CareCorpus: A Corpus of Real-World Solution-Focused Caregiver Strategies for Personalized Pediatric Rehabilitation Service Design

[*]Mina Valizadeh[1], [*†]Vera C Kaelin[2,3], Mary A Khetani[2], Natalie Parde[1]
[1]Department of Computer Science, University of Illinois Chicago
[2]Department of Occupational Therapy, University of Illinois Chicago
[3]Department of Computing Science, Umeå University
{mvaliz2, vkaeli2, mkhetani, parde}@uic.edu

Abstract

In pediatric rehabilitation services, one intervention approach involves using solution-focused caregiver strategies to support children in their daily life activities. The manual sharing of these strategies is not scalable, warranting need for an automated approach to recognize and select relevant strategies. We introduce CareCorpus, a dataset of 780 real-world strategies written by caregivers. Strategies underwent dual-annotation by three trained annotators according to four established rehabilitation classes (i.e., *environment/context*, n=325 strategies; a child's *sense of self*, n=151 strategies; a child's *preferences*, n=104 strategies; and a child's *activity competences*, n=62 strategies) and a *no-strategy* class (n=138 instances) for irrelevant or indeterminate instances. The average percent agreement was 80.18%, with a Cohen's Kappa of 0.75 across all classes. To validate this dataset, we propose multi-grained classification tasks for detecting and categorizing strategies, and establish new performance benchmarks ranging from F_1=0.53-0.79. Our results provide a first step towards a smart option to sort caregiver strategies for use in designing pediatric rehabilitation care plans. This novel, interdisciplinary resource and application is also anticipated to generalize to other pediatric rehabilitation service contexts that target children with developmental need.

Keywords: corpus, caregiver strategies, pediatric rehabilitation

1. Introduction

Pediatric rehabilitation services are delivered in diverse contexts (e.g., home, school, clinic). They can focus on outcomes such as a child's functional skills (e.g., motor or social skills) and *participation*, defined as their attendance and involvement in home, school, and community activities (Imms et al., 2017). Participation is a key rehabilitation outcome endorsed by the World Health Organization (WHO) (World Health Organization, 2001). Service providers and families typically develop care plans based on their child's functional skills and participation level, and they may exchange successful strategies with other caregivers to support their child's participation in meaningful activities (Jarvis et al., 2020, 2019; Kaelin et al., 2021a).

We introduce CareCorpus (Kaelin et al., 2023b), a new dataset of such solution-focused strategies submitted by real caregivers to a web-based healthcare application that supports pediatric rehabilitation service delivery. A solution-focused caregiver strategy is a plan or action designed by caregivers to support their child's participation in daily life activities (Khetani et al., 2013). We manually and exhaustively categorized these strategies according to a current framework for pediatric rehabilitation (Imms et al., 2017) that organizes strategies based on how they support child participation: strategies targeting a child's *environment/context*, *sense of self*, *preferences*, and *activity competencies*. Annotated strategies were also grouped into broader categories of extrinsic (i.e., strategies targeting a child's *environment/context*) and intrinsic (i.e., strategies targeting a child's *sense of self*, *preferences*, and *activity competencies*) strategies (Imms et al., 2017).

Manually determining the most relevant strategies for a family situation from a pool of options is not scalable. This task is complicated by the anticipated increase in available strategies generated when using web-based caregiver tools that are being developed for use in this domain (Jarvis et al., 2020; Kaelin et al., 2022a; Bosak et al., 2019; Khetani et al., 2017, 2015), as well as individual differences between families' rehabilitation goals and preferred strategies. It may be further complicated by structural and stylistic differences in strategy authoring across different caregivers when using these tools (Villegas et al., 2023). Unlike other healthcare tasks (Reyes-Ortiz et al., 2015; Farzana et al., 2020; Khanbhai et al., 2021; Valizadeh and Parde, 2022), exploring automated classification of strategies from caregivers of children accessing pediatric rehabilitation services remains rare in NLP, although recent innovations suggest a pathway to support this process (Zirikly et al., 2022; Albrecht et al., 2020; Jarvis et al., 2020; Kaelin et al., 2023a).

We implement predictive models for detecting and automatically classifying solution-focused caregiver strategies at multiple granularities, vali-

[*]Co-First Authorship.
[†]Work completed at University of Illinois Chicago.

