Using BERT for Qualitative Content Analysis in Psycho-Social Online Counseling

Philipp Grandeit, Carolyn Haberkern, Maximiliane Lang,
Jens Albrecht, Robert Lehmann
Nuremberg Institute of Technology Georg Simon Ohm, Nuremberg, Germany
{grandeitph64509, haberkernca76525, langma76539,
albrechtje, lehmannro}@th-nuernberg.de

Abstract

Qualitative content analysis is a systematic method commonly used in the social sciences to analyze textual data from interviews or online discussions. However, this method usually requires high expertise and manual effort because human coders need to read, interpret, and manually annotate text passages. This is especially true if the system of categories used for annotation is complex and semantically rich. Therefore, qualitative content analysis could benefit greatly from automated coding. In this work, we investigate the usage of machine learning-based text classification models for automatic coding in the area of psycho-social online counseling. We developed a system of over 50 categories to analyze counseling conversations, labeled over 10.000 text passages manually, and evaluated the performance of different machine learning-based classifiers against human coders.

1 Introduction

1.1 Psycho-Social Online Counseling

Online counseling has developed into a full-fledged psycho-social counseling service in Germany since the 1990s. Today, people can get advice on a wide variety of psycho-social topics in web forums and dedicated text-based counseling platforms. Online counseling is provided by psycho-social professionals who have received special training in this method. Similar to face-to-face psycho-social counseling, some aspects are known to make up high-quality online counseling, but there is few empirical evidence for special impact factors (Fukkink et al 2009, Dowling & Rickwood 2014).

Due to the complexity of the content, quantitative approaches have not been able to analyze the meaning and significance of methodical patterns in large numbers of consulting communications (Navarro et al. 2019). It is, however, possible to understand and describe the meaning of online counseling content with qualitative approaches (Bambling et al. 2008, Gatti et al. 2016).

This allows linking certain interventions of the counselors to the reactions of the clients on a case-by-case basis. But generalized statements on causal relationships are not possible with the small number of cases from qualitative studies (Ersahin & Hanley 2017).

An analysis of large numbers of counseling conversations using qualitative social research tools would help to better understand how successful online counseling works. Few related studies on these topics are available. Althoff et al. (2016) defined different models to measure general conversation strategies like adaptability, dealing with ambiguity, creativity, making progress or change in perspective and illustrated their applicability on a corpus of data from SMS counseling. Pérez-Rosas et al. (2019) analyzed the quality of consulting communications based on video recordings. Their automatic classifier used linguistic aspects of the content and could predict counseling quality with relatively good accuracy. However, neither of the mentioned approaches had the intention to recognize the meaning of individual phrases even though this deep understanding is crucial to eliminate weaknesses in the education of online counselors (Luitgaarden et al. 2016, Niuewboer et al. 2014). In addition, systems could be developed to provide online advisors with practical suggestions for improving their work.

1.2 Qualitative Content Analysis

Qualitative social research is a generic term for various research approaches. It attempts to gain a better understanding of people's social realities and to draw attention to recurring processes, patterns of

Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science, pages 11–23

https://doi.org/10.18653/v1/P17

interpretation, and structural characteristics (Kergel, 2018).

One such research approach deals with the content analysis of texts, the so-called qualitative content analysis according to Mayring (2015). It is a central source of scientific knowledge in qualitative social research. It tries to determine the subjective meaning of contents in texts. For this purpose, categories are formed based on known scientific theories on the topic and the discursive examination of the content. The definitions of those categories along with representative text passages are summarized in a codebook.

Then, human coders are coached in using the codebook. The coaching process and the implementation of the coding require high human expertise and manual effort because the coders must read, interpret, and annotate each text passage. Thus, qualitative studies can only be applied to a limited number of texts. Furthermore, it is hardly possible to define the categories so precisely that all coders find identical results, as human language is inherently ambiguous and its interpretation always partly subjective.

Machine learning could be a solution to the dilemma: If a trained model was able to categorize parts of the conversations according to a given codebook with similar accuracy as a human, the time-consuming text analysis could be automated.

1.3 Machine Learning for Qualitative Content Analysis

Previous studies have shown that supervised machine learning is generally suitable for qualitative content analysis (Crowston e.a. 2010, Scharkow 2013). However, these studies used only a few categories that could be distinguished relatively good, e.g. news categories like sports and business.

Online counseling, in contrast, is a complex domain. A detailed system of categories is necessary to identify impactful patterns in counseling conversations. Additionally, many categories such as “Empathy” or “Compassion” are quite similar in terms of the words used and can only be distinguished if the model is able to somehow "understand" the meaning of the texts.

Recent neural models have drastically outperformed previous approaches for sophisticated problems like sentiment analysis and emotion detection (Howard&Ruder 2018, Devlin e.a. 2018, Chatterjee e.a. 2019). We wanted to investigate if these models can be used for qualitative content analysis of online counseling conversations.

1.4 Research questions / Contribution

Our ***first research question*** is whether it is possible to train a model to identify psycho-social codes with a human-like precision. It also needs to be clarified whether a certain machine learning approach is particularly well suited for certain topics.

It is assumed that this training does not work equally well with all codes of the codebook. Therefore, ***the second question*** is which characteristics codes must have in order to be learned particularly well or particularly poorly.

In social science research, the discussion of different assessments of text passages is an important part of the scientific process. Therefore, the analysis of codes incorrectly assigned by a model is an important part of this work. The ***third research question*** is, therefore: What differences can be observed between the machine and human coding of text passages? If the deviations are plausible, they can be perceived as enriching the discursive process.

