pictures aim to influence the content of the subjects' fantasy and to provoke that such fantasy is projected onto the characters through a short (textual) story (see Table 1 for some examples of the type of produced stories for the top-left image from Fig. 1). Consequently, this motivational response emerges through the contents of the written imaginative material and can be coded for its motive imagery using standardized and validated content coding systems.

The most frequently used measures of implicit motives are the Picture Story Exercise (PSE) [7], the Thematic Apperception Test (TAT) [8], the Multi-Motive Grid (MMG) [9], and the Operant Motive Test (OMT) [10, 11]. Generally speaking, these tests are based on the operant methods, i.e., participants are asked ambiguous questions or are shown simple images, which they have to describe. Specifically, the

OMT test is a projective instrument in which participants are presented ambiguous pictures (e.g., sketched scenarios) and ask to think of a story that describes what is happening in the picture. Thus, participants are asked to first pick the main protagonist, think of a story involving this person, and then answer the following three questions as spontaneously as possible: “*What is important to this person in this situation and what is the person doing?, How does the person feel?, and Why does the person feel that way?*” [11, 12]. Then, trained psychologists label these textual answers with one of five motives, namely M-power, A-affiliation, L-achievement, F-freedom, and 0-zero; and each motive is associated with its corresponding level or emotion (from 0 to 5), resulting in a total of 30 ( $5\text{-motives} \times 6\text{-levels}$ ) different OMT categories. In Table 2 we briefly describe the meaning of the operant motives; interested reader is referred to [11, 12].

Even though nowadays there is a huge demand for psychological data and its automated analysis, see for example, works presented at forums such as the CLPsych workshop<sup>1</sup>, until recently, little research has been performed on the Operant Motive classification task [13–18]. The primary reason is the lack of available labeled psychological text data, as [19] point out, and the difficulty of capturing psychological traits from text data, especially from very short texts.

Accordingly, in this paper, we aim to mitigate the lack of research on this issue and we explore how the recent advances in natural language processing (NLP) can be applied in the task of automatically identify psychological

**Table 1** Example answers produced to the stimuli image on the top-left side of Fig. 1

Motive	Sample text (Closest English translation)
A	sie nimmt am Gespräch nicht teil und wendet sich ab. gelangweilt. es interessiert sie nicht, worüber die anderen beiden reden. schlecht. (she does not take part in the conversation and turns away. bored. they don't care what the other two talk about. bad.)
M	weicht ängstlich zurück. unterlegen. wird zurechtgewiesen. Gelegenheit den Fehler zu korrigieren (withdraws fearfully. inferior. is rebuked. Opportunity to correct the mistake)

<sup>1</sup> Computational Linguistics and Clinical Psychology Workshop (<https://clpsych.org/>)

**Table 2** Brief outline of the motives (imagery types) [11, 12]

Type	Definition
Power (M)	Subjects concerned about having impact, control, or influence on another person, group, or the world at large.
Freedom (F)	Subjects concerned about stories that include the themes self-joy, increases in self-esteem through praise and attention, self-growth and self-actualization, integration of negative experiences into the self, rigid self-protection, or expressions of the fear of self-devaluation.
Affiliation (A)	Subjects concerned about establishing, maintaining, or restoring friendship or friendly relations among persons, groups, etc.
Achievement (L)	Subjects concerned about a standard of excellence.

traits from short textual data, specifically, we performed an exhaustive evaluation on the classification of the Operant Motives from the text. We evaluate the impact of very recent deep learning architectures such as transformers [20] (BERT, XLM, DistilBert), recent generalization techniques as supervised autoencoders [21], traditional classification methods, e.g., fully connected neural networks and support vector machines. To perform our experiments, we use the dataset provided during the “GermEval 2020 Task on the Classification and Regression of Cognitive and Emotional Style from Text” [22].<sup>2</sup>

The present paper represents an important extension of our previous reported participation [18] at the GermEval 2020 [22]. The main addition relies upon the adaptation and evaluation of recent generalization techniques, namely, deep supervised autoencoders (SAE). To the best of our knowledge, this is the very first time such an exhaustive evaluation of SAE techniques is performed on the OMT classification task. Additionally, we perform a thorough analysis of how the attention mechanism from the transformers architectures is affected when solving the OMT task.

In summary, the main contributions of this paper are as follows:

1. To the best of our knowledge, this paper represents the very first systematic exploration, and comparison, of several recent NLP technologies as well as machine learning techniques on the OMT classification task;
2. We propose a supervised autoencoder architecture for solving the OMT classification task, which as per our literature review, none of the recent research has applied this type of technology on the posed task. At the same

time, we evaluate the impact of different feature types, ranging from char ngrams to recent contextual embeddings, as inputs to the proposed SAE;

3. Finally, we conducted an analysis of how the attention mechanism of the transformer-based architectures is adapted during the OMT classification task. This type of analysis provides, to some extent, transparency on how the classifier is making its decisions. During this process, we observed some strong connections between our obtained results and psychometrics research.

The rest of the paper is organized as follows. Related work is discussed in “[Related Work](#)”. “[Dataset](#)” describes the dataset used in this work and its main characteristics. Details of our applied methodology are given in “[Methodology](#)”. The experimental setup, results, and analysis are provided in “[Results and Discussion](#)”. Finally, we share our main conclusions and future work directions in “[Conclusion](#)”.

## Related Work

Nowadays, there is an acknowledged necessity for digital solutions for addressing the burden of mental health diagnosis and treatment. It is recognized that won’t be possible to treat people by professionals alone, and even if possible, some people might require to use alternative modalities to receive mental health support [23]; such situation has become more evident with the current COVID-19 pandemic. Examples of recent efforts building technology toward this direction are the Woebot [24] and Wysa [25] dialog systems for health and therapy support for patients that have depression symptoms; Expressive Interviewing [26], which is a conversational agent aiming at support users to cope with COVID-19 issues.

The underlying hypothesis of most these works relies on the notion of the language as a powerful indicator about our personality, social, or emotional status, and mental health [3, 27]. Accordingly, the NLP community has focused on proposing several methods to identify different psychological traits from texts, and to examine the connection between language and mental health. As a few examples of this type of research, we can mention dementia identification [28, 29], depression detection [27, 30, 31], crisis counselling [32], suicide risks identification [31, 33, 34], mental illnesses classification [35, 36], anxiety detection [37], personality traits identification [38, 39], etc.

Although plenty of research has been done in the field of mental disorders detection and personality traits detection, there has been very little research for identifying motivation, success, or similar characteristics from psychological projective tests. One representative work, that brought back the discussion of how these traits could be automatically detected through traditional NLP techniques, is the research

<sup>2</sup> <https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/germeval-2020-cognitive-motive.html>

reported in [16]. Authors performed a process of features engineering to train a logistic model tree (LMT) [40] to classify a reduced set of implicit motives (0, M, A, and L). An LMT is a decision tree, which performs logistic regression at its leaves. In their research, authors found that the perplexity of the language models for each motive, closed-class words, and ratios (words per sentence ratio, type/token ratio) were the most discriminating features during the classification process. In [15], authors proposed using a Long Short-Term Memory (LSTM) neural network combined with an attention mechanism for classifying OMT motives (0, M, A, and L) from text data. For their experiments, authors employed pre-trained German fastText word embeddings [41] to represent tokens in the OMT data. Authors mention that when reviewing tokens that have high associated attention weights and compared with the Linguistic Inquiry and Word Count (LIWC) tool [42], they found a weak connection between LIWC categories and the OMT theory.

More recently, during the 5th SwissText & 16th KONVENS Joint Conference 2020<sup>3</sup>, it was organized a shared task on the classification and regression of cognitive and motivational style from the text, GermEval 2020<sup>2</sup>. This represented the very first time a shared task for detecting and classifying OMT motives was organized, providing a huge dataset under which was carefully curated for this purpose [22]. An important characteristic of the provided dataset is the fact of including the F-motive, and the labeling of six emotions (or levels), representing the first time a dataset with these characteristics is ever released. Three different research teams participated during the shared task, showing as the main results, the pertinence of using recent Bidirectional Encoder Representations from Transformers (ie., BERT) for solving the posed task. The winning approach employed the pre-trained Digitale Bibliothek Munchener Digitalisierungszentrum (DBMDZ) German model, achieving an F-score of 70% in the prediction of motives, and levels [17]. Generally speaking, and based on the shared task submissions [22], it is possible to observe that the performance of systems using pre-trained BERT embeddings and attention-based models tend to perform better than linear models. We referred the

interested reader to the system description papers presented during the shared task [13, 17, 18].

Hence, and with the aim of providing a more detailed analysis of the performance of different Natural Language Processing techniques and Machine Learning approaches in the recently released GermEval 2020 dataset, this paper will positively impact future research done on the comprehension of the implicit motive theory and its automatic detection through recent techniques. Accordingly, in this paper, we present a substantial extension of our system description paper [18], presented during the GermEval 2020, with the following main differences: *i*) we introduce and perform a series of experiments using novel generalization techniques, namely deep supervised autoencoders; which have never been tested on the OMT task before; *ii*) we evaluate the performance of three different transformer-based architectures; and, *iii*) we perform an analysis on the attention mechanism of transformer-based technologies which helps to understand why this type of technologies are well suited for this particular problem, showing strong connections with previous findings from the psychometrics research field.

## Dataset

To perform our experiments, we employed the dataset available in the GermEval 2020 shared task on the “Classification and Regression of Cognitive and Motivational style from the text”.<sup>1</sup> The provided data, mostly written in standard German language, has been collected from around 14000 subjects that participated in the OMT test.<sup>4</sup> Each answer was manually labeled with the motives (0, A, L, M, F) and the levels (from 0 to 5). This annotation was performed by expert psychologists, trained by the OMT manual as described in [11]. The size of the data set is 209,000 texts, from which 167,200<sup>5</sup> are part of the training (*train*) partition, 20,900 are part of the development (*dev*), and partition and 20,900 for testing (*test*). Tables 3, 4, and 5 show the distribution of the instances across the different classes for the *train*, *dev*, and *test* partitions respectively.