LREC-COLING 2024, pages 2871–2882
20-25 May, 2024.

dating our dataset in settings in which we anticipate its real-world use. In doing so, we establish performance benchmarks ranging from F_1=0.79 for coarse-grained strategy detection to F_1=0.53–0.56 for finer-grained strategy classification. Our primary contributions include:

- We introduce CareCorpus, a dataset of 780 real-world caregiver strategies, organized into five classes including established pediatric rehabilitation categories (*environment/context*, *sense of self*, *preferences*, and *activity competence*) and an additional class (*non-strategy*) for submitted strategies with indeterminate or irrelevant content.

- We propose multi-grained classification tasks for use with CareCorpus, including strategy detection and finer-grained classification following anticipated need and an established pediatric rehabilitation framework.

- We present benchmark models for solving these tasks, achieving performance ranging from F_1=0.53–0.79 depending on task complexity.

- We conduct a clinically-informed analysis of these results to increase translational impact and recommend steps for future work.

This research was conducted by an interdisciplinary research team bringing together researchers with diverse technical and clinical backgrounds. This team includes two natural language processing experts holding PhDs in computer science, and two experts holding research doctorates in rehabilitation science, both of whom have prior experience as clinicians and research expertise on the development and/or use of assessments to capture data on constructs of interest (i.e., solution-focused caregiver strategies). We close by discussing the impact of these findings from both clinical and technical perspectives. Our dataset is publicly available (Kaelin et al., 2023b) to encourage follow-up work.

2. Related Work

Although research towards automatically and intelligently simplifying rehabilitation service delivery is of growing interest to the community, work examining caregiver strategy detection through predictive modeling has been limited (Kaelin et al., 2021b, 2022b). The closest contemporary line of related work may be in automated behavioral coding. For example, Cao et al. (2019) observed high performance in a behavioral coding model for motivational interviewing (MI), predicting counselor behaviors using a hierarchical gated recurrent unit (GRU) trained on a corpus of 353 MI sessions. Gupta et al. (2020) explored automated coding of Specific, Measurable, Attainable, Realistic, and Time-bound (Bovend'Eerdt et al., 2009, SMART) goal-setting behaviors by training a CRF model on 2,583 provider-patient text message exchanges, with promising success. However, the data included in these studies were transcribed interactions or text message exchanges that focused on goal-setting behavior broadly. In contrast, our work seeks to categorize caregiver strategies for supporting child participation specifically—a more targeted clinical task with unique and challenging nuances.

A recent NLP study that more closely focuses on participation sought to classify clinical documents (i.e., claims for federal disability benefits from the U.S. Social Security Administration) in two key areas (mobility and self-care/domestic life) within the International Classification of Functioning, Disability and Health (Newman-Griffis et al., 2021). Their highest performing model achieves strong performance (F_1>0.80), but it focuses on adult and not pediatric services, which differ in their content.

When examining research within participation-focused pediatric rehabilitation, most studies target robotics applications rather than opportunities leveraging language data specifically (Beaudry et al., 2019; Ljunglöf et al., 2011; Zhanatkyzy et al., 2020; Kaelin et al., 2021b). For example, So et al. (2020) demonstrated the use of a robot-based play-drama intervention to promote joint attention initiations and functional play behaviors of children with autism spectrum disorder (ASD). The few studies including participation-focused narrative data aim to capture engagement and rarely include caregiver-reported data on supporting child participation (Kaelin et al., 2022b, 2023a). For example, Chorianopoulou et al. (2017) experimented with detecting engagement of individuals with ASD and typically developing (TD) children through an utterance-level classification task using video-recorded sessions, achieving 62% and 59% unweighted recall for TD children and children with ASD, respectively. Thus, there is a well-motivated need for resources more tailored to supporting children's participation in pediatric rehabilitation contexts. With CareCorpus, we take concrete steps toward filling that gap.

3. Data

3.1. Data Collection

We sourced our dataset from content produced by caregivers of children with developmental need during two pilot implementation trials in a large early intervention (EI) program that includes pedi-

Characteristic, N=125	**n (%)**
Child Gender	
Male	70 (56.0)
Female	53 (42.4)
Missing	2 (1.6)
Child Age (Months)	
< 12	1 (0.8)
12-24	53 (42.4)
>24	70 (56.0)
Reason for EI Services	
Diagnosis	30 (24.0)
Developmental delay (no diagnosis)	85 (68.0)
Risk for developmental delay	10 (8.0)
Family Income($)	
<50,000	24 (19.2)
50,001-100,000	29 (23.2)
>100,001	69 (55.2)
Family Education Level	
High school/Some college	19 (15.2)
College degree	36 (28.8)
Graduate education	70 (56.0)

Table 1: Child and family characteristics alongside their statistics.

atric rehabilitation services such as occupational therapy (Albrecht et al., 2020; Kaelin et al., 2022a). Eligible participants: 1) were caregivers of children with developmental need; 2) were at least 18 years old; 3) had a child between 0-3 years old who had received services in the EI program for at least 3 months; 4) had Internet access; and 5) could read, write, and speak English. A total of 125 caregivers are represented in our dataset; descriptive statistics of our cohort are provided in Table 1. Multi-institutional ethics approval was obtained prior to participant recruitment and remained active throughout our study.

