Extracting Behaviors from German Clinical Interviews in Support of Autism Spectrum Diagnosis

Margareta A. Kulcsar
École Normale Supérieure Paris-Saclay
margareta.kulcsar@ens-paris-saclay.fr

Ian Paul Grant
Queen Mary University London
i.p.grant@qmul.ac.uk

Massimo Poesio
Queen Mary University London
Utrecht University
m.poesio@qmul.ac.uk

Abstract

Accurate identification of behaviors is essential for diagnosing developmental disorders such as Autism Spectrum Disorder (ASD). We frame the extraction of behaviors from text as a specialized form of event extraction grounded in the TimeML framework and evaluate two approaches: a pipeline model and an end-to-end model that directly extracts behavior spans from raw text. We introduce two novel datasets: a new clinical annotation of an existing Reddit corpus of parent-authored posts in English and a clinically annotated corpus of German ASD diagnostic interviews. On the English dataset, the end-to-end BERT model achieved an F1 score of 73.4% in binary behavior classification, outperforming the pipeline models (F1: 66.8% and 53.65%). On the German clinical dataset, the end-to-end model reached an even higher F1 score of 80.1%, again outperforming the pipeline (F1: 78.7%) and approaching the gold-annotated upper bound (F1: 92.9%). These results demonstrate that behavior classification benefits from direct extraction, and that our method generalizes across domains and languages. We release our code and dataset at : https://github.com/MaggieK410/Behavior_Extraction_from_Clinical_Interviews.git

1 Introduction

Accurate identification of behaviors and symptoms is essential for diagnosing developmental disorders such as Autism Spectrum Disorder (ASD), where behavior markers like repetitive movement, avoidant behaviors or absence of socially important behaviors are key diagnostic criteria (World Health Organization, 2023; American Psychiatric Association, 2013). However, existing tools for behavior and symptom identification rely heavily on qualitative analysis techniques (e.g., interviews, observations) that are time-consuming and subject to interpretation and whereby valuable information can be overlooked (Rutter et al., 2003; National Institute for Health and Care Excellence, 2023).

Although Event Extraction (EE) methods in NLP have been used in clinical contexts, such as extraction of symptoms and treatment decision in clinical records of medical disorders (Viani et al., 2020; Tung and Lu, 2016; Guzman-Nateras et al., 2022), general EE is ill-suited for ASD due to its heterogeneity and the nuanced presentation of behaviors and symptoms. Standard EE approaches identify general events, but not all events represent human behavior, particularly those that lack agentive or embodied action (Drury et al., 2022; Skinner, 1938). To support ASD diagnosis with behavior extraction, it is thus necessary to specify which events count as behaviors in the clinical sense. In this work, we frame binary behavior extraction as a specialized form of EE, and apply it to the analysis of actual clinical interviews in a novel corpus. Our contributions are threefold:
(1) We define a binary behavior classification scheme that embeds behavioral definitions within the TimeML framework.
(2) We present two new datasets for binary ASD behavior classification from events, annotated with our novel scheme: an English Reddit dataset from parents of autistic children and a German clinical interview corpus.
(3) We develop and compare pipeline and end-to-end models for behavior extraction both in English and German.
Our results show that end-to-end models trained directly on annotated behaviors outperform pipeline approaches.

In Section 2 we explore previous work on event extraction in clinical applications, in particular for behavior classification in ASD. We then describe the methods applied to both the English pilot study

Proceedings of the 16th International Conference on Computational Semantics, pages 143–155
September 22-23, 2025,

and the German clinical data, including data preprocessing, the pipeline, and the direct classification approach for behavior analysis. In the subsequent Section 5 about the English pilot study, we detail datasets, considered models, training and results for EE, and behavior classification. We conclude Section 5 with learned lessons from the pilot study. Next, we focus on our experiments with German data in Section 6, which is structured in the same way as Section 5 and details the new dataset, models, training, results and discussion. We present our conclusions in Section 7 and the limitations of our work in Section 8.

2 Previous Work

2.1 Event Extraction and Clinical Applications

EE is a subtask of Information Extraction that focuses on identifying and categorizing events, defined as actions or occurrences situated in time and space. EE models typically extract event triggers and associated arguments (e.g., agents, objects, time). Benchmark datasets include ACE 2005 (LDC, 2005), which categorizes events across predefined types (e.g., Life, Movement, Conflict), and TimeBank (Pustejovsky et al., 2003a), which emphasizes temporal properties of events. TAC-KBP (Ellis et al., 2015) extends these with knowledge base population objectives. These corpora are primarily based on newswire or forum data, limiting their direct applicability to clinical language.

EE has successfully been applied in medical NLP, where it supports the extraction of symptoms, clinical events, and diagnostic information from unstructured texts such as electronic health records (EHRs) and clinical notes. For instance, EE has been used to detect negative emotions, thoughts, and symptoms from patient narratives and social media (Tung and Lu, 2016; Guzman-Nateras et al., 2022). However, such applications often focus on mood disorders, such as depression or anxiety, where symptom expressions are relatively homogeneous, for example, fatigue, reduced activity, and increased sleep. In contrast, symptoms of ASD can strongly vary by patient: While one patient can be highly verbal and socially eager, another patient can be non-verbal and seeking sensory input through self-stimulatory behaviour.