1.5 Methodology and Structure of the Paper

For the experimental evaluation, the social scientists in our interdisciplinary team created a codebook consisting of over 50 fine-grained categories and labeled over 10.000 text sequences of psycho-social counseling conversations (described in Section 2). The computer scientists then trained and evaluated a support-vector machine and different state-of-the-art models (e.g. ULMFit and BERT) on the provided data set (Section 3). Finally, the team investigated how human coders from the social sciences perform in comparison to the BERT model on a subset of the data (Section 4).

2 Creating the Data Set

Online forums for psycho-social counseling provide a good basis for an empirical evaluation because they contain large amounts of publicly accessible data. For our study, we used posts from a German site for parent counseling. Here, parents who have problems in bringing up their children are seeking advice. Possible topics are, for example, drug abuse by the child or inadequate school performance. A user can start a new thread with a problem description. Professional counselors reply and discuss solution approaches with the initial

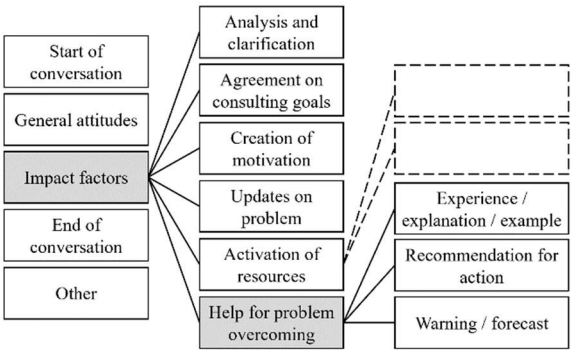

Figure 1: Illustration of the codebook with an exemplary breakdown of the categories

user and others. Thus, each thread contains a series of posts with questions and suggestions about the initially described problem. Since we are especially interested in counseling patterns, we focused on the posts of professional counselors in our analysis.

2.1 Development of the Codebook

Based on existing scientific theories (Fukkink et al 2009, Dowling & Rickwood 2014) on online counseling and first analyses of the text content, a first version of the codebook was created. The various aspects expected in counseling conversations were mapped to a logical hierarchical structure (see Figure 1). The top-level covers general counseling aspects, such as "General attitudes" or "Impact factors". On the intermediate level, these aspects were distinguished more finely, e.g. "Help for problem overcoming". The categories at the lowest level are the ones to be used for the annotation of the text passages, such as "Recommendation for action" or "Warning / forecast".

The different codes were defined as precisely as possible and provided with typical examples. The team of coders applied this codebook to the counseling texts in several turns and iteratively improved the codebook. The final version consists of 51 granular categories (see Appendix A).

2.2 Data Labeling

Based on the codebook described in Section 2.1, a team of coding social scientists manually labeled over 10.000 text sequences in 336 threads. Such a sequence can consist of only a few words (e.g. a greeting) or even multiple sentences (e.g. a recommended action). Sequences, however, do not overlap, i.e. each word should be part of only one labeled sequence. See Figure 2 to get an idea.

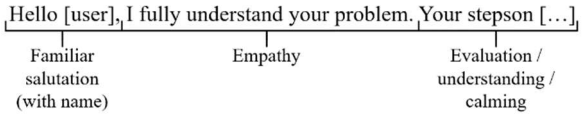

Figure 2: Example of three labeled sequences. The original texts are in German.

In the end, we obtained a heavily imbalanced data set: The average number of samples per category is about 200, but the numbers vary greatly (see Appendix A for more details). For some categories in the area "Impact factors", e.g. "Evaluation / understanding / calming" or "Experience / explanation / example" we obtained over 1000 samples, whereas other categories including "Change" or "Suggestion to put oneself in a problem situation physically" are barely represented. Such an unequal distribution of the frequencies of single codes is not unusual in the social sciences. Since there is no statistical analysis in qualitative research, this is usually not a problem. There are even some research approaches that consider the analysis of very rare codes, in particular, to be extremely insightful (Glaser 2017).

2.3 Data Preparation and Preprocessing

After labeling, we tested the impact of common preprocessing techniques like lemmatization and the removal of usernames. It turned out that both, the support-vector machine classifier as well as the BERT model work best without any of these techniques. Therefore, we used the labeled data without such modifications.

Sequence Length (WordPiece Tokens)	Number of Text Sequences in the Data Set
0-64	8846
65-128	814
129-256	310
257-512	101
>512	16

Table 1: Distribution of text sequence lengths (in WordPiece tokens) in the data set.

However, the BERT model can only process fixed-length sequences consisting of at most 512 subword units called WordPiece tokens (Vaswani et al., 2017). Thus, we restricted the sequence length for all training data. We decided to work with a limit of only 256 WordPiece tokens. This value provides a good trade-off between performance and resource consumption in our setting. Longer sequences yield potentially more accurate

results but generate a high overhead because all sequences must be padded to the specified length. Since only a little more than 1% of the complete data samples contain more than 256 WordPiece tokens, we did not lose much information (cf. Table 1). Instead, the trade-off in length allowed using higher batch sizes and faster training.

To make the results of the different classifiers comparable and to take the data set imbalance into account, a stratified 70-30-train test split was performed on the data set. This results in a training data set with 7169 samples and a test data set with 3072 samples in total. See Appendix A for the number of samples in each category.

3 Model-Based Classification of Psycho-Social Text Sequences

As a result of the created codebook and the collected data, our classification task consists of classifying psycho-social text sequences into one of 51 categories. For the training of the classifiers, the data set described in the previous section with 7169 samples is used. The created models are then evaluated against the 3072 samples in our test data set.