Participating caregivers completed the home and community sections of an electronic patient-reported outcome (e-PRO) measure designed for use in pediatric rehabilitation. Up to 24 open-ended items were administered about solution-focused strategies that caregivers have used to support their child's participation in activities for which change is desired (e.g., "Please describe a strategy that you have tried to help your child participate successfully in basic care routines"). Caregiver completion of these open-ended items resulted in 780 English-language solution-focused strategies, with an average of 6.24 and a range of 1-24 strategies per caregiver. Per IRB protocol and to preserve participants' privacy, we manually de-identified the data (e.g., we replaced any names appearing within the text with generic *[name]* tokens) and we release this de-identified version for public use.

3.2. Data Annotation

The collected data was dual-annotated and adjudicated according to strategy type by three trained annotators. The annotators were two undergraduate students and one graduate student, all of whom were native English speakers on an occupational therapy or pre-occupational therapy track and with expertise in participation-focused pediatric rehabilitation. Two annotators earned course credit and/or pay for conducting this work and one volunteered. Strategy types corresponded to known drivers of participation (Imms et al., 2017), *environment/context*, *sense of self*, *preferences*, and *activity competence*, detailed as follows:

- **Environment/Context (EC):** These strategies target "broad objective social and physical structures" (Imms et al., 2017, p. 20) and the setting where participation takes place, including the people, places, activity, objects, and time.
- **Sense of Self (SOS):** These strategies target "intrapersonal factors related to confidence, satisfaction, self-esteem, and self-determination" (Imms et al., 2017, p. 20).
- **Preferences (P):** These strategies target "interests or activities that hold meaning or are valued" (Imms et al., 2017, p. 20).
- **Activity Competence (AC):** These strategies target a child's "ability to execute the activity according to an expected standard" (Imms et al., 2017, p. 20), including cognitive, physical, and affective skills and abilities.

Any strategies that could not be assigned to one of those four classes were assigned a *no-strategy* label, indicating that they did not describe an identifiable action (Kaelin et al., 2021a). The four strategy classes can be collapsed into a simplified binary classification structure (Imms et al., 2017): *extrinsic strategies* (i.e., *environment/context*) and *intrinsic strategies* (i.e., *sense of self, preferences*, and *activity competence*). Table 2 presents corresponding examples for each class in our dataset. Annotators were trained on annotation procedures and guidelines prior to labeling. Each strategy was labeled independently by two annotators, and discrepancies were discussed and resolved with a third annotator based on majority rule and with feedback from a key informant, both with clinical expertise in pediatric rehabilitation and child participation.

Environment/Context (n=325)	
1.	*Preparing her for the activity and letting her know ahead of time.*
2.	*We've tried to add 10 minutes of tidy-up time into our evening routine.*
Sense of Self (n=151)	
1.	*We encourage him to play, say hello, give hugs.*
2.	*I wipe his mouth then give him the washcloth to wipe his own mouth.*
Preferences (n=104)	
1.	*He loves wooden puzzles.*
2.	*I put out costumes and allow her to choose to wear one if she wants.*
Activity Competence (n=62)	
1.	*Hand-over-hand brushing of teeth and washing of hands.*
2.	*Have her practice certain "moves" at home during the weeks so she hears about them.*
No-Strategy (n=138)	
1.	*None.*
2.	*She needs to be walking first.*

Table 2: Examples from each class.

Setting	**(%) Agreement**	**Kappa**
Multinomial		
Environment/Context	80.31	0.50
Sense of Self	82.78	0.56
Preferences	72.12	0.48
Activity Competence	75.81	0.48
No-Strategy	89.86	0.49
Binary		
Strategy/No-Strategy	95.90	0.80
ES/IS	88.47	0.71

Table 3: Inter-rater agreement, measured using percent agreement and Cohen's Kappa. ES/IS = extrinsic vs intrinsic strategies.

Of the 780 annotated data instances, 138 belonged to the *no-strategy* category. The remaining strategies were distributed across the *environment/context* (n=325), *sense of self* (n=151), *preferences* (n=104), and *activity competence* (n=62) categories. Instances had an average length of 14.62 tokens and a maximum length of 144 tokens. To measure annotation quality, we computed the average percent agreement (80.18%) and Cohen's Kappa (κ=0.75) across all classes. Table 3 presents the percent agreement per class and for binary categories.