**Table 3** Distribution of the *train* partition across the OMT’s motives and levels values

	0	1	2	3	4	5	Total:
0	7610	0	1	1	2	4	7,618
A	13	2838	9574	1362	7556	6798	28,141
M	21	9482	6870	10797	27175	14251	68,596
L	6	2396	12863	6289	7544	3747	32,845
F	8	1766	5539	4302	9059	9313	29,987
Total:	7,658	16,482	34,847	22,751	51,336	34,113	167,187

<sup>3</sup> <https://swisstext-and-konvens-2020.org/shared-tasks/>

<sup>4</sup> There are a few cases where answers were given in English or French.

<sup>5</sup> A total of 13 instances were removed from the training partition due to its lack of label, leaving 167,187 instances.

**Table 4** Distribution of the *dev* partition across the OMT's motives and levels values

	0	1	2	3	4	5	Total:
0	1004	0	0	0	0	0	1,004
A	0	386	1198	169	939	855	3,547
M	4	1203	886	1321	3335	1787	8,536
L	0	295	1531	833	951	493	4,103
F	0	200	647	541	1121	1201	3,710
Total:	1,008	2,084	4,262	2,864	6,346	4,336	20,900

As can be observed in tables (Tables 3, 4, and 5), the dataset is highly unbalanced, making the classification task more challenging. The majority of the instances ( $\approx 41\%$ ) are from the power motive (M), followed by the achievement (L) motive ( $\approx 20\%$ ). Regarding the levels, most of the instances are grouped among classes 4 ( $\approx 30.7\%$ ), 5 ( $\approx 20.4\%$ ), and 2 ( $\approx 20.84\%$ ). It is important to mention that the same distribution remains in the *dev* and *test* partitions (Tables 4 and 5). Table 6 shows some statistics of the GermEval 2020 dataset, for *train*, *dev*, and *test* partitions. We compute the average number of tokens, vocabulary, and lexical richness of each text in the dataset. Lexical richness (LR), also known as “type/token ratio” is a value that indicates how the terms from the vocabulary are used within a text. LR is defined as the ratio between the vocabulary size and the number of tokens from a text ( $LR = |V|/|T|$ ). Thus, a value close to 1 indicates a higher LR, which means vocabulary terms are used only once, while values near to 0 represent a higher number of tokens used more frequently (i.e., more repetitive).

Two main observations can be done at this point. On the one hand, notice that for the three partitions (i.e., *train*, *dev*, and *test*), textual descriptions are very short, on average 20 tokens with a vocabulary of 18 words, resulting in a very high LR (0.92). The high LR value means that very few words are repeated within each textual description, i.e., very few redundancies. On the other hand, globally speaking, the complete dataset has a low LR (0.08 for *train* and 0.13 for *dev* and *test*). Although these values are not directly comparable due to the size of each partition, they indicate, to

analysis helped us to envision the complexity and nature of the data. Finally, we measure the coverage ratio of the German word embeddings (EmbC)<sup>6</sup> into our dataset, resulting in a 68.26% of coverage for the *training* partition, 83.93% for the *dev* set, and 84.10% for the *test* set. Similar to the LR score, we can not directly compare EmbC results due to the size of each partition. However, it was expected not to have a 100% coverage due to the noise contained in the OMT data, e.g., the many spelling and grammar errors present. Nevertheless, it is relevant to highlight the low coverage in the training set ( $\sim 68\%$ ). As it will be explained in “Results Analysis”, this low coverage has an important impact on the experiments based on these word embeddings.

## Methodology

Figure 2 shows the general view of the applied methodology for performing our experiments in this paper. As shown, given a textual description, we extract different types of features: character n-grams, word n-grams, non-contextual word embeddings (FastText), and pre-trained contextual embeddings (BERT). Then, depending on the selected learning strategy, computed features are feed to a specific learning technique, e.g., a supervised autoencoder (SAE), a fully connected network (FC), or into a transformer-based architecture (ST); that is trained to detect motives and levels from text, i.e., the OMT task. It is important to mention that instead of facing the OMT task as a 30 class classification

**Table 5** Distribution of the *test* partition across the OMT's motives and levels values

	0	1	2	3	4	5	Total:
0	963	0	0	0	0	1	964
A	0	367	1182	164	936	888	3,537
M	0	1224	821	1383	3396	1766	8,590
L	0	315	1525	815	906	437	3,998
F	1	192	715	555	1109	1239	3,811
Total:	964	2,098	4,243	2,917	6,347	4,331	20,900

some extent, that information across texts is very repetitive, i.e., similar types of words are being used by tested subjects for describing different images, even though they belong to different classes (motives and levels). Overall, this initial

<sup>6</sup> For some of our experiments, we used word embeddings trained with FastText on 2 million German Wikipedia articles, available at: <https://www.spinningbytes.com/resources/wordembeddings/>

**Table 6** Statistics of the OMT dataset in terms of number of tokens, vocabulary size and lexical richness. The minimum length of the texts is 1 token, while the maximum length is 99, 90, and 96 tokens for *train*, *dev*, and *test* partitions, respectively. In all partitions, the 75% of the data has a length of 27 tokens

Partition	Metric	Average ( $\sigma$ )	Total
Training	Tokens	20.27 ( $\pm 12.08$ )	3,389,945
	Vocabulary	18.07 ( $\pm 9.82$ )	63,133
	LR	0.92 ( $\pm 0.08$ )	0.08
	EmbC		43,095 (68.26%)
Development	Tokens	20.38 ( $\pm 12.17$ )	425,880
	Vocabulary	18.17 ( $\pm 9.94$ )	19,571
	LR	0.92 ( $\pm 0.08$ )	0.13
	EmbC		16,426 (83.93%)
Test	Tokens	20.24 ( $\pm 12.01$ )	423,018
	Vocabulary	18.05 ( $\pm 9.76$ )	19,780
	LR	0.92 ( $\pm 0.08$ )	0.13
	EmbC		16,636 (84.10%)

problem, we split the problem into two separate classification tasks: motives (5 classes), and levels detection (6 classes). This decision was made in accordance with the operant motive (OMT) theory [11], which states that motives and levels are disjoint orthogonal and thus not directly connected. Thus, at the end of our methodology, we fuse the predicted labels in order to get the motive-level combination of the given instance. Notice the dashed lines that go from BERT to the FC and SAE; these lines represent a series of experiments done after fine-tuning BERT on the posed task, where the newly computed embeddings are used as

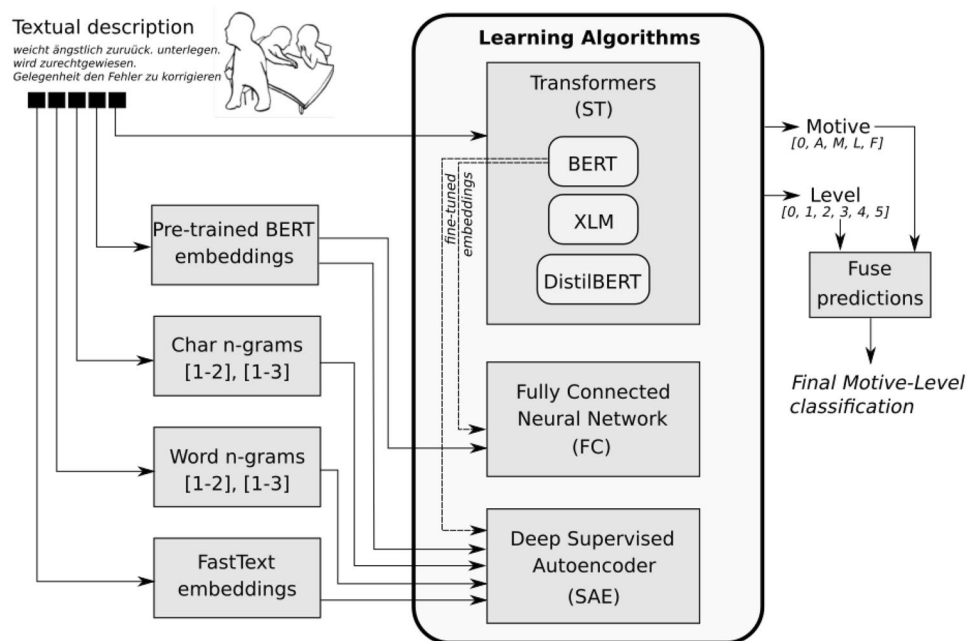
features to the FC and SAE learning techniques (see “Simple Transformer” to know the details of the fine-tuning process). Finally, the solid line that goes directly from the input textual description to the transformers block, it represents a series of experiments where the ST’s are configured as classifiers by adding a simple dense layer at the end.

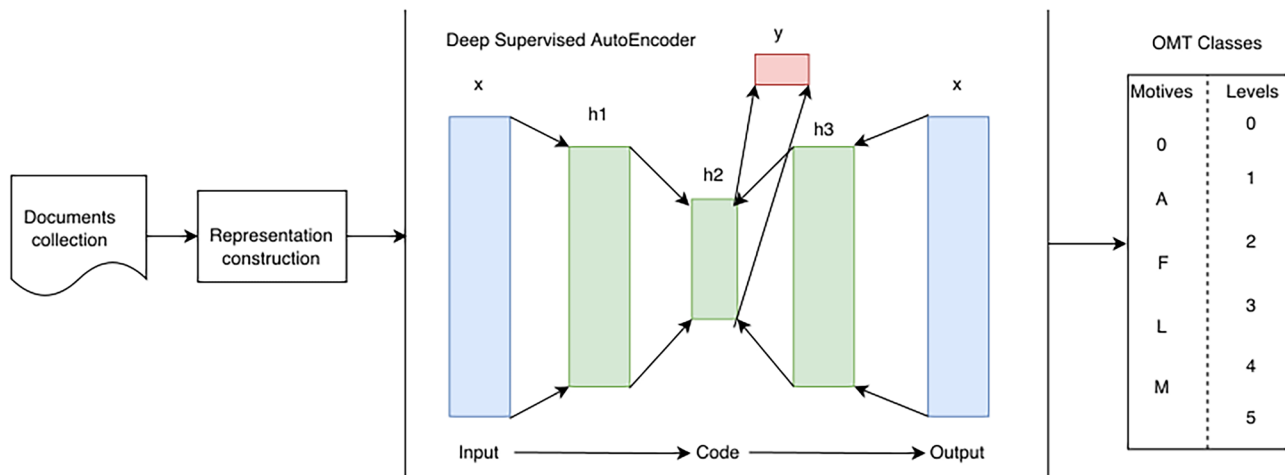
Following sections describe in detail the applied methodology for performing all our experiments. The proposed supervised autoencoder is detailed in “Supervised Autoencoders”. The description of the fine-tuning of transformer-based architectures is depicted in “Simple Transformer”, and the configuration of the traditional fully connected neural network is described in “Fully Connected Neural Network”. “Preprocessing” details the preprocessing operations to the dataset. “Evaluation Metrics” describe the considered evaluation metrics. An finally, “Baseline and Validation Approaches” defines the considered baseline and employed validation strategies (see Fig. 3).