3.1 Support-Vector Machine as a Baseline

The support-vector machine (SVM) is a commonly used classifier due to being lightweight, benefitting from fast training times, and still achieving good results in text classification tasks (Aggarwal, 2018, pp. 12). Therefore, the SVM was chosen as a baseline model. The prepared data was transformed into TF-IDF vectors (bag-of-words) for training and evaluation (Aggarwal, 2018, pp. 24-26).

The model was implemented using the scikit-learn library. The hyperparameters used were chosen according to the results of our hyperparameter tuning. Apart from the default parameters of the TF-IDF-vectorizer, a max_df-value of 0.5 and a min_df-value of 0 was used. Additionally, the inverse-document-frequency reweighting was enabled and unigrams, as well as bigrams, were considered. The support-vector classifier itself used a sigmoid kernel with the gamma value set to "scale", a C-value of 10, and enabled probability estimates which internally enables 5-fold cross-validation.

The SVM achieved a total accuracy of 68.8% on the test data (cf. Table 2). Due to the heavily imbalanced data set, however, the total accuracy is not a good indicator of the model's performance. Thus, we also calculated the macro and weighted F1 scores. The SVM achieves a weighted F1 score of 68.0% (close to the accuracy) and a macro F1 score of 39.7%. The low macro F1 indicates, that classes with little support are frequently misclassified.

Metric	SVM	BERT
Accuracy	68.8%	75.8%
Macro F1 score	39.7%	29.2%
Weighted F1 score	68.0%	74.4%

Table 2: Evaluation metrics on the test data set

A detailed analysis of the results shows that the SVM achieves quite good results in categories with a large number of training samples. For instance, an F1 score of 76.2 % is achieved in the category "Experience / explanation / example" with 1398 training and 599 test sequences. Furthermore, simple sequences that only contain few keywords, such as greeting phrases in the category "Start of conversation", can also be identified quite well, even though only a few training samples exist. In particular, the category "General salutation" achieves an F1 score of 75.0% while only having 22 training and 9 test samples. More complex categories, such as the expression of "Empathy for others", however, achieve lower F1 scores of 59.8% even with a relatively high number of 118 training and 51 test samples. Other categories like "Warning / forecast" achieve even lower F1 scores of only 29.3% even though having 71 training and 30 test samples.

3.2 BERT as Advanced Classifier

BERT is a multi-layer bidirectional Transformer encoder based on the original Transformer implementation described in Vaswani et al. (2017). BERT is typically pre-trained on two unsupervised learning tasks. After the pre-training, the model can be fine-tuned according to the downstream task (Vaswani et al., 2017).

For the classification task in our approach, we used the BertForSequenceClassification implementation from the Hugging Face's Transformers library (Wolf et al., 2019) that combines the BERT Transformer model with a sequence classification head on top (Hugging Face, 2020).

In total, we tested thirteen pre-trained BERT models. Among the ten tested German language models, the results varied between a weighted F1 score of 69.3% and 74.4% on the test data set, whereby the best result was achieved with the pre-

trained uncased language model of the Bavarian State Library (DBMDZ, 2019). The three multilingual models achieved weighted F1 scores as high as 71.0% with the pre-trained language model by DeepPavlov (DeepPavlov, n. d.).

All of the following analyses are, therefore, based on the best performing DBMDZ BERT model.

The hyperparameters used for the fine-tuning were taken from the original BERT publication (Devlin e.a., 2018). Since we are using text sequences with a length of 256 WordPiece tokens, a batch size value of no more than 16 was possible due to GPU memory limitations. Larger models, especially multi-lingual models, even only allowed a batch size of 8. Further testing has shown that the best results can be achieved with a learning rate of 2e-5 and 4 epochs.

3.3 Analyzing the Classification Results

Table 2 shows the different evaluation metrics for both, the SVM and the best BERT classifier.

The low macro F1 score with 29.2% of the BERT classifier compared to the 39.7% of the SVM classifier shows that the BERT classifier performs significantly worse on classes with few samples than the SVM classifier. The result of the weighted F1 score of 74.4% of the BERT model compared to the 68.0% of the SVM model, however, indicates that the BERT classifier outperforms the SVM if the whole data set is considered.

Category	F1 score		Support (Training)
	SVM	BERT	
Other introduction	27.3%	11.8%	37
Activation of resources (professional level)	43.2%	42.1%	49
Wish	63.8%	75.9%	80
Empathy for others	59.8%	49.5%	118
Evaluation / understanding / calming	59.0%	67.0%	1136
Experience / explanation / example	76.2%	83.1%	1398

Table 3: Extract of the classification report

Table 3 shows an extract from the classification report. In general, the BERT classifier improves in its performance with the increase in available training samples for each class.

In specific categories, such as “Empathy for others”, this observation is not true. Categories with this behavior often contain previously mentioned category-specific keywords or phrases which is why the simple bag-of-words approach outperforms the more complex BERT techniques from a statistical point of view. A detailed analysis of the misclassified sequences by the BERT model, however, has shown that the classification of these sequences is not inherently wrong but rather shows suitable alternative affiliations to categories. This behavior is examined in greater detail in Section 3.6.

3.4 Examining other Classification Models

In addition to BERT, other classification models, such as DistilBERT (Sanh et al., 2019), XLM-RoBERTa (Conneau et al., 2019), XLM (Lample and Conneau, 2019), and ULMFit (Howard and Ruder, 2018) were examined in our study as well.