3.3. Unique Qualities of the Dataset

To our knowledge, this is the first publicly available dataset of solution-focused strategies written by caregivers of young children with developmental needs. It consists of short and topically- and stylistically-varied content, raising complex and intriguing questions about classifying these strategies. To better understand the caregiver strategies represented within each class, we conducted topic modeling using Latent Dirichlet Allocation (LDA) (Blei et al., 2003). We automatically generated 15 topics per class using LDA. These topics were carefully analyzed by a researcher with expertise in pediatric rehabilitation and child participation, who created topic titles for the top three topics for each class based on common themes observed within each topic. Via our topic analyses, we identified specific topics represented within each class as outlined in Table 4. For *activity competence* we only observed two relevant topic titles. For the *no-strategy* class, we did not observe any relevant topic titles, which was expected; instead of content related to participation strategies, the identified topics referred to different aspects of caregiver and child life contexts.

The shared dataset reveals an imbalance among classes (see Table 2). Specifically, the class *environment/context* has the highest number of cases (n=325) while the class *activity competence* has the fewest (n=62). When grouping the entries into broader classes, such as *strategies* and *no-strategies* or *extrinsic* and *intrinsic* strategies, for less complex classification tasks, an imbalance persists between the classes *strategies* (n=642) and *no-strategies* (n=138). However, there is no imbalance between *extrinsic* (n=325) and *intrinsic* strategies (n=317).

4. Proof of Concept

To validate our dataset in settings in which we anticipate its real-world use, we implemented classification models to establish new benchmarks for caregiver strategy detection and classification. In Section 4.1 we describe our preprocessing steps for both classical machine learning and Transformer-based models, and in Section 4.3 we outline our modeling algorithms in detail. In Section 4.4, we describe three classification tasks for use with this dataset, of varying complexity.

4.1. Data Preprocessing

We applied the following preprocessing techniques to the dataset prior to training our models:

- **Spelling Correction:** We used

Environment/Context	
1.	*Parent is involving toys, setting a time and day and showing the child how to do an activity*
2.	*Using pictures, preparing ahead*
3.	*Going to the park or taking the child to the park*
Sense of Self	
1.	*Asking the child to help with an activity (e.g., cleaning, putting dishes away, vacuum cleaning, holding something)*
2.	*Using positive reinforcement and including the child into an activity*
3.	*Encouraging new things, activities, and experiences*
Preferences	
1.	*Including the child's ideas and choosing activities (e.g., play) that a child loves and enjoys*
2.	*Making activities fun*
3.	*Making activities a game and letting the child pick activities or toys*
Activity Competence	
1.	*Working on gross motor and fine motor skills and vision*
2.	*Providing instructions about the activity or hand-over-hand support*

Table 4: Topics per class.

`pyspellchecker`[1] to check and validate the spelling of words in our dataset, and replace errors with suggested corrections (Wankhede et al., 2018).

- **Punctuation Mark Removal:** We removed all punctuation marks from our dataset.
- **Number Replacement:** To prevent our models from overfitting to specific number values or introducing unnecessary noise to the learning process, we replaced all numbers in our dataset with generic *[number]* tokens.
- **Case Normalization:** We converted all text to lowercase to avoid introducing noise associated with capitalization inconsistencies to the learning process.

When training statistical classification models, we applied two extra preprocessing steps to the data. First, we removed stopwords (Kaur and Buttar, 2018) using the NLTK English stopwords list (Bird, 2006). Then, we lemmatized (Plisson et al., 2004) words in our dataset, using NLTK's WordNetLemmatizer, to base forms from the *WordNet* dictionary (Fellbaum, 2010) when available to reduce data sparsity.

4.2. Feature Representations

We represented our feature spaces differently for statistical and Transformer-based models. For statistical models, we used TF-IDF vectors with a vocabulary size of the 5000 most-frequent words in our dataset (Zhang et al., 2011). For Transformer-based models, we used contextual word embeddings generated by the pretrained language model.