### Supervised Autoencoders

An autoencoder (AE) is a neural network that learns a representation (encoding) of input data and then learns to reconstruct the original input from the learned representation. The autoencoder is mainly used for dimensionality reduction or feature extraction [43]. Normally, AE are used in an unsupervised learning fashion, meaning that we leverage the neural network for the task of representation learning. By learning to reconstruct the input, the AE extracts the underlying abstract attributes that facilitate accurate prediction of the input.

**Fig. 2** Proposed methodology. Given a textual description, we evaluated several text representations as input for different recent learning algorithms, including simple transformers, fully connected neural networks, and our proposed deep supervised autoencoder. In the end, predicted motive and level are combined to produce the final motive-level classification output





**Fig. 3** An example of a Supervised Autoencoder, where the supervised component ( $y$ -labels) is included. The input of the SAE is any type of pre-defined features computed over the document collection, e.g., character  $n$ -grams, word embeddings, or sentence encodings

Thus, a supervised autoencoder (SAE) is an autoencoder with the addition of a supervised loss on the representation layer (see Fig. 3). For the case of a single hidden layer, a supervised loss is added to the output layer and for a deeper autoencoder, the innermost (smallest) layer would have a supervised loss added to the bottleneck layer that is usually transferred to the supervised layer after training the autoencoder.

In supervised learning, the goal is to learn a function for a vector of inputs  $\mathbf{x} \in \mathbb{R}^d$  to predict a vector of targets  $\mathbf{y} \in \mathbb{R}^m$ . Consider a SAE with a single hidden layer of size  $k$ , and with weights for the first layer defined as  $\mathbf{F} \in \mathbb{R}^{k \times d}$ . The function is trained on a finite batch of independent and identically distributed (i.i.d.) data,  $(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_t, \mathbf{y}_t)$ , with the goal of a more accurate prediction on new samples generated from the same distribution. The weight for the output layer consists of weights  $\mathbf{W}_p \in \mathbb{R}^{m \times k}$  to predict  $\mathbf{y}$  and  $\mathbf{W}_r \in \mathbb{R}^{d \times k}$  to reconstruct  $\mathbf{x}$ . Let  $L_p$  be the supervised loss and  $L_r$  be the loss for the reconstruction error. In the case of regression, both losses might be represented by a squared error, resulting in the objective:

$$\frac{1}{t} \sum_{i=1}^t \left[ L_p(\mathbf{W}_p \mathbf{F} \mathbf{x}_i, \mathbf{y}_i) + L_r(\mathbf{W}_r \mathbf{F} \mathbf{x}_i, \mathbf{x}_i) \right] = \frac{1}{2t} \sum_{i=1}^t \left[ \|\mathbf{W}_p \mathbf{F} \mathbf{x}_i - \mathbf{y}_i\|_2^2 + \|\mathbf{W}_r \mathbf{F} \mathbf{x}_i - \mathbf{x}_i\|_2^2 \right] \quad (1)$$

The addition of supervised loss to the autoencoder loss function acts as regularizer and results (as shown in Eq. 1) in the learning of the better representation for the desired task [21]. Although SAE have been extensively evaluated on image classification tasks [21], its pertinence on thematic and non-thematic text classification tasks has not been extensively evaluated, being this an important contribution of this work.

Consequently, in order to perform a broad evaluation of this approach, we passed as input features to the SAE different types of text representation, namely pre-trained BERT encodings and also fine-tuned BERT encodings, in both cases using as representation the information extracted from the last hidden layer (LastHL), and the concatenation of the 4 last hidden layers (Concat4LHL).<sup>7</sup> Additionally, we also tested several traditional text representation techniques: word and char  $n$ -grams (with ranges 1–2 and 1–3). Finally, we also evaluate the performance of the SAE architecture using as a representation type non-contextual embeddings, in particular, we tried the German FastText embeddings trained on 2 million German Wikipedia articles.<sup>8</sup> All these variations can be observed at the bottom part of Fig. 2. For all our performed experiments, the overall configuration of the SAE model was done using nonlinear activation function (ReLU) with 3 hidden layers, the number of nodes in the representation layer was set to 300, and we trained to a maximum of 100 epochs.

### Simple Transformer

The transformer model [44] introduces an architecture that is solely based on attention mechanism and does not use any recurrent networks but yet produces results superior in quality to Seq2Seq [45] models, incorporating the advantage of addressing the long term dependency problem found in Seq2Seq model.

<sup>7</sup> More details on how these encodings are extracted in "Fully Connected Neural Network"

<sup>8</sup> <https://www.spinningbytes.com/resources/wordembeddings/>

For our experiments using simple transformers (ST) architectures, we setup three different state-of-the-art configurations:

1. Bert [46]: we use a pre-trained model referred as bert-base-german-cased, with 12-layer, 768-hidden, 12-heads, 110M parameters.<sup>9</sup> The model is pre-trained on German Wikipedia dump (6GB of raw text files), the OpenLegal-Data dump (2.4 GB), and news articles (3.6 GB). We refer to this configuration as ST-Bert in our experiments.
2. XLM [47]: for this configuration we use a model with 6-layer, 1024-hidden, 8-heads, which is an English-German model trained on the concatenation of English and German Wikipedia documents (bert-base-german-cased). We refer to this configuration as ST-XLM in our experiments.
3. DistilBert [48]: for this model we used a model with 6-layer, 768-hidden, 12-heads, 66M parameters (distilbert-base-german-cased). We refer to this configuration as ST-DistilBert in our experiments.

For all the previous configurations, in order to perform the fine-tuning of the ST architecture as a classifier, a simple dense layer with softmax activation is added on top of the final hidden state  $h$  of the first token [CLS], through a weight matrix  $W$ , and we predict the probability of label  $c$  the following way:

$$p(c|\mathbf{h}) = \text{softmax}(W\mathbf{h}) \quad (2)$$

Then, all weights, including the model's ones and  $W$ , are adapted, in order to maximize the log-probability of the correct label. The training is done using a Cross-Entropy loss function. To perform these experiments, we used the Simple Transformers library which allows us to easily implement this setup.<sup>10</sup>

As main configuration parameters, we set the max\_length parameter to 90,<sup>11</sup> and we re-trained the models up to two epochs. From here after, we refer as Bert-(FT) to the fine-tuned experiments. It is important to mention that the considered models, i.e., BERT, XLM, and DistilBERT, represent state-of-the-art language models available in the German Language. Although there are many other recent technologies, none of these are trained in German. Further details of employed models can be found at huggingface web page.<sup>12</sup>

<sup>9</sup> <https://deepset.ai>

<sup>10</sup> <https://pypi.org/project/simpletransformers>

<sup>11</sup> For deciding this value, we consider the information obtained in the data analysis described in "Dataset".

<sup>12</sup> [https://huggingface.co/transformers/pretrained\\_models.html](https://huggingface.co/transformers/pretrained_models.html)

**Table 7** Fully connected neural network configuration parameters. Notice that the number of input neurons depends on the representation type, while the number of output neurons is determined by the classification task, e.g., for the motive task, there are 5 output neurons

Hyper Parameter	Range
number of layers	3
number of hidden layers	1
nodes in hidden layer	16
activation function	ReLU

## Fully Connected Neural Network

As an additional classification method, we configured a fully connected neural network (FC). This type of artificial neural network is configured such that all the nodes, or neurons, in one layer, are connected to all neurons in the next layer. The topology of the employed network and its configuration parameters are mentioned in Table 7.

For the performed experiments using FCs, we passed as input features to the FC the sentence representation generated using BERT encodings. Thus, to generate the representation of the input text, we evaluate several configurations, namely: last hidden layer (LHL), concatenation of the 4 last hidden layers (Concat4LHL), min, max and mean pool of the last hidden layers. However, we only report the best performances obtained during the validation stage, i.e., LHL and Concat4LHL configurations. On the one hand, for generating the Concat4LHL representation we concatenate the last four layers values from the token [CLS]. As known, the [CLS] token at the beginning of the sentence is treated as the sentence representation. On the other hand, for the LHL configuration, we preserve as the text representation the values of the last hidden layer from the token [CLS].

For the reported experiments under the FC method (see Fig. 2), two configurations of BERT were tested for generating the LHL and Concat4LHL representation: i) pre-trained German encodings of BERT (distilbert-base-german-cased), referred as Bert(pre-trained); and ii) resultant fine-tuned BERT encodings from the re-training we explained in "Simple Transformer", referred as Bert(fine-tuned).

## Preprocessing

For performing our experiments, we perform different preprocessing operations depending on the selected representation type, i.e., word/char n-grams or contextual/non-contextual word embeddings (see Fig. 2).

In particular, for all the experiments performed with the pre-trained BERT embeddings, or in the fine-tuning process of the transformers architectures, we did not perform any type of preprocessing operation. Contrastly, when char/word n-grams or FastText embeddings are



employed, we apply the following preprocessing operations to the input text: 1) we remove all non-alphabetical symbols, e.g., numbers, strange symbols; 2) every word is lower cased. Other preprocessing techniques, like removing stopwords, punctuation or replacing German umlauts (ä, Ä, ö, Ö, ü, Ü) and ligatures (e.g., ß) were not applied as previous research indicates that no improvement is obtained from doing it [17].

Finally, it is also worth mentioning that we did not apply any type of spelling or grammar corrector. We decide not to do it given that such types of errors have shown to be important style markers in several tasks of authorship analysis [49].