Classification Model	Weighted F1 score
SVM (baseline)	68.0%
BERT (best model)	74.4%
DistilBERT	70.4%
XLM-RoBERTa	70.5%
XLM	65.1%
ULMFit	71.2%

Table 4: Weighted F1 scores of all evaluated classification models

Table 4 shows the best weighted F1 scores of each model. The DistilBERT model performs around 4% worse than the best BERT model on our test data set. This difference lies around the range described by the authors of the DistilBERT paper (Sanh et al., 2019). In addition to that, both the XLM-RoBERTa and XLM models also perform worse than the best BERT classifier. Apart from the Transformer approaches, the bidirectional RNN model called ULMFit was also analyzed. The results show that the different Transformer models as well as the ULMFit model generally perform quite similar on our classification task, except for the XLM model that performs even worse than the simple SVM approach.

3.5 Explaining the Classifiers

Since predictions of BERT, or Transformer models in general, are often untransparent and difficult to

Expert assessment	Number of Samples	Percentage
(I) Both, actual and predicted label would fit	62	32.4%
(II) Predicted label fits better than actual label	49	25.7%
(III) Similar choice of words between actual and predicted classes	29	15.2%
(IV) Sequence contains keywords from other classes	16	8.4%
(V) Assignment cannot be explained by the experts	14	7.3%
(VI) Incorrect sequence	12	6.3%
(VII) Special sequence (uncommon words; not enough context)	6	3.1%
(VIII) Multiple sentences with multiple categories	3	1.6%

Table 5: Expert assessment of incorrectly classified text sequences

justify, different approaches, such as LIME (Ribeiro et al., 2016) or Attention Flow (Abnar and Zuidema, 2020), can be used to generate model insights.

While LIME takes a retrospective approach that can be applied to any classification model, Attention Flow tries to visualize the actual attention maps of Transformer models. Both approaches provide insights that can be used to explain the classification predictions of the models. Since we want to generate model insights regardless of the approach used to create the model, we decided to use LIME as our analyzing tool of choice.

For example, the analysis of the sentence "Have you ever spoken to the kindergarten teachers?" (cf. original German sentence in Figure 3) helps to further understand the model. Originally, the sequence was coded as "Follow-up question" by the expert coders. The BERT classifier did correctly classify this sequence, whereas the SVM classifier classified this sequence as a "Questions about possible support resources".

While both assignments might sound reasonable at first, the question arises why each classifier performed its prediction. To answer this question, the text-heatmaps in Figure 3 were generated with LIME. The percentage values indicate how important the LIME model considers the corresponding word for the classification.

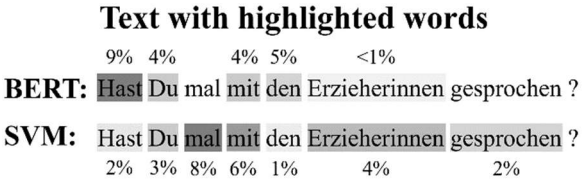

Figure 3: Text heatmaps highlighting the determining words for the classification decision

The BERT heatmap shows that the model mainly focuses on the words that form the question "Hast", "Du", "mit", "den", "Erzieherinnen" (Engl. "have", "you", "with", "kindergarten teachers") while the SVM heatmap shows that the SVM classifier considers all words as important for the classification but with high focus on the word "Erzieherinnen" (Engl. kindergarten teachers) which is a possible support resource.

This strong focus on individual keywords from the SVM can be explained by the operating principle of the bag-of-words approach and verifies the assumption from Section 3.3 that the SVM performs well in classes with distinctive keywords. But examples like this show that this simple approach can also be misled when such distinctive keywords appear in more complex sequences in which the keyword is not decisive for the correct class and the context has to be considered as well for the correct classification.

Since LIME follows a bag-of-words evaluation model, it cannot provide additional insights on how our BERT model exactly handles context. Thus, we can only use LIME to illustrate whether the models' decisions are reasonable, or not.

3.6 Analyzing Misclassified Sequences

To better understand our model and to identify further potential for improvement, the incorrectly classified test data were analyzed.

Out of the 3072 test sequences, the BERT model classified 2325 sequences correctly. Out of the 747 incorrectly classified sequences, our team of social scientists manually examined a sample of 191 sequences. The inspected samples were randomly chosen based on conspicuous categories that were not in the diagonal of the confusion matrix. The summarized results of this examination are shown in Table 5.

The general conclusion of this analysis is that 58.1% (Table 5, I+II) of the incorrectly classified sequences are not inherently wrong but their assigned category depends on the different points of view of the coders. For example, the sequence "Have you ever talked to a pediatrician? Or do you

have a family counseling center?" was initially encoded as a "Question about possible support resources" by the human encoder, whereas the BERT model associated the sequence with a "Follow-up question". In our analysis, the experts concluded that both categories would fit. Another example in which the predicted label would fit even better than the actual label is the sequence "This has to be done consequently, even if screaming is annoying. You have to go through it – sometime." This sequence was initially encoded as a "Warning / forecast" by the human experts. The BERT model, however, assigned this sequence to the category of "Recommendation for action". Since these different interpretation options are not only a technical issue but can also be observed in human coders, the inter-coder reliability between an expert coder, an untrained human coder ("novice"), and BERT is analyzed in Section 4.

For another 23.6% of the analyzed sequences (Table 5, III+IV), we were able to trace back the incorrect classification to the use of keywords or similar terms between different categories. For example, the simple sequence "good luck" is considered to be a "Wish" by the human encoders, whereas our BERT model mistakes this sequence for a traditional farewell phrase (category "Other farewell"). This behavior of the BERT model can be explained by the fact that some sequences in the training data contain closing phrases, such as "Good luck [user]".