2.2 Event Extraction and ASD Detection

ASD presents unique challenges for automated analysis due to the heterogeneity of symptom presentation and atypical use of language.

Most digital tools for behavior and symptom identification in ASD remain underdeveloped. Standard NLP pipelines often fail to accommodate the idiosyncratic and context-dependent nature of autism-related behaviors (Calvo et al., 2017; Themistocleous et al., 2024).

Existing EE models assume relatively consistent linguistic patterns across populations, which makes them poorly suited for capturing the diverse behavioral descriptions in ASD, particularly from caregiver accounts (Zhang et al., 2022; Jurafsky and Martin, 2013). Due to data protection constraints, most text-based ASD studies rely on social media corpora (e.g., Reddit, Twitter) (Zirikly et al., 2019; Amir et al., 2019), which differ markedly from clinical interviews or third-party reports.

Although some work has focused on detecting discrete behaviors from text (Yates et al., 2017; Tadesse et al., 2019), these are typically surface-level behaviors and not grounded in a conceptual model of behavior relevant to developmental disorders (Skinner, 1938). Despite the similarity in the conception of events and behaviors, there exists a research gap in applying EE specifically to behavior detection for ASD within diagnostic settings.

2.3 Temporal Annotation with TimeML

TimeML is an annotation framework that supports fine-grained event labeling, including temporal categories (Pustejovsky et al., 2003b). Its application in mental health NLP has enabled the construction of patient timelines (e.g., tracking the duration of untreated psychosis from EHRs) (Viani et al., 2020). These tools help model behavioral onset and change over time, which is highly relevant for developmental disorders. While our work does not yet focus on temporal reasoning, TimeML's structured event taxonomy forms the basis of our behavior classification system.

2.4 Behavior Extraction and ASD

Our approach addresses this gap by adapting EE to behavior extraction, using third-party reports (e.g. transcripts of parent interviews) and applying a behavior-specific classification scheme grounded in TimeML, supporting the extraction of diagnostically relevant information for ASD.

3 Defining Behaviors in Terms of Events

Behaviors and events share key properties: they are observable, unfold over time, involve agents, and can be causally linked to outcomes. However, not all events describe behaviors. For behavior extraction in text, we require a more specific definition grounded in linguistic and psychological theory.

The TimeML annotation framework (Pustejovsky et al., 2003b) categorizes events into types such as *Occurrence* (actions that happen), *Perception* (sensory experiences), *Reporting* (communication acts), *Aspectual* (beginning, ending, or continuing another event), *I_Action/I_State* (intentions or mental states), and *State* (persistent conditions). These categories offer a rich base for distinguishing between behaviors and other event types.

Our behavior annotation follows a two-step process: (1) identify TimeML-style events in text; and (2) classify which of these constitute behaviors and which do not in a binary fashion. For instance, only agentive and embodied actions (e.g., a child "makes eye contact" or "repeats phrases") qualify as behaviors. In contrast, mental states or results of actions (e.g., "was upset", "was ignored") are excluded. Some event types like *State* and *Aspectual* never meet the definitional criteria for behavior as action by definition.

This filtered annotation is used to train both the pipeline and end-to-end behavior classification models. By grounding our behavior definition in the TimeML schema, we bridge the gap between generic event detection and clinically meaningful behavior identification.

4 Methods

We experiment with two datasets: a publicly available English Reddit dataset newly annotated for events and behaviors, and a novel German clinical interview dataset. The first enables comparison with existing methods, while the second provides real-world clinical insights. We discuss here the common aspects of the methods used in the two experiments.

4.1 Pre-processing

The pre-processing for both English and German data is identical and differs only by task and model. The BERT model classifies each token of the input, while generative models such as T5 and Phi3 get textual input with prompts and generate an output sequence. Since the generative models might produce outputs of different length than the input, we chose to set the maximum output length to be equal to length of the input sentence plus one additional token. This constraint promotes precision by excluding tokens beyond the input length from the loss calculation, thereby increasing the influence of earlier errors in the sequence.

For the EE task, both BERT and T5 receive a raw sentence as input. We embed the input sentence into one of two prompt templates with varying amounts of contextual information about events, and examples for Phi3 model (see Appendix A.1). The outputs are post-processed to ensure consistency for evaluation. The BERT token classification model outputs a predicted class label for each input token. The generative T5 and Phi-3 models produce event-tagged sentences that include event delimiters. An example of the raw input sentence, the desired event-tagged sentence and the tokenwise classifications by BERT is given below:

Raw input sentence: Aber wir beginnen mal. (Translation: "But let's get started.")
Event-tagged sentence: Aber wir [ASPECTUAL] beginnen [END ASPECTUAL] mal.
Tokenized raw sentence:["Aber", "wir", "beginnen", "mal"]
List of token event classifications: [0, 0, 3, 0, 0]

Behavior classification can be approached either end-to-end, using the raw input sentence, or in a pipeline setting, using an event-tagged sentence as input. In behavior classification using BERT as a token classifier (BERT BC), each token is labeled as either not an event (number 2), an event but not a behavior (number 0), or an event and also a behavior (number 1). For the Phi3 behavior classification model (Phi3 BC), we add definition of behavior, extract the mentions in the tagged sentence and instruct it to classify each mention into behavior and non-behavior (see Appendix A.2). Phi3 outputs a list of mentions and their corresponding classifications. For evaluation, we disregard specific event categories by replacing them with "[EVENT]", "[END EVENT]" and "[EVENT, Bx]", since the end-to-end model can only classify into behavior and non behavior tokens. "[EVENT, Bx]" refers to the beginning of an event that is also a behavior, while "[EVENT]" marks the beginning of an event that is not a behavior. "[END EVENT]" is the end delimiter for both types of behaviors. This is illustrated in the following example:

Behavior annotated event-tagged sentence: Er [OCCURRENCE, Bx] spricht [END OCCURRENCE] mit dem Hund meiner Schwester Englisch. (Translation: "He speaks English with my sister's dog")
Event-tagged input sentence: Er [OCCURRENCE] spricht [END OCCURRENCE] mit dem Hund meiner Schwester Englisch.
Tokenized event-tagged sentence: ['Er', '[OCCURRENCE]', 'spricht', '[END OCCURRENCE]', 'mit', 'dem', 'Hund', 'meiner', 'Schwester', 'Englisch
Gold token map: [2, 2, 1, 2, 2, 2, 2, 2, 2, 2]

4.2 Behavior Classification: Pipeline vs. End-to-End

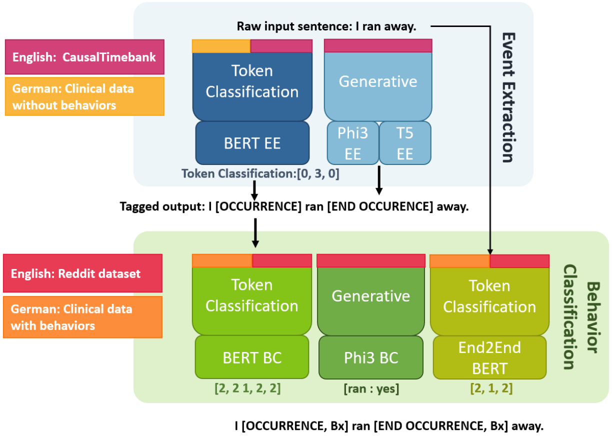

Figure 1: Overview of our experiment design with both English and German data.

For pipeline models, we consider only events as potential behaviors: we first extract events using the BERT EE token classification model, and then classify some of these events as behaviors in a second step with either BERT BC or Phi3 BC as the behavior classification model. By incorporating semantic information about event classes and mentions, the pipeline approach allows the behavior classifier to benefit from correlations between behaviors and specific event types. We investigate whether these benefits transfer equally across structurally different languages, such as German and English.

In the end-to-end model, each token in the raw input sentence is classified without any information about events. While the raw input provides no additional event information to the behavior classifier, avoiding a pipeline architecture reduces the risk of error propagation.

In a pilot study on publicly available english data, we compare generative models with token classifiers at both event extraction and behavior classification level in our pipeline. We contrast the pipeline approach with direct behavior classification on token level. The pilot study allows us to draw preliminary conclusions about which model types are successful before applying them to German clinical data. Figure 1 shows a visual overview of the pipeline and the direct approach to the behavior classification task and highlights, which models were trained with English data and which with German clinical data.

4.3 Post-Processing and Evaluation

Our evaluation for both tasks focuses on the mentions that are extracted from a model output. Evaluating EE models is inherently challenging due to the possibility of partial correctness, for example, extracting the correct text span but assigning the wrong event class (Peng et al., 2023; Zheng et al., 2021). Generative models such as Phi3 and T5 introduce additional complexity. While they are effective for tasks without strict output constraints, they are prone to hallucinating mentions or entire event classes.

In our EE evaluation, we employ F1, precision, and recall to evaluate the models' performances in identifying event classes, mentions, and spans. We extract mentions by aligning the model's token map outputs with the original sentences for BERT and extract mentions using regex from the tagged sentences generated by Phi3 and T5.

As mentioned in Section 4.1, we disregard specific event classes in the behavior classification task. Instead, we introduce the "Bx" addition ([EVENT, Bx]) to the class-neutral event delimiter [EVENT] to indicate that an event has been classified as a behavior. We then compare the extracted mentions, their spans, and the associated behavior classifications. Additionally, we compare the EE part of the pipeline to the end-to-end model by omitting the "Bx" addition and place a delimiter wherever the token map has a value that is not 2.

5 Pilots with Existing (English) Data

While developing the new German clinical dataset, we piloted our approach on an existing English dataset of non-clinical texts, which was newly annotated for behaviors and events by our clinical collaborators. This allowed us to evaluate different