## Evaluation Metrics

For measuring the overall effectiveness of the classification process, we use standard set-based evaluation measures, such as precision, recall and macro F-score. This decision was based on agreement with previous work and the official OMT classification task that reports and rank results with these metrics [22], specifically using the macro F1.

Generally speaking, when evaluating a classification task, there are four types of outcomes that occur:

1. True positives (TP) refer to the case when the classifier predicts an observation belongs to class  $c$  and it actually belongs to that class.
2. True negatives (TN) refers to the case when the classifier predicts an observation not belonging to class  $c$  and it actually does not belong to that class.
3. False positives (FP) occur when the classifier predicts an observation belongs to class  $c$  when in reality it does not.
4. False negatives (FN) occur when the classifier predicts an observation as not belonging to class  $c$  when in fact it does.

Thus, precision ( $P$ ) and recall ( $R$ ) are defined as shown in expression Eqs. 3 and 4 respectively.

$$P = \frac{TP}{TP + FP} \quad (3)$$

$$R = \frac{TP}{TP + FN} \quad (4)$$

The F-score (or F1), also known as the harmonic mean of  $P$  and  $R$ , is computed as follows:

$$F1 = 2 \times \frac{P \times R}{P + R} \quad (5)$$

Although the F1 score is a good metric to compare the performance of classifiers, the macro F1-score (F1-macro) is recommended to assess the quality of problems with multiple binary labels or multiple classes. Accordingly, the F1-macro is defined as the mean of the class-wise F1-scores (Eq. 6):

$$F1\text{-macro} = \frac{1}{N} \sum_{i=0}^N F1_i \quad (6)$$

where  $i$  is the class index and  $N$  the total number of classes. Notice that the F1-macro is not affected by the classes imbalance.

Although the Accuracy ( $Acc = \frac{TP}{TP+FP+TN+FN}$ ) score is a common metric to compare performance results, the Acc is not recommended in classification problems where there is a large class imbalance. In such particular scenario, it is very likely that a model tends to predict the value of the majority class for all predictions and achieve a high classification accuracy, however, this does not mean that such model is useful in the posed task, a phenomenon known as the accuracy paradox [50].

## Baseline and Validation Approaches

As a baseline, we replicated the approach proposed by the GermEval 2020 OMT task organizers, i.e., a linear Support Vector Classifier (SVC) using as a form of representation of the documents a traditional *tf-idf* strategy. Proposed baseline consists of a 30 (combined motive/level labels) binary SVCs (one-vs-all) classifiers.

In order to report robust and stable results, we implemented two different validation strategies. On the one hand, we performed a stratified  $k$  cross-fold validation strategy with  $k = 5$  using the entire dataset (*train+dev+test*); we refer to this configuration as ‘5CFV’ experiments. And, on the other hand, we report results on the *dev* and *test* partitions, which allows direct comparisons with the GermEval 2020 shared task participants.

## Results and Discussion

The results of each of the considered approaches (see Fig. 2) are reported in Tables 8 and 9 for the fivefold validation strategy and for the *dev* partition, respectively. Results are reported in terms of F1-macro, precision, and recall metric.

For the results reported in Table 8, we can observe that the proposed baseline is able to reach an F1-macro of 64.5%, even though this baseline faces the OMT task as a 30 class problem. A similar behavior is observed in Table 9, where the SVC baseline classifier yields good performance on the

**Table 8** Average performance ( $\mu$ ) obtained across the 5-cross-fold-validation strategy (5CFV); the number between parenthesis represents the standard deviation ( $\sigma$ )

Method	Configuration type	Configuration sub-type	F1-macro $\mu(\pm\sigma)$	Precision $\mu(\pm\sigma)$	Recall $\mu(\pm\sigma)$
ST <sup>†</sup>	Bert	base-german-cased	0.701 ( $\pm 0.003$ )	0.701 ( $\pm 0.003$ )	0.702 ( $\pm 0.004$ )
ST <sup>†</sup>	XLM	mlm-ende-1024	0.691 ( $\pm 0.002$ )	0.693 ( $\pm 0.003$ )	0.693 ( $\pm 0.002$ )
ST <sup>†</sup>	Distilbert	base-german-cased	0.695 ( $\pm 0.004$ )	0.696 ( $\pm 0.003$ )	0.696 ( $\pm 0.004$ )
FC	Bert	LHL	0.589 ( $\pm 0.002$ )	0.603 ( $\pm 0.004$ )	0.586 ( $\pm 0.003$ )
FC	Bert	Concat4LHL	0.610 ( $\pm 0.006$ )	0.621 ( $\pm 0.003$ )	0.608 ( $\pm 0.007$ )
FC <sup>†</sup>	Bert-(FT)	LHL	0.693 ( $\pm 0.002$ )	0.689 ( $\pm 0.002$ )	0.697 ( $\pm 0.003$ )
FC <sup>†</sup>	Bert-(FT)	Concat4LHL	0.687 ( $\pm 0.002$ )	0.682 ( $\pm 0.001$ )	0.693 ( $\pm 0.003$ )
SAE	char	n-grams(1-2)	0.556 ( $\pm 0.001$ )	0.555 ( $\pm 0.004$ )	0.569 ( $\pm 0.003$ )
SAE	char	n-grams(1-3)	0.628 ( $\pm 0.006$ )	0.624 ( $\pm 0.006$ )	0.634 ( $\pm 0.005$ )
SAE	word	n-grams(1-2)	0.630 ( $\pm 0.004$ )	0.628 ( $\pm 0.006$ )	0.634 ( $\pm 0.005$ )
SAE	word	n-grams(1-3)	0.634 ( $\pm 0.002$ )	0.632 ( $\pm 0.002$ )	0.638 ( $\pm 0.002$ )
SAE	FastText	German-Wiki	0.614 ( $\pm 0.004$ )	0.614 ( $\pm 0.006$ )	0.618 ( $\pm 0.003$ )
SAE	Bert	LHL	0.558 ( $\pm 0.010$ )	0.558 ( $\pm 0.011$ )	0.567 ( $\pm 0.008$ )
SAE	Bert	Concat4LHL	0.563 ( $\pm 0.011$ )	0.572 ( $\pm 0.009$ )	0.568 ( $\pm 0.010$ )
SAE <sup>†</sup>	Bert-(FT)	LHL	0.674 ( $\pm 0.020$ )	0.669 ( $\pm 0.021$ )	0.683 ( $\pm 0.021$ )
SAE <sup>†</sup>	Bert-(FT)	Concat4LHL	0.656 ( $\pm 0.007$ )	0.651 ( $\pm 0.007$ )	0.664 ( $\pm 0.008$ )
Baseline	SVM	tf-idf	0.646( $\pm 0.003$ )	0.650( $\pm 0.003$ )	0.644( $\pm 0.003$ )

Symbol ‘†’ indicates that the obtained performance is statistically significant according to the Wilcoxon test with  $p < 0.05$

*dev* partition (F1 = 63.9%). Hence, we can conclude that the proposed SVC represents a hard baseline, showing stable results in both validation strategies.

In general, we can observe that our proposed supervised autoencoder is not able to generalize well in comparison to ST and FC methods. Observe that while in the 5CFV

configuration the best result is achieved when fine-tuned encodings are used as text representation technique (Table 8, SAE(Bert-FT)) F1 = 67.4%; in the *dev* partition the best performance is obtained when input features are defined by word n-grams from range 1 to 2 (F1 = 63.4%). As known, word n-grams are useful to capture the identity of a word and

**Table 9** Obtained results on the *dev* partition of the OMT classification task

Method	Configuration type	Configuration sub-type	F1-macro	Precision	Recall
ST <sup>†</sup>	Bert	base-german-cased	0.698	0.699	0.699
ST <sup>†</sup>	XLM	mlm-ende-1024	0.688	0.689	0.690
ST <sup>†</sup>	Distilbert	base-german-cased	0.692	0.694	0.693
FC	Bert	LHL	0.589	0.601	0.585
FC	Bert	Concat4LHL	0.616	0.620	0.617
FC	Bert-(FT)	LHL	0.673	0.675	0.673
FC <sup>†</sup>	Bert-(FT)	Concat4LHL	0.675	0.673	0.678
SAE	char	n-grams(1-2)	0.554	0.556	0.564
SAE	char	n-grams(1-3)	0.629	0.624	0.636
SAE	word	n-grams(1-2)	0.634	0.631	0.639
SAE	word	n-grams(1-3)	0.630	0.627	0.635
SAE	FastText	German-Wiki	0.474	0.485	0.476
SAE	Bert	LHL	0.540	0.555	0.547
SAE	Bert	Concat4LHL	0.571	0.569	0.575
SAE	Bert-(FT)	LHL	0.633	0.631	0.636
SAE	Bert-(FT)	Concat4LHL	0.624	0.623	0.629
Baseline	SVM	tf-idf	0.639	0.644	0.638

Symbol ‘†’ indicates that the obtained performance is statistically significant according to the Wilcoxon test with  $p < 0.05$

its context. Thus, these results indicate, to some extent, that the SAE attempts to exploit this information when solving the classification task. A similar performance is obtained when character n-grams are used as input features, specifically n-grams of size 1-3. These results are also interesting, as they are aligned with previous research findings, demonstrating the relevance of character n-grams in different non-thematic classification tasks [51, 52]. Char n-grams are capable of providing an excellent trade-off between sparseness and word's identity, while at the same time they combine different types of information: punctuation, morphological makeup of a word, lexicon, and even context. As main observations of the SAE performance we can highlight that, using fine-tuned BERT encodings produced the best results under the 5CFV strategy (outperforming the proposed baseline), but word and character n-grams are not capable to improve the baseline performance. Similarly, for the experiments on the *dev* partition (Table 9) where even though the results obtained with the fine-tuned BERT encodings are similar to those obtained with word and char n-grams, none of these configurations were able to improve the SVC baseline (63.9%).