In 14 more cases (Table 5, V) the experts were unable to identify any distinctive features that caused the sequences to be classified incorrectly by the BERT model.

Apart from these technical insights, in 12 cases (Table 5, VI) weaknesses in the training data set were identified, such as incorrect assignments of the actual label previously made by the human coder, sequences composed by clients rather than counselors, or sequences that only contain single characters.

Furthermore, in a total of nine sequences (Table 5, VII+VIII), the experts declared the sequences as "hard to assign for humans" due to the usage of uncommon words, not enough context, or since the sequence consists of multiple sentences with multiple categories.

To estimate the impact of the interpretation options during the classification regarding the evaluation metrics, an adjusted accuracy can be estimated. This adjusted accuracy is calculated by transferring the proportion of analyzed incorrectly classified sequences that are not inherently wrong (Table 5, I+II) to the total of the 747 incorrectly classified sequences. This means that 58.1% of the originally incorrectly classified sequences can be considered as correct. This leads to an increase of the correctly classified sequences from 2325 to 2759 which corresponds to a more than satisfying accuracy of 90%, respectively. Since this is only an overall estimation, adjusted F1 scores cannot be calculated.

3.7 Discussion about Improving the Model

To understand the influence of the availability of training samples, we ran multiple tests in which the number of training samples in a specific category was reduced. Hereby, we tested all categories that achieve an F1 score of 70% or higher. For each of the categories, six models were trained with a restricted number (10, 20, 50, 100, 250, and 500) of randomly selected training samples. All models were then evaluated on our test data set. Results have shown that simple categories, such as "General salutation", "Familiar salutation (without name)", "Welcoming", or "Follow-up question", only require about 50 training samples to achieve F1 scores of 0.71 or higher. However, categories that contain text sequences with more complex structures, such as "Experience / Explanation / Example" or "Recommendation for action", still show significant improvements when using 250, 500, or all available text sequences for training.

As described in Section 2, our training data set is unevenly distributed. Data set imbalance is a well-known problem in machine learning (He and Garcia, 2009) and in our case is due to the annotation process. Hereby, available forum posts were annotated without specifically having the category distribution in mind. Typical techniques to reduce the data set imbalance, such as random oversampling or synthetic sampling with data generation (He and Garcia, 2009), cannot easily be applied to textual data, especially not when precise phrasing and wording is important for the classification as in our case. One technique that might, however, lead to improvements is generating new text sequences by randomly combining sentences from other sequences of the same category. Other possible approaches such as aggregating categories with few examples to their superset-level were also considered but dismissed since our goal is to predict categories on a detailed level.

With the approximate number of required samples per category, we think that manually creating additional training data in especially underrepresented classes and edge-cases will, therefore, help to improve the model in the future.

Another idea to improve the model is by taking the model's first and second prediction into account. Human coders can then be supported with suggestions by the model during coding tasks and choose the best fitting label. This feedback can then be used to further improve the model.

4 BERT vs. Human Coders

Coding of text passages is to some degree dependent on the subjective perception of the coders. Especially for similar categories like "Empathy" and "Compassion", different coders will sometimes assign different labels to the same text. Thus, even human coders which were trained on the usage of the codebook will not reach 100% agreement. To get a better understanding of the applicability of our model for automatic coding, we compared the coding performance of BERT against a trained human coder familiar with the codebook ("expert") and an untrained human coder ("novice").

4.1 Intercoder Reliability between Experts

The degree of consensus among coders, the intercoder reliability, is often measured by Cohen's κ (kappa) coefficient (Cohen 1960, Burla et al. 2008). The maximum value of κ is 1, $\kappa > 0.8$ indicates almost perfect, and $\kappa > 0.6$ indicates substantial agreement.

During the creation of the training data, our experts regularly coded the same texts and aligned their coding style. After coding was finished, we calculated the κ coefficient between those two coders who had coded the most samples. Thereby, we considered only posts coded by both coders and text sequences with at least 75% overlap regarding the first and last word. We determined a κ coefficient of 0.73 between those two experts. This value is relatively high given our complex codebook with over 50 categories.

4.2 Intercoder Reliability between an Expert, a Novice, and BERT

To understand how our BERT model performs compared to human coders, we benchmarked the performance of the following three participants: The expert was one of the coders observed in the intercoder reliability measurement. The novice had only a little experience in text annotation and had just recently familiarized herself with the codebook and typical examples for each category. The third participant was our BERT classification model.

All participants had the task to annotate the same 50 text passages. Each text passage was randomly chosen from the set of previously unlabeled forum posts.

Besides measuring the intercoder reliability among the participants, we also wanted to generate indications about which sequence length is best suited for the application of the BERT model. For typical coding tasks in the social sciences, the length of a sequence to be coded is defined by a change in the occurring category. This contrasts with most machine-based classifiers which expect a defined sequence of words as input. The choice of start and end for a label in continuous text is usually not part of the classification task.

Therefore, we generated three variants of the 50 text sequences for coding: The first data set consists of single sentences only, the second data set includes, if existing, the following sentence for each sample, and the third data set contains sequences of at most three consecutive sentences. Figure 4 illustrates the breakdown of an exemplary post.