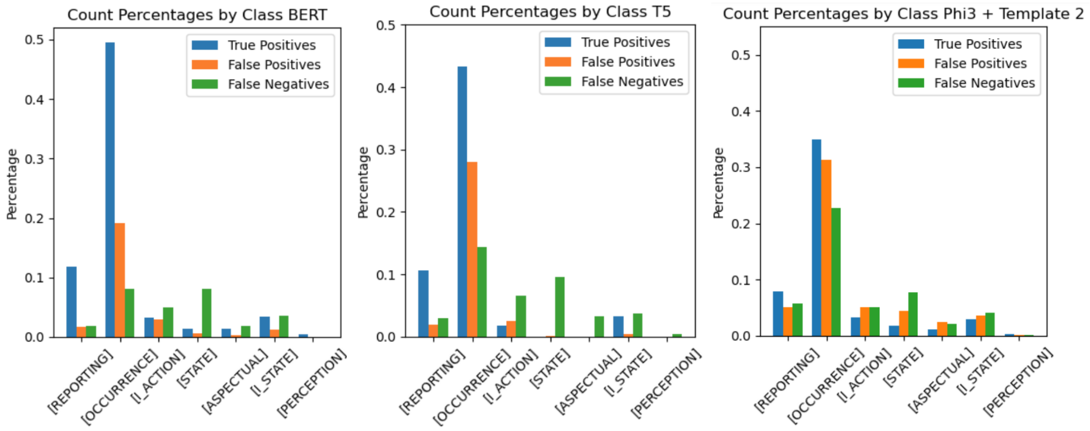

Figure 2: Normalized counts of false positives, false negatives, and true positives by class. We exclude "None" classifications, which refer to tokens that do not correspond to any event, since all models handle this majority class similarly. Phi3 shows the highest rate of false positives due to hallucinations and also the most false negatives, particularly in the *Occurrence* class.

methods in comparison to existing event extraction approaches. We report these preliminary experiments in this section.

5.1 TimeML Event Extraction

As a pilot experiment on English data, we evaluate a diverse set of model families to assess their suitability for downstream application to the German dataset. Specifically, we compare generative models Phi3-mini-128k-instruct (referred to as Phi3 hereafter)[1] and T5-base[2], with token classification model BERT-base-cased[3].

5.1.1 Data

For training and testing EE on English data we used CausalTimebank,[4] a freely available subset of the TimeML-annotated TimeBank dataset. We split the 6811 articles into 2655 sentences and produce train, test and validation sets with a 8:1:1 ratio (2123, 266, 266 sentences). All TimeML event classes are present in the dataset. These are *Occurrence*, *Reporting*, *I_Action*, *State*, *I_State*, *Aspectual* and *Perception*. This mirrors our German clinical dataset, which also includes all event classes.

5.1.2 Models and Training

We train the BERT-base-cased model for ten epochs on our pre-processed CausalTimebank dataset using the token classification objective with a learning rate of 2e-5. The generative models T5 and Phi3 are trained over 5 epochs. For Phi3 we set the learning rate to 2e-5 and the maximum length of the output to 1500 characters to accommodate the prompts. We set the learning rate to 5e-5 for T5. All models were trained on a single A40 GPU with 48GB RAM.

5.1.3 Results and Discussion

In Table 1 we report the performance of the models based on exact matches of mention, span and event class. We compared the weighted true positive, false positive, and false negative counts by event class for the BERT and T5 models with the Phi3 model using template 2 in Figure 2.

Model	Prec.	Rec.	F1
BERT-base-cased (BERT EE)	69.90%	72.19%	71.41%
T5	65.15%	56.12%	60.27%
Phi3 + template 1	72.61%	59.55%	65.08%
Phi3 + template 2	73.56%	65.63%	69.37%

Table 1: Recall, precision and F1 values for exact match for span, mention and event class for the models employed.

The results show BERT outperforming T5 and Phi3 in F1 and recall value. While Phi3 achieves slightly higher precision values then BERT and T5, but lower recall values lead to overall lower F1 values. The T5 model is outperformed by both Phi3 and BERT. The two template variants of Phi3 alter the F1 score by 4% and have a greater effect on recall, and consequently on the false negative rate, than on precision. This difference in performance highlights the importance of prompt engi-

[1] https://huggingface.co/microsoft/Phi-3-mini-128k-instruct

[2] https://huggingface.co/google-t5/t5-base

[3] https://huggingface.co/google-bert/bert-base-cased

[4] https://github.com/paramitamirza/Causal-TimeBank

neering. Recent research (Shiri et al., 2024) shows that providing more event information improves model performance. However, this approach increases training time and memory costs. Reward functions have also been shown to enhance LLM performance in EE (Gao et al., 2024) and could be a starting point for future work.

Elevated levels of false positives in Figure 2 indicate that generation errors influence the results of T5 and Phi3. The generative models also exhibit higher false negative counts especially in the high occurring classes compared to BERT. BERT achieves the highest true positive rate for the most common class *Occurrence*. T5 underperforms on the *Occurrence* class and shows higher false positive rates across all classes, a pattern even more pronounced in the Phi3 results. Phi3 includes more false negatives in the *Occurrence* class, but has similar false negatives levels in the rarer classes, showing that the performance difference mainly stems from the most prevalent *Occurrence* class. Limiting output to the input length plus one token simplifies alignment and prevents drifting, but may lower the number of true positives in T5. When false positives are generated at the beginning of the sentence, the cutoff potentially eliminates correctly tagged mentions later in the sentence.

5.2 Behavior Classification

We train BERT BC and Phi3 BC as behavior classifiers on our English Reddit dataset and compare a pipeline approach with an end-to-end approach.

5.2.1 Data

Although our primary evaluation uses German clinical data, access to high-quality medical datasets is often limited. To train models for behavior annotation in English, we use the publicly available Reddit dataset[5] with posts collected between December 2022 and March 2024 from Autism related subreddits. These posts, primarily written by parents detailing their autistic children's behaviors and experiences, are shorter and less structured than professional consultations, but serve as a valuable resource for testing the abilities of models to extract useful information from third-party descriptions of behavior. Leveraging publicly available data, shows that our approach generalizes to non-clinical settings and may enable future cross-lingual analyses.