Regarding the performance of the FCs, the best performance is obtained when we use as features the fine-tune BERT encodings extracted from the last hidden layer (F1 = 69.3%) for the 5CFV experiments, and when the concatenation of the 4 last hidden layers is used (F1 = 67.5%), in the *dev* partition. In both cases, Tables 8 and 9, we can observe an important difference on the performance of the FC when pre-trained or fine-tuned encodings are used. Generally speaking, the impact of the fine-tuning allows a better performance of the neural networks (as expected), outperforming the proposed baseline in both cases.

Overall, based on these experiments (Tables 8 and 9), the best performance (in terms of classification F1 score) was obtained by a simple transformer using BERT embedding. The attention-based architecture was found effective in comparison to FC and SAE methods. Consequently, during GermEval 2020 competition, we submitted a subset of what we found were the most effective configurations. Table 10 shows the performance of our submitted systems.

As can be observed in Table 10, our best performing system was the simple transformer architecture using BERT encodings. Specifically, this was our configuration that obtained the second place during the GermEval 2020 competition [22]. As a reference, we put at the bottom of the table the performance of the baseline system, and the performance obtained by the first and second places. As expected, the SAE were not able to improve the baseline system. However, our configuration based on the fully connected network, using the fine-tuned BERT encodings was able to outperform the proposed baseline. It is worth mentioning that the winning approach during GermEval 2020 is based

**Table 10** Obtained results on the *test* partition of the OMT classification task. Performance results are reported as given by the GermEval 2020 organizers [22]

Method	Configuration type	Configuration sub-type	F1-macro
ST <sup>†</sup>	Bert	base-german-cased	0.698
ST <sup>†</sup>	XLM	mlm-ende-1024	0.686
ST <sup>†</sup>	Distilbert	base-german-cased	0.688
FC	Bert-(FT)	LHL	0.671
SAE	char	n-grams(1-3)	0.628
SAE	word	n-grams(1-2)	0.634
Baseline	SVM	tf-idf	0.644
1st place [17]	-	-	0.704
3rd place [13]	-	-	0.678

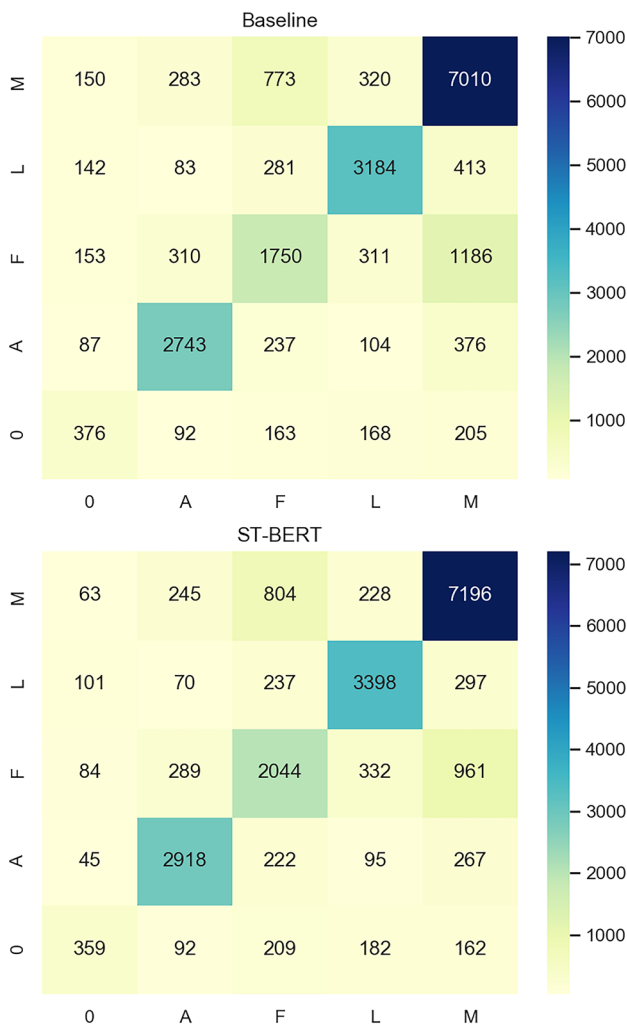
Symbol ‘†’ indicates that the obtained performance is statistically significant according to the Wilcoxon test with  $p < 0.05$

on a BERT methodology as well [17], with the pre-trained Digitale Bibliothek Münchener Digitalisierungszentrum (DBMDZ<sup>13</sup>) German model, validating the positive impact of transformer-based methods. Although the employed methodology by the winning approach [17] and our best configuration is the same, there are a few variations that are important. We consider as base model the BERT pre-trained on German Wikipedia, Open-Legal-Data and news articles. On the contrary, [17] used the DBMDZ<sup>13</sup> model. We did not apply any data correction process, while [17] made a data exploration to find all non-German texts, they applied an automatic translation process of all of these into German, and applied a spellchecker to correct spelling mistakes. Nevertheless, in spite of these extra effort, obtained results are very close for both techniques.

## Results Analysis

In this section, we present a more detailed analysis of the obtained results by our best configuration. Accordingly, Figs. 4 and 5 show detailed classification results between the SVC baseline, and our best configuration (ST-BERT). For ease of understanding we split the problem into detection of motives (Fig. 4) and detection of levels (Fig. 5). It can be observed that the ST architecture significantly increases the number of correctly classified instances in motives M (+2.5%), L (+6.2%), F (+14.38%), and A (+5.9%), however, this situation is not the same for motive zero (-4.5%). A similar situation occurs in the levels detection task (Fig. 5). For all the levels' categories, the ST is able to increase

<sup>13</sup> <https://github.com/dbmdz/berts>

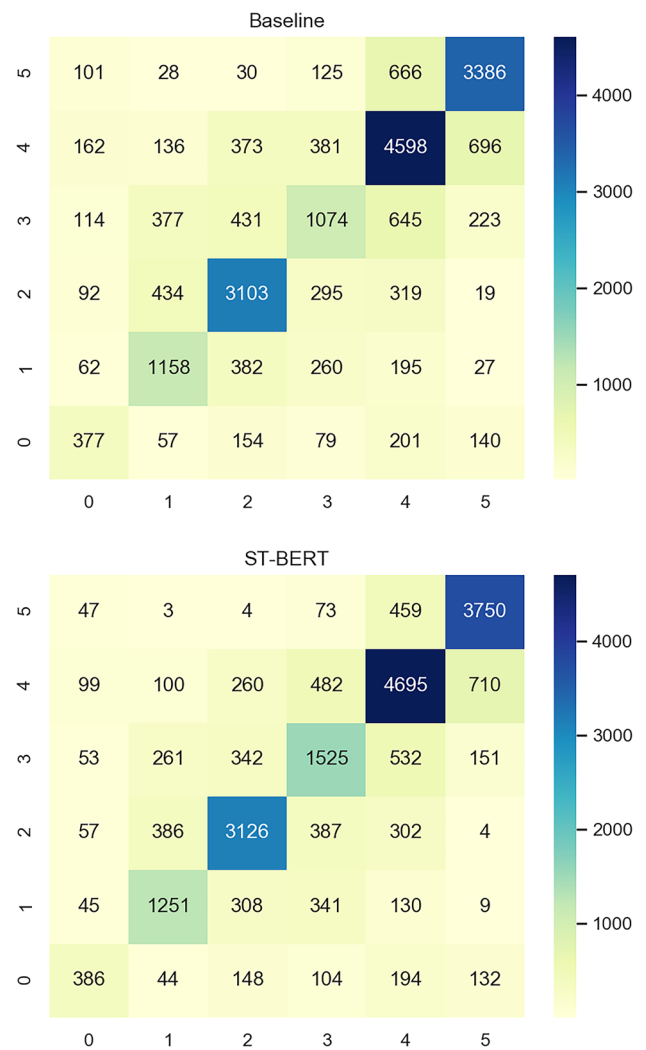


**Fig. 4** Confusion matrices for the MOTIVES classification task: top-baseline performance; bottom-ST performance

the number of correctly classified instances: 5 (+9.7%), 4 (+2.0%), 3 (+29.5%), 2 (+0.07%), 1 (+7.4%), 0 (+2.3%).

As mentioned in “[Methodology](#)”, instead of solving a 30 class problem, we split the OMT classification task in two separated problems, i.e., motives and levels classification problems. Thus, obtained results are aligned with the OMT theory [11], since according to our experiments it is possible to detect motives and levels separately, reinforcing the fact that motives and levels are not directly connected. Nevertheless, some of the methods presented during the GermEval 2020 did face the OMT problem as a 30 class classification task [22], which indicates that the OMT theory has to be revised and compared against what the NLP community has found.

In addition to the previous analysis, and given that a common concern is the lack of transparency of many deep learning architectures, we perform an analysis of what is the







**Fig. 5** Confusion matrices for the LEVELS classification task: top-baseline performance; bottom-ST performance

attention mechanism focusing on when solving the OMT task. The result of this analysis is shown in Table 12. The main intention of this type of analysis is to provide a better understanding of the connection between machine learning algorithms and language usage. To perform this analysis, we randomly select 5 sample texts produced by evaluated subjects in the *dev* partition. We show the results only for the distinct motives (A, F, L, M, 0). For a fair comparison, we selected textual samples belonging to the same level class, in this case, all samples belong to level 4.

To visualize the attention mechanism, we extract the attention weights based on [CLS] token from the last layer, average each token weight across all attention heads, and finally normalize weights across all tokens so that each weight is a value between 0.0 (very low attention) and 1.0

**Table 11** Followed criteria of highlight colors used in the visualization of the attention mechanism

Attention weights	Description	Highlight color
0.00 - 0.20	Very Low Attention	N/A
0.20 - 0.40	Low Attention	
0.40 - 0.60	Medium Attention	
0.60 - 0.80	High Attention	
0.80 - 1.00	Very High Attention	

(very high attention)<sup>14</sup>. The highlight criteria of the words are shown in Table 11.

Table 12 demonstrates the visualization of how the attention mechanism works in the Operant Motive Test classification task. Attention weights are extracted after the fine-tuning of the BERT method. As an important observation, notice the attention of functional words: *ist, und, an, der, das, zu, sie, sind, ein, einer, von (is, and, at, the, the, to, they, are, a, one, from)*. This indicates that for the simple transformer architecture the writing style is becoming more relevant at the moment of solving the classification task.