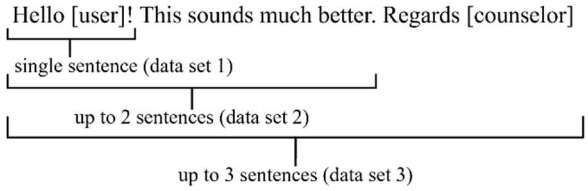

Figure 4: Exemplary structure of the sequences within the different data sets

All three data sets were then coded independently by the participants. As before, the agreement between the different coders was measured using the κ coefficient (see Table 6).

Surprisingly, the intercoder reliability between BERT and the human expert is higher than the intercoder reliability between the expert and the novice, regardless of the sequence length. In its best case, the BERT classifier achieves nearly expert-expert-like intercoder reliability with a value as high as 0.64 in comparison to the earlier calculated expert intercoder reliability of 0.73. It seems that the BERT model has learned the expert style of

Sequence Length	k coefficient		
	Expert-Novice	Expert-BERT	Novice-BERT
Single sentence	0.55	0.64	0.53
Up to 2 sentences	0.53	0.60	0.43
Up to 3 sentences	0.50	0.50	0.38

Table 6: κ coefficient of the participants

coding from the training data better than an untrained human coder using the codebook.

While classifying sequences that contain only one sentence was rated difficult by the human coders due to the missing context, sequences with up to 3 sentences were rated as too long since they often contained patterns from multiple categories. Therefore, sequences with the length of two sentences were rated as best fitting lengths for classifying sequences by both the novice and the expert coder. In contrast to the ratings of the coders, the intercoder reliability shows the highest values when encoding sequences with the length of only one sentence.

5 Conclusion

It has been shown that machine-based classifiers can reach human-like performance for the annotation of complex categories in psycho-social texts. The results indicate that the models learn to mimic the coding style of the initial creators of the training data. The trained BERT model was even better in coding than a human novice. As in other areas of machine learning, this bears the risk that a model also learns the bias from the training data. Therefore, it is important to understand and regularly check the decisions of the model by human experts.

High coding quality could not be achieved for all codes, however. Especially underrepresented categories, which are common in social sciences, are problematic. Thus, a sufficient number of training samples is an obvious prerequisite for good results.

The typical approach of social sciences in analyzing text corpora consists of coding one text after the other and ignoring unequal frequencies of the individual codes. Our study shows that when using machine learning methods, it is better to generate training examples for as many categories as possible and pay less attention to the complete coding of individual texts. This is an important finding for the organization of future studies in this field.

The investigation of misclassified sequences showed that many recorded misclassifications actually were minor mistakes. The model frequently chose not the actual but a very similar category such that even human experts would regard the assignment plausible. Thus, codes with very similar meanings must be distinguished more sharply to give the model a chance to learn to differentiate.

The analysis of the misclassified sequences of BERT opens up new perspectives for the social sciences: More than half of the "incorrectly classified sequences" appeared to the human expert to be plausible or at least worthy of consideration. Since the discussion of the understanding of individual text passages is an important element of social science research, such plausible misinterpretations can enrich the research process. They offer an alternative way of looking at reality and force the human coder to either rethink his assessments or to better justify them.

Currently, we are working on improving the classification performance. One approach is the generation of additional training data for underrepresented categories. Another idea is using an ensemble of SVM and BERT as a classifier to better utilize the individual strengths of the different models. In any case, the findings on how the models work and perform help to consider such technical aspects in future social science research.

With regard to the application domain, we can conclude that it is definitely possible to analyze online counseling conversations with the help of machine learning. We intend to use machine learning in future research projects to investigate correlations between the different techniques used by counselors and the characteristics and reactions of clients. In addition to the question of whether successful counselors use certain techniques significantly more often than others, it can now be clarified if certain approaches are particularly promising for certain target groups or specific problems. These findings can be integrated into the education of online counselors. Furthermore, assistance systems are conceivable that support online counselors in real-time with information generated from this data.

In any case, the results of this study have shown that it is possible to merge the advantages of qualitative and quantitative approaches in social science with the help of machine learning. Automated data annotation for qualitative analysis is the cornerstone for future insights on an unprecedented level.

References

Abnar, S. and Zuidema, W. (2020) Quantifying Attention Flow in Transformers [Online]. http://dx.doi.org/10.18653/v1/2020.acl-main.385.

Aggarwal, C. C. (2018) Machine Learning for Text, Springer [Online]. https://doi.org/10.1007/978-3-319-73531-3.

Althoff, T., Clark, K. and Leskovec, J. (2016) Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health [Online]. http://dx.doi.org/10.1162/tacl_a_00111.

Bambling, M., King, R., Reid, W. & Wegner, K. (2008) Online counselling: The experience of counsellors providing synchronous single-session counselling to young people, *Counselling and Psychotherapy Research*, vol. 8, no. 2, pp.110–116 [Online]. https://doi.org/10.1080/14733140802055011.

Burla, L., Knierim, B., Barth, J., Liewald, K., Duetz, M. and Abel, T. (2008) From Text to Codings: Intercoder Reliability Assessment in Qualitative Content Analysis, *Nursing research*, vol. 57, no. 2, pp. 113–117 [Online]. https://doi.org/10.1097/01.nnr.0000313482.33917.7d.

Cameron, G., Cameron, D. M., Megaw, G., Bond, R. B., Mulvenna, M., O'Neill, S. B. et al. (2017) Towards a chatbot for digital counselling, *Proceedings of the 31st International BCS Human Computer Interaction Conference (HCI 2017).* BCS Learning & Development [Online]. https://doi.org/10.14236/ewic/HCI2017.24.