Two clinical psychology experts, trained in the TimeML scheme, classified events and behaviors in 1,000 posts from raw text, creating a new English dataset used for the experiments in this section. We obtained a Fleiss kappa value of 0.53 for inter annotator agreement, which in the psychological literature is considered fair to good.

We consider the 743 sentences that contain events, and obtain a total of 2159 events from the annotated Reddit data. The data was split into train, test and validations splits with a ratio of 8:1:1 resulting in 216, 221 and 1722 mentions for test, validation and training, respectively.

This Reddit dataset, the first behavior classification dataset grounded in the psychological definition of behavior, will be released alongside this paper.

5.2.2 Models and Training

For the pipeline behavior extractor, we trained BERT for token classification (BERT BC) and Phi3 (Phi3 BC) for mention-level classification (see Appendix A.2) on our expert-annotated event sentences, using 10 epochs and a learning rate of 2e-5. For evaluation, we combine BERT BC with human-annotated events to estimate an upper bound, and use both Phi3 BC and BERT BC with events extracted by the BERT EE model from our previous experiment. We compare these pipeline models with an end-to-end BERT token classification model, predicting behaviors directly from the raw input.

5.2.3 Results and Discussion

We report precision, recall and F1 values for exact match of span and mention with and without behavior classification in Table 2. We also display an error analysis on token level using confusion matrices for each classification in Figure 3 for the two best performing pipeline models. Overall, the end-to-end BERT model outperforms the pipeline approach with a BERT EE model and a subsequent BERT BC or Phi3 BC behavior classifiers. The upper bound results using gold event annotations show that with a perfect event extraction model, a pipeline approach would significantly improve behavior classification over an end-to-end model. This suggests that the semantic information carried in the tagged events could enhance performance if captured accurately with the EE model. However, the performance gap between pipelines using gold versus predicted events illustrates the difficulty of

[5] https:\/\/huggingface.co\/datasets\/Osondu\/reddit_autism_dataset

accurate event extraction and how errors in this step reduce the pipeline's overall effectiveness. Additionally, Phi3 BC performed poorly as a behavior classifier and introduced further errors, possibly due to the limited size of the Reddit dataset or a suboptimal prompt.

The confusion matrices indicate that models can learn to distinguish behaviors from non-behaviors, which indicates the presence of identifiable patterns that make behavior extraction statistically feasible.

With behavior classification

Model	Prec.	Rec.	F1
Gold EE + BERT BC	82.50%	82.50%	82.50%
End-to-end: BERT BC	**73.27%**	**73.61%**	**73.44%**
BERT EE+BERT BC	67.17%	66.50%	66.83%
BERT EE+Phi3 BC	54.56%	52.78%	53.65%

Without behavior classification

Model	Prec.	Rec.	F1
End-to-end: BERT BC	**85.25%**	**85.65%**	**85.45%**
BERT EE+BERT BC	83.73%	81.02%	82.35%

Table 2: Recall, precision and F1 value for exact match of span and mention with and without behavior classification on the English Reddit dataset.

5.3 Lessons Learned

Our experiments show that generative models are less suited for EE, as they often produce false positives and hallucinations that compromise performance and complicate evaluation. We compared a pipeline using BERT EE for event extraction and BERT BC or Phi3 BC for behavior classification with an end-to-end BERT model that labels tokens directly. We find that while the pipeline approach can outperform direct token classification under perfect EE, errors from the EE step accumulate and degrade performance. Additionally, since large clinical datasets are often unrealistic in real-world settings and BERT performs significantly better, we use BERT for downstream analysis on the German data. We conclude that token level behavior classification from raw input sentences performs best on the English dataset. To assess how well our approach generalizes across languages, we apply it to structurally different German data, by comparing a behavior classification pipeline (using both gold and BERT-extracted events) to an end-to-end BERT token classification model.

6 Experiments on German Clinical Data

Our main experiment involves the same steps as the pilot with English data, but this time applied to real clinical data in German.

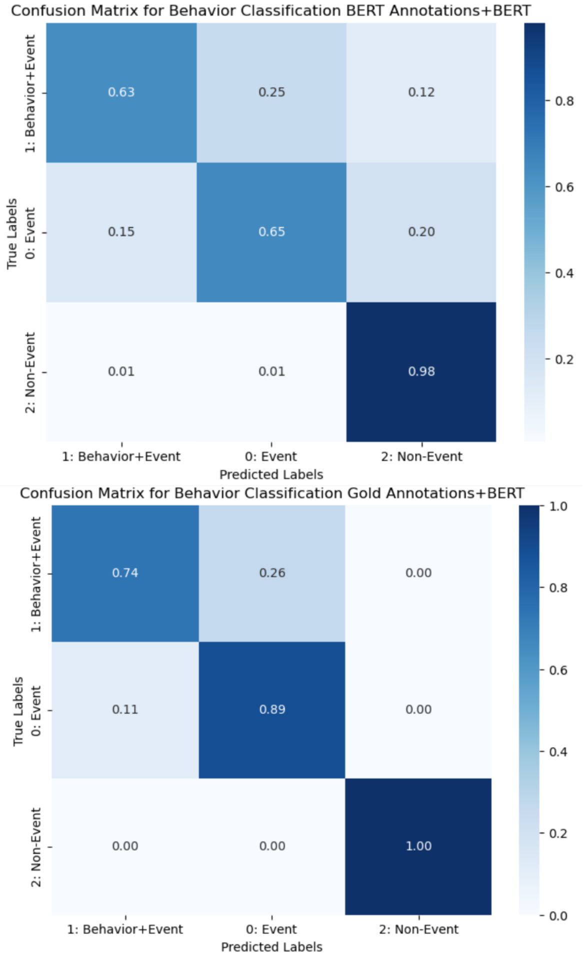

Figure 3: Normalized confusion matrices for behavior classification with BERT event annotations (top) and the gold annotations (bottom) on the Reddit dataset.