And additional observation is the attention paid to the word's ending. As known, during the tokenization process, unknown words are split into smaller tokens. When this is the case, the symbol '#' is added to the generated tokens. Especially for motive L, we can notice many cases where tokens with the symbol '#' are receiving the attention from the simple transformer architecture. Notice that negation words are resulting very important, e.g., *nicht*, as well as some punctuation marks, e.g., '!', '?'.

Furthermore, in Fig. 6, we show the usage given to the top 25 terms with higher attention values. For performing this analysis, we obtained the top most important words, i.e., words with higher attention values, for each motive category (A, F, L, M). Then, to obtain these 25 words, we intersected

the corresponding sets. The figure illustrates the relative frequency given to each of these words according to the motive class. As can be seen in the figure, subjects from different categories use these words with different frequency values. This frequency analysis also explains the good performance of the SVM classifier; which is based on a traditional *tf-idf* vectorial representation. However, even though frequency counts are helping the SVM to accurately separate among classes, the context in which these words appear is important, i.e., how users are employing these lexical units is relevant for solving the task.

Accordingly, in Figs. 7 and 8, we illustrate the context in which words *nicht (not)* and *sind (are)* are employed in our dataset. For this analysis, we took all the text generated by users from the same category (i.e., M, A, F, and L), and perform a collocation analysis fixing a target word (in this case: *nicht/sind*). As known, a collocation is a sequence of words that co-occur with high frequency within some corpus. Thus, for generating the visualization of each tree, we kept the most frequent collocations from each category. From this analysis, it is possible to observe that, even though these target words are frequently used by subjects, the employed contexts by each category are very different from each other. For example, subjects labeled with motive M (power) use the *nicht* words in an imperative/control fashion, e.g., *kommen nicht auf* (do not come up), while subjects categorized with A (affiliation) motive use it to show concern about others, e.g., *sie nicht alleine* (they are not alone). Similarly, for F (freedom) class, subjects use this negation to indicate concern about themselves, e.g., *möchte nicht mitbekommen* (don't want to notice); and in the L (achievement) motive, the common context denotes insecurity, e.g., *sie nicht weiss* (she doesn't know). Hence, the good performance of transformer-based NN architectures is explained by this analysis, as the attention mechanism of BERT is able to learn contextual relations between words (and sub-words) from the input text. Notice that these findings are aligned with previous psychometrics research (see Table 2).

This analysis provided interesting insights that we can summarize as: ST architecture pays higher attention to the use of personal pronouns, stop words, negation, punctuation marks, unknown words, and some conjugation styles, filtering out most of the unimportant elements such as content words. But not just isolated words, the context in which these words appear are providing important information to the transformer-based NN methods at the moment of detecting motives-levels. In other words, the writing style (*how we write*) is more relevant than content words (*what we write*) for solving the OMT classification task. Additionally, it is particularly interesting how the usage of negations words (*nicht*, and *##t* which correspond to “-n't” contractions) are frequently used by the

<sup>14</sup> We adapted the implementation from: <https://github.com/uzaymacar/comparatively-finetuning-bert/tree/f579ad55bf7afee5292f40a0943eb0ef018abe83>

**Table 12** Attention mechanism visualization for the OMT classification task

Motive	Sample text (closest English translation)
A	<p>sie braucht Verständnis und wendet sich an jemand der ihr zu ##hört und sie versteht. sie fühlt sich geborgen und angenommen und erzählt , was sie belastet. sie <b>ist</b> angenommen so wie sie <b>ist</b>.</p> <p>(she needs understanding and turns to someone who listens to and understands her; she feels safe and accepted and tells what is burdensome; she is accepted as she is.)</p>
F	<p>alles richtig gemacht zu haben und dem gegenüber Par ##oli bietet . trotz ##ig. weil <b>das</b> gegenüber <b>unein</b> ##sichtig <b>scheint</b>.</p> <p>(to have done everything correctly and to stand up to the other person, despite the fact that the other party seems unreasonable.)</p>
L	<p>sie kl ##etter ##t ein berg hoch, und darf <b>das</b> gleich ##gewicht nicht verlieren. ein bisschen in <b>siner</b> unangen ##hem ##ne position. sie dar ##chte nicht dass es so schwer wäre. sie sch ##af ##t es mit mut h ##ich zu kommen, und <b>ist</b> erstaun ##t <b>das</b> sie es <b>gesch</b> ##af ##t hat. glücklich <b>eine</b> neue seit ##e von ihr selbst zu finden.</p> <p>(she climbs a mountain and must not lose her balance. a little in an uncomfortable position. she does not think that it would be so difficult. she manages to do it with courage and is amazed that she did it. fortunately a new one to find her side.)</p>
M	<p>sie möchte <b>der</b> anderen Person zeigen, <b>dass</b> sie enttäuscht <b>ist</b>. enttäuscht über <b>die</b> andere Person. weil <b>die</b> andere Person <b>nicht</b> ihren Erwartungen stand gehalten hat.</p> <p>(she wants to show the other person that she is disappointed, disappointed with the other person, because the other person has not lived up to her expectations.)</p>
O	<p><b>die</b> Personen überlegen nach <b>siner</b> Lösung, verzwei ##felt und pa ##nisch. weil sie kurz vor dem ert ##rin ##ken <b>sind?</b>. gut . . . alle <b>drei</b> werden gerettet.</p> <p>(the people are thinking of a solution, desperate and panicked, because they are about to drown? ... good ... all three are saved.)</p>

power (M) and the freedom (F) motives. This finding is partially aligned with previously reported from the psychological theory [53], where it has been showed that so-called activity inhibition (AI) trait is mainly described as negations in combination with the power motive. Finally, our performed collocations analysis helped to understand and visualize the type of context that is helping the transformer-based NN architecture to solve the problem more accurately. Overall, these findings could foster implicit psychometrics theory, and consequently, advanced aptitude diagnostics supported by NLP technologies.

The aptitude test (i.e., OMT) is a type of psychological test that could affect the subjects' lives, specially if performed automatically without any human intervention [54]. Hence, there is an important urgency for understanding how this type of automatic decisions are being done by recent machine learning technologies. As stated in [55], explainable methods are becoming more relevant, particularly in the health-care domain. Thus, it is necessary to consider many aspects when designing explainable ML methods, e.g., *who is the domain expert?*, *who are the affected users?*, among others [56, 57]. Accordingly,

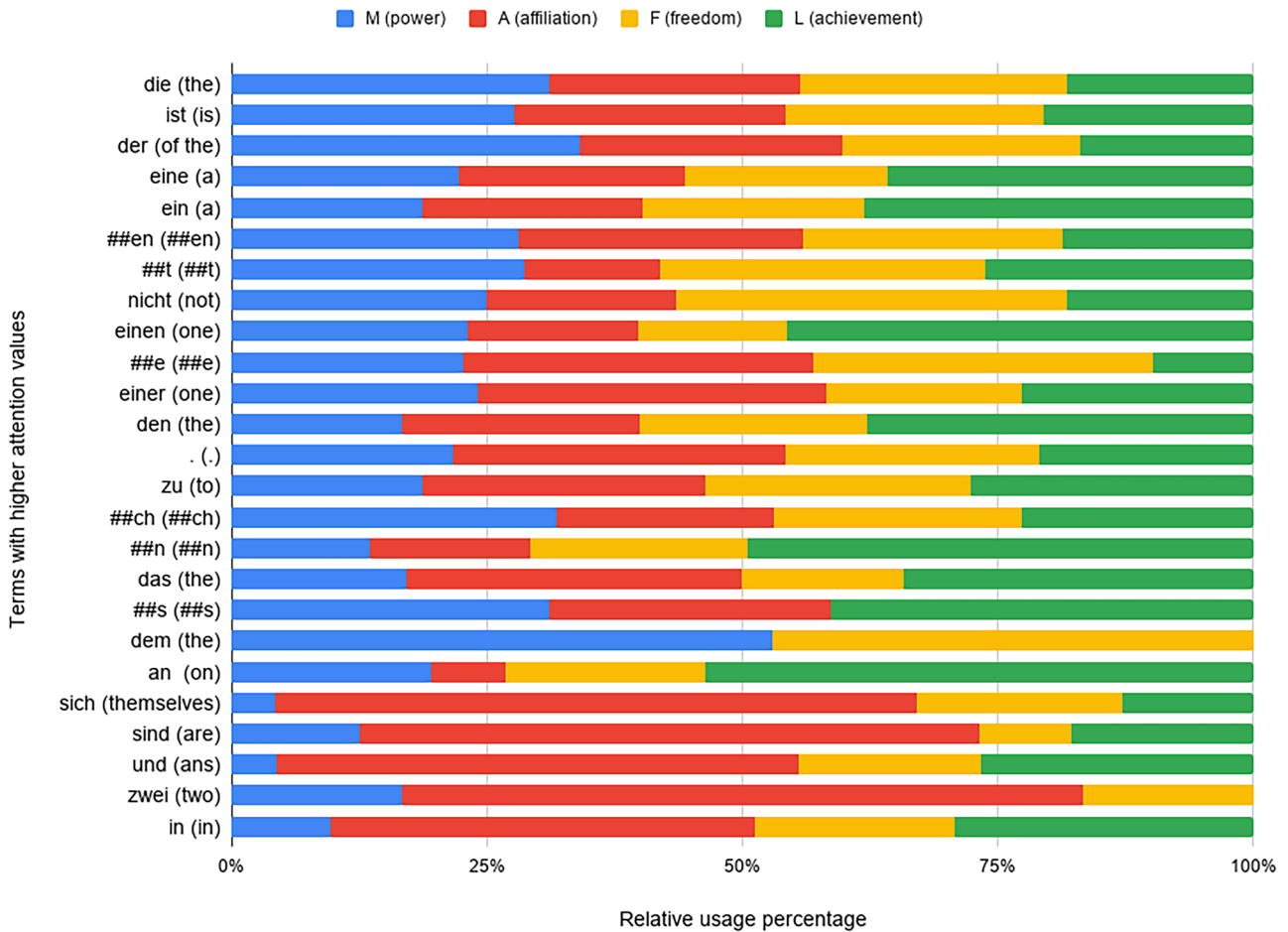


Fig. 6 Top 25 terms with higher attention across MOTIVES categories

and as part of our future work, we plan to extend our interpretability analysis towards the design of responsible artificial intelligence algorithms (i.e., explainable and

transparent) in the context of mental health automated analysis by applying some of the proposed recommendations in [56].