Chardon, L., Bagraith, K. S. & King, R. J. (2011) Counseling activity in single-session online counseling with adolescents: An adherence study, *Psychotherapy Research*, vol. 21, no.5, pp. 583–592 [Online]. https://doi.org/10.1080/10503307.2011.592550.

Chatterjee, A., Narahari, K.N., Joshi, M., and Agrawal, P. (2019) SemEval-2019 Task 3: EmoContext - Contextual Emotion Detection in Text. In: *Proceedings of the 13th International Workshop on Semantic Evaluation* (SemEval-2019), pp. 39–48 [Online]. http://dx.doi.org/10.18653/v1/S19-2005.

Cohen J. (1960) A Coefficient of Agreement for Nominal Scales. In: *Educ Psychol Meas.* 20:37–46 [Online]. https://doi.org/10.1177/001316446002000104.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. and Stoyanov, V. (2019) Unsupervised Cross-lingual Representation Learning at Scale [Online]. http://dx.doi.org/10.18653/v1/2020.acl-main.747.

Crowston, K., Liu, X. and Allen, E.E. (2010) Machine learning and rule-based automated coding of qualitative data, *Proc. Am. Soc. Info. Sci. Tech*., vol. 47, pp. 1-2 [Online]. https://doi.org/10.1002/meet.14504701328.

DBMDZ (2019) German BERT [Online]. https://huggingface.co/dbmdz/bert-base-german-uncased.

DeepPavlov (n. d.) Sentence Multilingual BERT [Online]. https://huggingface.co/DeepPavlov/bert-base-multilingual-cased-sentence.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Online]. http://dx.doi.org/10.18653/v1/N19-1423.

Dowling, M. J. & Rickwood, D. J. (2014) Experiences of counsellors providing online chat counselling to young people, *Journal of Psychologists and Counsellors in Schools*, vol. 24, no.2, pp. 183–196. Cambridge University Press, Cambridge, UK [Online]. https://doi.org/10.1017/jgc.2013.28.

Dowling, M. & Rickwood, D. (2015) Investigating individual online synchronous chat counselling processes and treatment outcomes for young people, *Advances in Mental Health*, vol. 12, no. 3, pp. 216-224 [Online]. https://doi.org/10.1080/18374905.2014.11081899.

Ersahin, Z. & Hanley, T. (2017) Using text-based synchronous chat to offer therapeutic support to students: A systematic review of the research literature, *Health Education Journal*, vol. 76, no. 5, pp. 531–543 [Online]. https://doi.org/10.1177/0017896917704675.

Fukkink, R. G. & Hermanns, J. M. A. (2009) Children's experiences with chat support and telephone support, *Journal of Child Psychology and Psychiatry*, vol. 50, no. 6, pp.759-766 [Online]. https://doi.org/10.1111/j.1469-7610.2008.02024.x.

Gatti, F. M., Brivio, E. & Calciano, S. (2016), "Hello! I know you help people here, right?": A qualitative study of young people's acted motivations in text-based counseling, *Children and Youth Services Review*, vol. 71, pp. 27-35 [Online]. https://doi.org/10.1016/j.childyouth.2016.10.029.

Glaser, B. G. & Strauss, A. L. (2017). Discovery of grounded theory: Strategies for qualitative research. Routledge.

Hanley, T. (2012) Understanding the online therapeutic alliance through the eyes of adolescent service users, *Counselling and Psychotherapy Research*, vol. *12, no.*1, pp. 35–43 [Online]. https://doi.org/10.1080/14733145.2011.560273.

He, H. and Garcia, E. A. (2009) Learning from Imbalanced Data, IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284 [Online]. https://doi.org/10.1109/TKDE.2008.239.

Howard, J. and Ruder, S. (2018) Universal Language Model Fine-tuning for Text Classification [Online]. http://dx.doi.org/10.18653/v1/P18-1031.

Hugging Face (2020) BertForSequenceClassification [Online]. https://huggingface.co/transformers/model_doc/bert.html#transformers.BertForSequenceClassification.

Kergel D. (2018) *Qualitative Bildungsforschung – Ein integrativer Ansatz*. Wiesbaden, Springer VS [Online]. https://doi.org/10.1007/978-3-658-18587-9.

King, R., Bambling, M., Reid, W. & Thomas, I. (2006) Telephone and online counselling for young people: A naturalistic comparison of session outcome, session impact and therapeutic alliance, *Counselling and Psychotherapy Research*, vol. 6, no. 3, pp. 175–181 [Online]. https://doi.org/10.1080/14733140600874084.

Lample, G. and Conneau, A. (2019) Cross-lingual Language Model Pretraining [Online]. http://arxiv.org/pdf/1901.07291v1.

van de Luitgaarden, G. & van der Tier, M. (2018) Establishing working relationships in online social work, *Journal of Social Work*, vol.18 no.3, pp. 307–325
[Online]. https://doi.org/10.1177/1468017316654347.

Mayring, P. (2015) Qualitative Content Analysis: Theoretical Background and Procedures (Advances in Mathematics Education), *in* A. Bikner-Ahsbahs, C. Knipping & N. Presmeg (eds)*, Approaches to Qualitative Research in Mathematics Education: Examples of Methodology and Methods.* Dordrecht, Springer Netherlands, pp. 365-380 [Online]. https://doi.org/10.1007/978-94-017-9181-6_13.

Navarro, P., Bambling, M., Sheffield, J. & Edirippulige, S. (2019) Exploring Young People's Perceptions of the Effectiveness of Text-Based Online Counseling: Mixed Methods Pilot Study, JMIR Mental Health, vol. 6, no. 7, e13152 [Online]. https://doi.org/10.2196/13152.