6.1 A New Dataset

The English pilot study only includes newspaper articles from CausalTimebank (for EE) and non-clinical Reddit posts (for behavior extraction). For proper evaluation in a clinical setting, we create and annotate four transcribed sessions with parents of autistic children and qualified psychologists asking directed questions about the child's behavior and development using the ADI-R interview. This data was first TimeML annotated by the same clinical experts that annotated the English dataset, and subsequently behavior annotated using the same scheme used for the Reddit dataset. In total, the dataset contains 6566 events. We split our data by patient since we want to be able to generalize to other patients and simulate a realistic training environment. We select two patients for the training dataset, leading to 4,254 events, and the two remaining ones for test and validations set, with 1123 and 1189 events, respectively. We have enough events to train an EE model and a subsequent behavior classification model, as well as a end-to-end model on this data.

This second clinical dataset will also be made publicly available after publication.

6.2 Models and Training

Based on our experiments on English data, we select BERT-base-multilingual-cased[6] for the pipeline, and compare it with end-to-end classification from raw input. We do not use generative models, as we saw on the English data that they achieve lower performance compared to the token classification models. We prepare three distinct versions of the dataset for our experiments: (1) one with raw inputs and event-tagged outputs for training EE models; (2) one with raw inputs and behavior-tagged outputs for training the direct behavior classification model; and (3) one with event-tagged inputs and behavior-tagged outputs for training the behavior classification model using extracted event information. All models were trained for 5 epochs.

6.3 Results and Discussion

Table 3 reports the results for four setups: (1) the end-to-end BERT model, (2) a pipeline using BERT for event extraction (EE) and either BERT or (3) Phi3 for behavior classification, and (4) behavior classification on human-annotated sentences.

The best performing model from raw inputs is the end-to-end model, while the pipeline approaches suffer from error accumulation and performance decline. The upper bound for the subsequent behavior classification model is set by the EE component of the pipeline, which explains the weaker overall performance of the full pipeline.

The results on the clinical German data reflect the same pattern found in the English piloting experiment, which shows that the pilot on non-clinical, easily available data did yield valuable insights for this task that can be expanded to other languages.

However, we observed a notable improvement of $\sim$ 10% across all metrics in the German dataset compared to the English dataset. This is most likely due to the fact that our large clinical dataset contains three times as many events as the English Reddit dataset, enabling more reliable learning for the behavior classification model and resulting in better downstream performance, particularly evident in the BERT BC using human annotations. Additionally, its size of 6,566 events is of a similar scale to the 6,811 events in the CausalTimeBank dataset, allowing the German event extraction models to perform similarly to their English counterparts. A more detailed comparison between the German clinical dataset and CausalTimebank can be found in Figure 4. Since we split the German data by patient to ensure a more realistic clinical setting, event class distributions vary, potentially affecting the EE model's performance on the test set.

With behavior classification

Model	Prec.	Rec.	F1
Gold EE + BERT BC	92.93%	92.93%	92.93%
End-to-end: BERT BC	**79.13%**	**81.11%**	**80.10%**
BERT EE+BERT BC	77.48%	79.98%	78.71%

Without behavior classification

Model	Prec.	Rec.	F1
BERT EE	89.16%	92.03%	90.57%

Table 3: Recall, precision and F1 value for exact match of span and mention with and without behavior classification on the German clinical dataset.

7 Conclusions

We introduce a novel approach for identifying behaviors in text to support ASD diagnosis, by formulating behavior classification as a refinement of EE. Our analysis focuses on ASD behaviors which are described in third person by caretakers. Our approach was tested both on a newly created German dataset of clinical interviews with caretakers of potential ASD patients–to our knowledge, the first clinical German dataset with event and behavior annotations–as well on an existing, publicly available English dataset, which we also newly annotated using the same scheme.

Both of the new behavior classification datasets created for this work, and annotated by psychologists with extensive training in TimeML annotation, will be released.

Our results on both datasets show that the end-to-end model outperforms pipeline models that use an EE model followed by a behavior classifier, primarily due to error accumulation in the EE step. However, with optimal annotations in the EE step, a pipeline approach can outperform the end-to-end