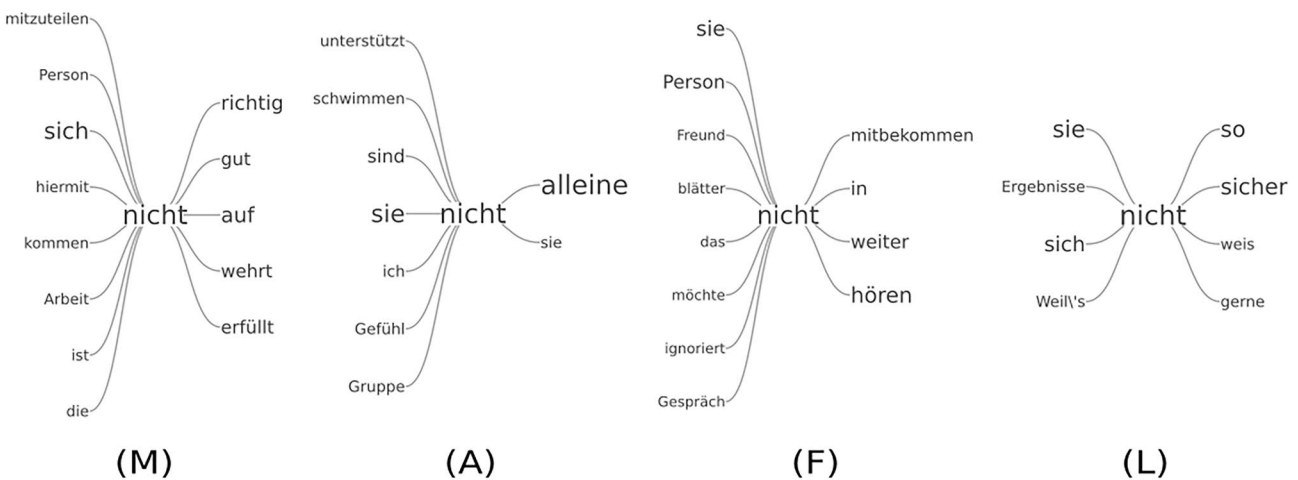
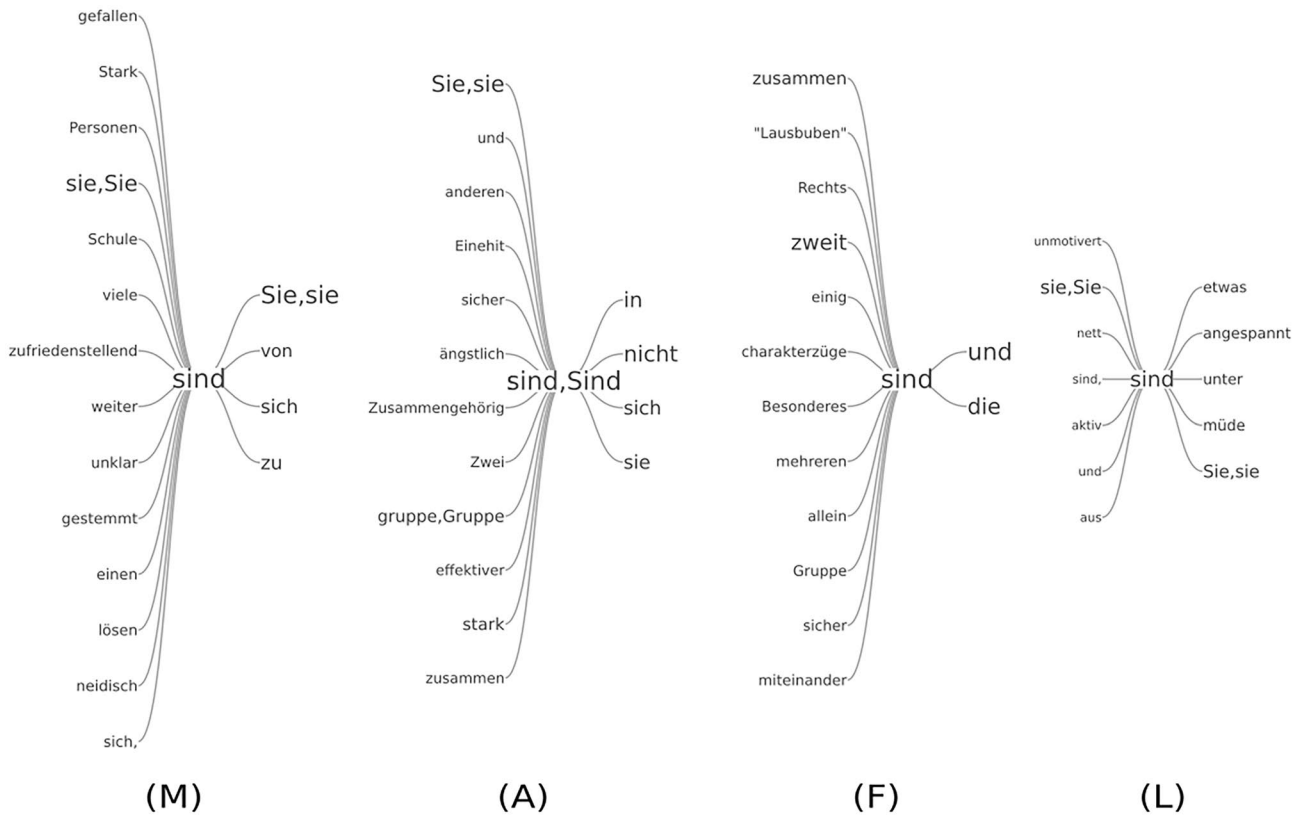


Fig. 7 Contextual tree of word *nicht*. Words in the leaves represent the most frequent words appearing next to the target word, in this case, *nicht*. Below each tree, its corresponding motive class is mentioned



**Fig. 8** Contextual tree of word *sind*. Words in the leaves represent the most frequent words appearing next to the target word, in this case, *sind*. Below each tree, its corresponding motive class is mentioned

**Conclusion**

This paper represents a first step towards the analysis of recent NLP technologies for solving the OMT classification task. To this end, we performed a comparative analysis among state-of-the-art simple transformer-based architectures, e.g., BERT, XLM, and DistilBert, very recent generalization techniques as supervised autoencoders and traditional machine learning techniques. Notably, transformer-based methods exhibit the best empirical results, obtaining a relative improvement of 7.9% over the baseline suggested as part of the GermEval 2020 challenge [22]. We performed an exploration on how the attention mechanism is working in this particular task, and obtained results revealed that features associated with the writing style are more important than content-based words. Some of these findings shown strong connections to behavioral research made on the implicit psychometrics theory. For example, as the result of our performed analysis, we observed that the usage of negations in combination with the power motive it is supported by the research made by [53]. As future work, we plan to evaluate the impact of hyperparameter tuning

through optimization methods, such as Bayes optimizer [58], evaluate the impact of early-fusion strategies in the performance of the SAE, and to perform further analysis on how the attention mechanism from the transformers architecture is working in the OMT task.

Finally, we would like to emphasize the importance of the ethical necessity of carefully understanding the research being done in the field of NLP & psychology. Although NLP technologies indicate that solving this type of tasks is, to some extent, possible, further research needs to be conducted to carefully explain the relation between psychological tests and subjects aptitudes. The authors would like to clearly state that we are against the use of this type of technology to discriminate against people in any type of our daily life situations. Even though we believe that this research is important, as can be useful for psychologists professionals, claiming that the NLP/ML community is able to accurately classify users according to their professional aptitudes and personality traits is not something we agree on. We support the idea that this type of research can help to validate previous theories as well as to support mental health care practitioners to evaluate or get important insights from closed and controlled studies.



**Acknowledgements** Esaú Villatoro-Tello was partially supported by Idiap Research Institute, SNI-CONACyT Mexico, and UAM-Cuajimalpa Mexico during the elaboration of this work. We would like to thank the GERMEVAL 2020 organizers, in particular, Dirk Johannßen for all the provided help.

## Declarations

**Ethical approval** This article does not contain any studies with human participants or animals performed by any of the authors. The authors declare that they have no conflict of interest. The authors do not see any significant ethical or privacy concerns that would prevent the processing of the data used in the study. All datasets containing personal data are processed in compliance with the GDPR and national law.