Nieuwboer, C. C., Fukkink, R. G. & Hermanns, J. M. A. (2015) Single session email consultation for parents: an evaluation of its effect on empowerment, *British Journal of Guidance & Counselling*, vol. *43*, no. 1, pp. 131–143 [Online]. https://doi.org/10.1080/03069885.2014.929636.

Pérez-Rosas, V., Wu, X., Resnicow, K. and Mihalcea, R. (2019) What Makes a Good Counselor? Learning to Distinguish between High-quality and Low-quality Counseling Conversations, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy. Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 926–935 [Online]. https://doi.org/10.18653/v1/P19-1088.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier [Online]. http://dx.doi.org/10.18653/v1/N16-3020.

Rodda, S. N., Lubman, D. I., Cheetham, A., Dowling, N. A. & Jackson, A. C. (2015) Single session web-based counselling: a thematic analysis of content from the perspective of the client, *British Journal of Guidance & Counselling*, vol. 43, no.1, pp. 117–130
[Online]. https://doi.org/10.1080/03069885.2014.938609.

Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [Online]. http://arxiv.org/pdf/1910.01108v4.

Sefi, A. & Hanley, T. (2012) Examining the complexities of measuring effectiveness of online counselling for young people using routine evaluation data, *Pastoral Care in Education*, vol. 30, no. 1, pp. 49–64 [Online]. https://doi.org/10.1080/02643944.2011.651224.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017) Attention Is All You Need [Online]. http://arxiv.org/pdf/1706.03762v5.

Wolf, T., e.a. (2019) HuggingFace's Transformers: State-of-the-art Natural Language Processing [Online]. http://arxiv.org/pdf/1910.03771v5.

Appendix A. Codebook Including Number of Samples and Classification Results

Top Level	Superset Level	Category Level	Training Support	Test Support	SVM F1 Score	BERT F1 Score
Start of conversation	**Salutation**	General salutation	22	9	75.0%	82.4%
		Formal salutation (without name)	2	1	0.0%	0.0%
		Formal salutation (with name)	4	1	100.0%	0.0%
		Familiar salutation (without name)	139	59	71.7%	98.3%
		Familiar salutation (with name)	579	248	89.5%	98.6%
		Other salutation	8	3	0.0%	0.0%
	Welcoming	Welcoming	102	44	91.3%	93.5%
		Introduction institution / consultant	5	2	0.0%	0.0%
		Other introduction	37	16	27.3%	11.8%
	Conversation management	Conversation management	1	0	0.0%	0.0%
		Organizational issues	14	6	0.0%	0.0%
		Technical issues	6	2	66.7%	0.0%
		Reference to post	190	81	53.2%	52.5%
		Explanatory modalities	4	1	0.0%	0.0%
General attitudes	**Empathy**	Empathy	1	0	0.0%	0.0%
		Empathy for others	118	51	59.8%	49.5%
		Compassion	17	7	72.7%	0.0%
		Concern for others	3	2	0.0%	0.0%
	Appreciation	Congratulations	1	1	0.0%	0.0%
		Praise / acknowledgement	43	18	38.5%	30.8%
		Gratitude / appreciation	8	3	0.0%	0.0%
	Congruence	Wish	80	34	63.8%	75.9%
Impact factors (1 of 2)	**Analysis and clarification**	Analysis and clarification	3	2	66.7%	0.0%
		Follow-up question	491	211	65.3%	85.1%
		Communication of grasp	176	75	42.5%	49.3%
		Evaluation / understanding / calming	1136	487	59.0%	67.0%
	Agreement on consulting goals	Demand for concern	1	1	0.0%	0.0%
		Encouragement to think about the concern	2	1	0.0%	0.0%
		Definition of the objective	1	1	0.0%	0.0%
	Creation of motivation	Encouragement	134	57	36.6%	34.2%
		Change	1	0	0.0%	0.0%
	Update on problem	Request for detailed description	10	5	25.0%	0.0%
		Suggestion to put oneself in a problem situation physically	1	1	100.0%	0.0%
		Suggestion to put oneself in a problem situation mentally	2	1	0.0%	0.0%

Top Level	Superset Level	Category Level	Training Support	Test Support	SVM F1 Score	BERT F1 Score
Impact factors (2 of 2)	**Activation of resources**	Question about possible support resources	15	6	16.7%	0.0%
		Activation of resources (family)	4	2	0.0%	0.0%
		Activation of resources (friends)	2	1	100.0%	0.0%
		Activation of resources (professional level)	49	21	43.2%	42.1%
		Activation of resources (uncertain level)	3	2	0.0%	0.0%
	Help for problem overcoming	Experience / explanation / example	1398	599	76.2%	83.1%
		Recommendation for action	1372	588	68.1%	75.5%
		Warning / forecast	71	30	29.3%	12.1%
End of conversation	**Suggestion for private exchange**	Suggestion for private exchange	8	4	0.0%	0.0%
	Suggestion for further forum exchange	Suggestion for further forum exchange	51	22	52.9%	50.0%
		Formal farewell (with a name)	3	2	50.0%	0.0%
		Familiar farewell (without a name)	21	9	57.1%	71.4%
		Familiar farewell (with a name)	629	269	90.2%	92.7%
		Other farewell	135	58	52.7%	53.2%
Other	**Typographical error**	Typographical error	1	1	0.0%	0.0%
	Inappropriate comment	Inappropriate comment	10	4	0.0%	0.0%
	Emotional clarification	Emotional clarification	55	23	64.6%	90.9%