[6] https://huggingface.co/google-bert/bert-base-multilingual-cased

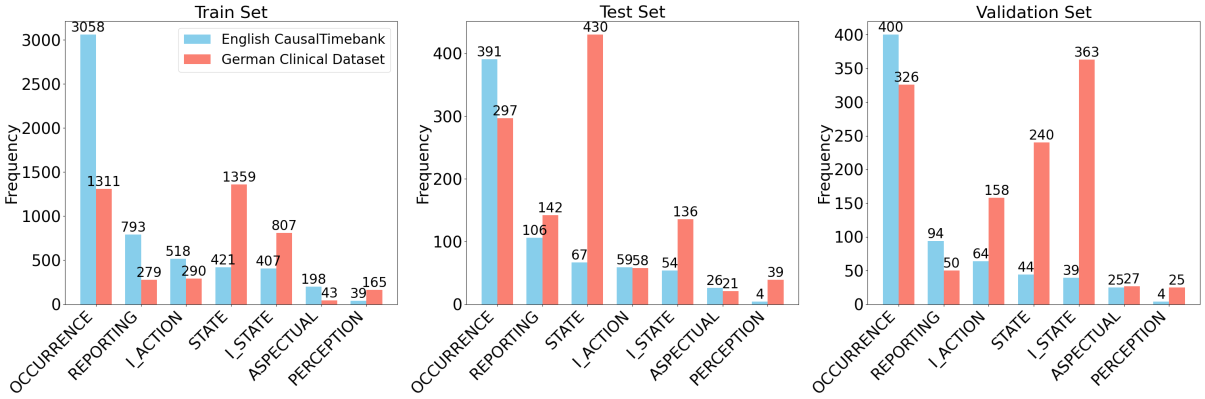

Figure 4: Frequencies of different event classes in the English CausalTimebank dataset and our German clinical dataset. The German data contains a more mixed profile, since we split by patients and not in a stratified way like we did in for CausalTimebank.

model. These results are similar in both English and German, suggesting that our approach is rooted in semantics of events and behaviors.

We can also infer from the results that the application of LLMs in the EE field is still challenging. Overall, Phi3 outperformed T5 with a slight margin, but different prompts for Phi3 had a notable impact on the performance indicating that prompt engineering needs to be further improved. On behavior classification, Phi3 performed notably worse than on EE, possibly because of the smaller dataset and an non-optimized prompt.

These results show that the extraction of behaviors conceptualized in terms of EE especially coupled with token classification has promise for the further development of this technology as well as implications for the development of clinical tool for disorders with idiosyncratic descriptions of behaviors. For example, we presented a prototype visual platform at HealTac2025, where clinicians can upload texts such as session transcripts, and our models extract and highlight events and behaviors in the submitted text.

8 Limitations

Our work explores the application of NLP in area of behavior classification in support of behavior analysis and is aimed at descriptions by parents of autistic children. Although we hope this work helps clinicians focus on important parts of the treatment and save time looking over transcripts and notes, we emphasize that these models do not have a perfect accuracy and are subject to not highlighting important parts of the text. Therefore, close analysis of the outputs by clinicians remains crucial.

Our work covers English and German data, but leaves many languages that might be syntactically different, and therefore more difficult to annotate, open for future work. Especially agglutinative languages might highlight the propagation of errors in EE. Additionally, we release the first German clinical dataset for behavior and event annotations, but there is currently a lack of large scale clinical datasets analyzing behavior and events in other languages.

9 Acknowledgements

Massimo Poesio's research was in part funded by ARCIDUCA, EPSRC EP/W001632/1. Ian Grant's research was funded through Queen Mary University London's Principal Scholarship.

References

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5)*. American Psychiatric Publishing, Arlington, VA. Autism Spectrum Disorder.

Silvio Amir, Mark Dredze, and John W. Ayers. 2019. Mental health surveillance over social media with digital cohorts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 114–120, Minneapolis, Minnesota. Association for Computational Linguistics.

Rafael A. Calvo, David N. Milne, Mohammed S. Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Benjamin Drury, Hugo Gonçalo Oliveira, and António de Araújo Lopes. 2022. A survey of the extraction and applications of causal relations. *Natural Language Engineering*, 28(3):361–400.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie Strassel. 2015. Overview of linguistic resources for the tac kbp 2015 evaluations: Methodologies and results. *Theory and Applications of Categories*.

Jun Gao, Huan Zhao, Wei Wang, Changlong Yu, and Ruifeng Xu. 2024. Eventrl: Enhancing event extraction with outcome supervision for large language models.

Luis Guzman-Nateras, Viet Lai, Amir Pouran Ben Veyseh, Franck Dernoncourt, and Thien Nguyen. 2022. Event detection for suicide understanding. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1952–1961, Seattle, United States. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2013. *Speech and Language Processing*, pearson new international edition edition. Pearson Education, London, UK.

LDC. 2005. ACE (Automatic Content Extraction) 2005 Multilingual Training Corpus. Linguistic Data Consortium, Philadelphia. LDC Catalog No.: LDC2006T06.

National Institute for Health and Care Excellence. 2023. Autism spectrum disorder in under 19s: support and management. Accessed: Jul. 04, 2023.

Hao Peng, Xiaozhi Wang, Feng Yao, Kaisheng Zeng, Lei Hou, Juanzi Li, Zhiyuan Liu, and Weixing Shen. 2023. The devil is in the details: On the pitfalls of event extraction evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9206–9227, Toronto, Canada. Association for Computational Linguistics.

James Pustejovsky, Jose Castano, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir Radev. 2003a. Timebank: Robust annotation of event and temporal expressions. In *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5)*, pages 225–264, Tilburg, The Netherlands.

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Rob Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir Radev. 2003b. Timeml: Robust specification of event and temporal expressions in text. pages 28–34.

Michael Rutter, Ann LeCouteur, and Catherine Lord. 2003. *Autism Diagnostic Interview–Revised (ADI®-R)*. Western Psychological Services.