## References

- Burdick L, Mihalcea R, Boyd RL, Pennebaker JW. Analyzing connections between user attributes, images, and text. *Cogn Comput*. 2020;1–20.
- Pennebaker JW. The secret life of pronouns. *New Scientist*. 2011;211(2828):42–5.
- Pennebaker JW, Mehl MR, Niederhoffer KG. Psychological aspects of natural language use: Our words, our selves. *Annu Rev Psychol*. 2003;54(1):547–77.
- Reece AG, Reagan AJ, Lix KL, Dodds PS, Danforth CM, Langer EJ. Forecasting the onset and course of mental illness with twitter data. *Sci Rep*. 2017;7(1):1–11.
- Pennebaker JW, Chung CK, Frazee J, Lavergne GM, Beaver DI. When small words foretell academic success: The case of college admissions essays. *PLoS ONE*. 2014;9(12):e115844.
- Gawronski B, De Houwer J. Implicit measures in social and personality psychology. *Handbook of research methods in social and personality psychology*. 2014;2:283–310.
- McClelland DC, Koestner R, Weinberger J. How do self-attributed and implicit motives differ? *Psychol Rev*. 1989;96(4):690.
- Murray, H. *Thematic Apperception Test*. Harvard University Press. 1943. <https://books.google.co.in/books?id=r9a0CsRpEzYC>
- Sokolowski K, Schmalt H-D, Langens TA, Puca RM. Assessing achievement, affiliation, and power motives all at once: The multi-motive grid (mmg). *J Pers Assess*. 2000;74(1):126–45.
- Denzinger F, Brandstätter V. Stability of and changes in implicit motives. a narrative review of empirical studies. *Frontiers in psychology*. 2018;9:777.
- Kuhl J, Scheffer D. Der. operante multi-motiv-test (omt): Manual [the operant multi-motive-test (omt): Manual]. Germany: University of Osnabrück; 1999.
- Baum IR, Baumann N. Autonomous creativity: The implicit autonomy motive fosters creative production and innovative behavior at school. *Gifted and Talented International*. 2018;33(1–2):15–25.
- Çöltekin C. Predicting educational achievement using linear models. In *Proceedings of the GermEval 2020 Task 1 Workshop in conjunction with the 5th SwissText & 16th KONVENS Joint Conference*. 2020;23–29.
- Johannßen D, Biemann C. Between the Lines: Machine Learning for Prediction of Psychological Traits - A Survey. 192–211. [https://doi.org/10.1007/978-3-319-99740-7\\_13](https://doi.org/10.1007/978-3-319-99740-7_13), <https://hal.inria.fr/hal-02060047> Part 2: MAKE-Text.
- Johannßen D, Biemann C. 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)*: Long Papers (Erlangen, Germany, 2019), German Society for Computational Linguistics & Language Technology. 2019;68–78.
- Johannßen D, Biemann C, Scheffer D. Reviving a psychometric measure: Classification and prediction of the operant motive test. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. 2019;121–125.
- Schäfer H, Idrissi-Yaghir A, Schimanowski A, Bujotzek MR, Damm H, Nagel J, Friedrich CM. Predicting cognitive and motivational style from german text using multilingual transformer architectures. In *Proceedings of the GermEval 2020 Task 1 Workshop in conjunction with the 5th SwissText & 16th KONVENS Joint Conference*. 2020;17–22.
- Villatoro-Tello E, Parida S, Kumar S, Motliceck P, Zhan Q, Idiap & uam participation at germeval 2020: Classification and regression of cognitive and motivational style from text. In *Proceedings of the GermEval 2020 Task 1 Workshop in conjunction with the 5th SwissText & 16th KONVENS Joint Conference*. 2020;11–16.
- HusseinOrabi A, Buddhitha P, HusseinOrabi M, Inkpen D. Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (New Orleans, LA, June 2018)*, Assoc Comput Linguistics. 2018;88–97. <https://doi.org/10.18653/v1/W18-0609>, <https://www.aclweb.org/anthology/W18-0609>
- Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Brew J. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv:abs/1910.03771*. 2019.
- Le L, Patterson A, White M. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. In *Advances in Neural Information Processing Systems*. 2018;107–117.
- Johann D, Biemann C, Remus S, Baumann T, Scheffer D. GermEval 2020 task 1: Classification and regression of cognitive and motivational style from text. In *Proceedings of the GermEval 2020 Task 1 Workshop in conjunction with the 5th SwissText & 16th KONVENS Joint Conference*. 2020;1–10.
- Wykes T, Lipshitz J, Schueller SM. Towards the design of ethical standards related to digital mental health and all its applications. *Curr Treat Options Psychiatry*. 2019;6(3):232–42.
- Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*. 2017;4(2):e19.
- Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*. 2018;6(11):e12106.
- Welch C, Lahnala A, Perez-Rosas V, Shen S, Seraj S, An L, Resnicow K, Pennebaker J, Mihalcea R. Expressive interviewing: A conversational system for coping with COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020 (Online, Dec. 2020)*, Association for Computational Linguistics. 2020. <https://www.aclweb.org/anthology/2020>
- Tackman AM, Sbarra DA, Carey AL, Donnellan MB, Horn AB, Holtzman NS, Edwards TS, Pennebaker JW, Mehl MR. Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis. *J Pers Soc Psychol*. 2019;116(5):817.
- Masrani V, Murray G, Field T, Carenini G. Detecting dementia through retrospective analysis of routine blog posts by bloggers with dementia. In *BioNLP 2017 (Vancouver, Canada, Aug. 2017)*, Association for Computational Linguistics. 2017;232–237. <https://doi.org/10.18653/v1/W17-2329>, <https://www.aclweb.org/anthology/W17-2329>
- Szatloczki G, Hoffmann I, Vincze V, Kalman J, Pakaski M. Speaking in alzheimer’s disease, is that an early sign? importance of

- changes in language abilities in alzheimer's disease. *Front Aging Neurosci.* 2015;7:195.
30. Aragón ME, López-Monroy AP, González-Gurrola LC, Montes M. Detecting depression in social media using fine-grained emotions. 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)*. 2019;1481–1486.
  31. Losada DE, Crestani F, Parapar J. Erisk 2020: Self-harm and depression challenges. In *Advances in Information Retrieval (Cham, 2020)*, J.M. Jose, E.Yilmaz, J.Magalhães, P.Castells, N.Ferro, M.J. Silva, and F.Martins, Eds., Springer International Publishing. 2020;557–563.
  32. Demasi O, Hearst MA, Recht B. Towards augmenting crisis counselor training by improving message retrieval. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology (Minneapolis, Minnesota, June 2019)*, Association for Computational Linguistics. 2019;1–11. <https://doi.org/10.18653/v1/W19-3001>, <https://www.aclweb.org/anthology/W19-3001>
  33. Losada DE, Crestani F, Parapar J. Overview of erisk: Early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer. 2018;343–361.
  34. Matero M, Idnani A, Son Y, Giorgi S, Vu H, Zamani M, Limbachiya P, Guntuku SC, Schwartz HA. Suicide risk assessment with multi-level dual-context language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology (Minneapolis, Minnesota, June 2019)*, Association for Computational Linguistics. 2019;39–44. <https://doi.org/10.18653/v1/W19-3005>, <https://www.aclweb.org/anthology/W19-3005>
  35. Villatoro-Tello E, Dubagunta P, Fritsch J, Ramirez-de-la Rosa G, Motlicek P, Magimai-Doss M. Late fusion of the available lexicon and raw waveform-based acoustic modeling for depression and dementia recognition. In *INTERSPEECH 2021 (Brno, Czech Republic, August 2019)*. ISCA-International Speech Communication Association. 2021.
  36. Zomick J, Levitan SI, Serper M. Linguistic analysis of schizophrenia in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology (Minneapolis, Minnesota, June 2019)*, Association for Computational Linguistics. 2019;74–83. <https://www.aclweb.org/anthology/W19-3009>
  37. Shen JH, Rudzicz F. Detecting anxiety through Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality (Vancouver, BC, Aug. 2017)*, Association for Computational Linguistics. 2017;58–65. <https://doi.org/10.18653/v1/W17-3107>, <https://www.aclweb.org/anthology/W17-3107>
  38. Ramírez-de-la Rosa G, Villatoro-Tello E, Jiménez-Salazar H. Txp-i-u: A resource for personality identification of undergraduates. *J Intell Fuzzy Syst.* 2018;34(5):2991–3001.
  39. Souri A, Rahmani A, Hosseinpour S. Personality classification based on profiles of social networks users and the five-factor model of personality. *HCIS.* 2018;8:8–24.
  40. Landwehr N, Hall M, Frank E. Logistic model trees. *Machine Learning.* 2005;59:161–205. <https://doi.org/10.1007/s10994-005-0466-3>
  41. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. *Trans Assoc Comput Linguistics.* 2017;5:135–46.
  42. Pennebaker JW, Francis ME, Booth RJ. *Linguistic inquiry and word count: Liwc 2001*. Mahway: Lawrence Erlbaum Associates. 2001;71.
  43. Zhu Q, Zhang R. A classification supervised auto-encoder based on predefined evenly-distributed class centroids. 2019. arXiv:1910.00220.
  44. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In *Advances in neural information processing systems.* 2017;5998–6008.
  45. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems.* 2014;3104–3112.
  46. Devlin J, Chang MW, Lee K, Toutanova K. 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)*. 2019;4171–4186. <https://doi.org/10.18653/v1/N19-1423>, <https://www.aclweb.org/anthology/N19-1423>
  47. Conneau A, Lample G. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems.* 2019;7057–7067.
  48. Sanh V, Debut L, Chaumond J, Wolf T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. 2019. arXiv:1910.01108.
  49. Stamatatos E. A survey of modern authorship attribution methods. *J Am Soc Inform Sci Technol.* 2009;60(3):538–56.
  50. Valverde-Albacete FJ, Carrillode Albornoz J, Pelaez-Moreno C. A proposal for new evaluation metrics and result visualization technique for sentiment analysis tasks. In: P. Forner, H. Muller, R. Paredes, P. Rosso, *Attention Based Models for Detecting Cognitive Processes 31 B. Stein (eds.) Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pp. 41-52. Springer Berlin Heidelberg, Berlin, Heidelberg. 2013
  51. Sánchez-Vega F, Villatoro-Tello E, Montes-y Gómez M, Rosso P, Stamatatos E, Villaseñor-Pineda L. Paraphrase plagiarism identification with character-level features. *Pattern Anal Appl.* 2019;22(2):669–81.
  52. Sapkota U, Bethard S, Montes M, Solorio T. Not all character n-grams are created equal: A study in authorship attribution. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies.* 2015;93–102.
  53. Winter DG. The role of motivation, responsibility, and integrative complexity in crisis escalation: comparative studies of war and peace crises. *J Pers Soc Psychol.* 2007;92(5):920.
  54. Johannan D, Biemann C, Scheffer D. 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;30–44.
  55. Ieracitano C, Mammone N, Hussain A, Morabito FC. A novel explainable machine learning approach for eeg-based brain-computer interface systems. *Neural Comput Appl.* 2021;1–14.
  56. Barredo Arrieta A, Diaz-Rodriguez N, Del Ser J, Bannetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Info Fus.* 2020;58:82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>, <https://www.sciencedirect.com/science/article/pii/S1566253519308103>
  57. Holzinger A, Malle B, Saranti A, Pfeifer B. Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai. *Info Fus.* 2021;71:28–37. <https://doi.org/10.1016/j.inffus.2021.01.008>, <https://www.sciencedirect.com/science/article/pii/S1566253521000142>
  58. Snoek J, Larochelle H, Adams RP. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems.* 2012;2951–2959.