Fatemeh Shiri, Van Nguyen, Farhad Moghimifar, John Yoo, Gholamreza Haffari, and Yuan-Fang Li. 2024. Decompose, enrich, and extract! schema-aware event extraction using llms. *ArXiv*, abs/2406.01045.

B.F. Skinner. 1938. *The Behavior of Organisms: An Experimental Analysis*. Appleton-Century-Crofts, New York.

Mihretab Molla Tadesse, Hongmin Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893.

Charalambos K. Themistocleous, Maria Andreou, and Eleni Peristeri. 2024. Autism detection in children: Integrating machine learning and natural language processing in narrative analysis. *Behavioral Sciences*, 14(6):459.

Chun-Hung Tung and Wen-Hsiung Lu. 2016. Analyzing depression tendency of web posts using an event-driven depression tendency warning model. *Artificial Intelligence in Medicine*, 66:53–62.

Nicholas Viani, Judy Kam, Liang Yin, et al. 2020. Temporal information extraction from mental health records to identify duration of untreated psychosis. *Journal of Biomedical Semantics*, 11(2):1–10.

World Health Organization. 2023. *ICD-11: International Classification of Diseases, 11th Revision*. World Health Organization. Autism spectrum disorder (6A02).

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.

Tong Zhang, Anne M. Schoene, Shi Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *NPJ Digital Medicine*, 5(1):1–13.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2021. Revisiting the evaluation of end-to-end event extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4609–4617, Online. Association for Computational Linguistics.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.

A Appendix

A.1 EE prompt

The first prompt template (referred to as **template 1** in the paper) includes no extra information about the events and is similar to the T5 input, apart from it containing a simple instruction:

″<|user|>Your task is to extract the events in a sentence. There are 7 event types to consider: OCCURRENCE, I_ACTION, I_STATE, ASPECTUAL, REPORTING, STATE, PERCEPTION. In the following sentence, please extract all the events based on the above classes. Remember, there can be multiple events in a sentence:″

We add the sentence, followed by <|end|><|assistant|> as instructed on the model's webpage to generate the outputs. This prompt yielded the best results.
The second prompt (**template 2**) includes example annotations for each class and performed slightly worse than template 1.

″<|user|>Your task is to extract the events in a sentence. There are 7 event types to consider: OCCURRENCE, I_ACTION, I_STATE, ASPECTUAL, REPORTING, STATE, PERCEPTION.
EXAMPLES:
REPORTING:<|user|>He said that the volcano was spewing gases.<|end|> <|assistant|>He [REPORTING]said[END REPORTING] that the volcano was spewing gases.<|end|>

OCCURRENCE:<|user|>Two moderate eruptions shortly before 3 p.m. Sunday appeared to signal a larger explosion<|end|>
<|assisstant|> Two moderate [OCCURRENCE]eruptions[END OCCURRENCE] shortly before 3 p.m. Sunday appeared to [OCCURRENCE]signal[END OCCURRENCE] a larger [OCCURRENCE]explosion[END OCCURRENCE]<|end|>

I_ACTION:<|user|>Israel has been scrambling to buy more masks abroad.<|end|>
<|assistant|>Israel has been [I_ACTION]scrambling[END I_ACTION] to buy more masks abroad.<|end|>

STATE: <|user|>No injuries were reported over the weekend.<|end|>
<|assistant|>No [STATE]injuries[END STATE] were reported over the weekend<|end|>

I_STATE:<|user|>The agencies fear they will be unable to crack those codes to eavesdrop on spies and crooks.<|end|>
<|assistant|>The agencies [I_STATE]fear[END I_STATE] they will be unable to crack those codes to eavesdrop on spies and crooks.<|end|>

ASPECTUAL:<|user|>The volcano began showing signs of activity in April for the first time in 600 years.<|end|>
<|assistant|>The volcano [ASPECTUAL]began[END ASPECTUAL] showing signs of activity in April for the first time in 600 years<|end|>

PERCEPTION:<|user|>Witnesses tell Birmingham police they saw a man running.<|end|>
<|assistant|>Witnesses tell Birmingham police they [PERCEPTION]saw[END PERCEPTION] a man running.<|end|>

In the following sentence, please extract all the events based on the above class descriptions.″

After this, we add the desired sentence followed by <|end|><|assistant|>. The model performed slightly worse with this prompt. A possible explanation could be the lost in the middle problem, where elements in the middle of a long prompt are forgotten.

A.2 Behavior Classification Prompt

We experiment with only one prompt for behavior classification. It includes a psychological definition and three example sentences:

″Behavior in psychology is defined as: ”That portion of an organism's interaction with its environment that is characterized by detectable displacement in space through time of some part of the organism and that results in a measurable change in at least one aspect of the environment”

Examples:
<|user|> My son is 5 years old & is said to have level 1 autism In this sentence, does "said" describe behavior?<|end|>
<|assistant|> said: yes<|end|>

<|user|> Key words I should be looking for on their websites that are green flags or red flags? In this sentence, does "looking" describe a behavior?<|end|>
<|assistant|> looking: no<|end|>

<|user|> He also likes books and reads books to himself in his own " In this sentence, do "likes" and/or "reads" describe behavior?<|end|>
<|assistant|> likes: no: yes<|end|>″

To integrate the sentences from the dataset, we extract the mentions and create a sentence listing them as in the example. We exclude the three example sentences from the dataset.