4.3. Models

We experimented with both statistical and Transformer-based classification models to validate our dataset, compare model performance, and establish a performance benchmark. Statistical machine learning models often work well for low-resource tasks or problems leveraging smaller training datasets (Meetei et al., 2021), but recently Transformer-based models have also performed competitively in those settings (Cruz and Cheng, 2020; Wu and Dredze, 2020). For our initial model comparison experiments, we performed a multinomial classification of caregiver strategies across our full five-class data distribution, comparing logistic regression, naïve Bayes, BERT, and Bio-ClinicalBERT models to a most-frequent-class baseline. We briefly motivate each below:

- **Logistic Regression (LR):** LR is a straightforward, widely-utilized feature-based model that often achieves strong results for text classification problems (Lee et al., 2006; Ganesan and Subotin, 2014). We applied LR because it has been often used as an approach in related research (Le Glaz et al., 2021).
- **Naïve Bayes (NB):** NB is based on the Bayes' probability theorem (Joyce, 2003) and is often useful for classifying documents and text segments (Kim et al., 2006). We applied NB because it is known to perform well with small datasets (Jurafsky and Martin, 2008).
- **BERT:** BERT is a powerful pre-trained Transformer-based model that has achieved success on a wide variety of NLP tasks, in many cases outperforming previously introduced models (Devlin et al., 2018).
- **Bio-ClinicalBERT:** BERT is pre-trained on general-domain text, but the patterns learned from these data may not generalize well to

[1] `https://pypi.org/project/pyspellchecker`

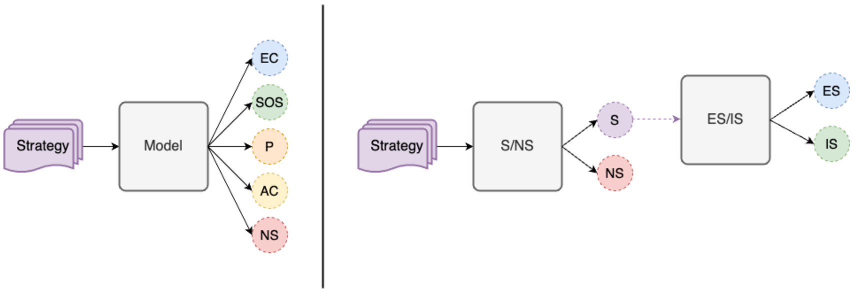

Figure 1: Classification settings considered in our experiments. EC=environment/context; SOS=sense of self; P=preferences; AC=activity competence; S/NS=strategy vs. no-strategy; ES/IS=extrinsic vs. intrinsic strategies.

some clinical problems. Bio-ClinicalBERT (Alsentzer et al., 2019) is a BERT-based model that is pre-trained on two million clinical notes from the MIMIC-III v1.4 database (Johnson et al., 2016). We hypothesized that its use may carry advantages similar to, but perhaps better than, those expected with BERT.

- **Baseline:** Since there is no existing benchmark for our task, to set a performance floor we predicted the most frequent class from the training set for each instance. This permitted us to illustrate the learnability of our dataset and associated task at a higher rate than chance.

We split the data into training (90%) and test (10%) sets to train and evaluate our statistical models. We optimized model parameters via 10-fold cross-validation on the training set (Refaeilzadeh et al., 2009). The best-performing model was then retrained using the full training set and applied to the held-out test set.

For the Transformer-based models, we split the data into training (80%), validation (10%), and test (10%) subsets, training our models on the training set and fine-tuning models and optimizing hyperparameters using the validation set. The collective training and validation sets used for Transformer-based models were the same data used in the the training set for our statistical models, and likewise the test sets were identical for statistical and Transformer-based models. To ensure no unintentional biases, we kept all strategies authored by the same participant in the same set for all models. We trained Transformer-based models using a learning rate of 3e-5 and batch size of two. We trained BERT for three epochs and Bio-ClinicalBERT for four epochs. We used our top-performing model from our model comparison for our later experiments.

4.4. Classification Settings

Our dataset yields high real-world clinical relevance by enabling automated classification of solution-focused caregiver strategies into different rehabilitation-relevant categories. This includes 1) strategy classification into all five classes (i.e., *environment/context*, *sense of self*, *preferences*, *activity competence*, and *non-strategies*), and 2) pipelined filtering of *non-strategies* followed by classification of the remaining strategies into *extrinisic strategies* (i.e., strategies focusing on the *environment/context*) and *intrinsic strategies* (i.e., strategies focusing on a child's *sense of self*, *preferences*, or *activity competencies*). We compared models across both scenarios using 1) the model comparison described for our preliminary experiments, and 2) a pipelined approach where we experimented with a sequential classification paradigm (see Figure 1).

For our pipelined approach, the first *strategy versus no-strategy (S/NS)* task was a binary strategy detection task designed to filter out instances that do not qualify as caregiver strategies. We trained and evaluated our models using a *strategy* class (containing instances from *environment/context*, *sense of self*, *preferences*, and *activity competence*) and a *no-strategy* class. This separation facilitates further analyses in our second pipelined task including only caregiver strategies. Our second *extrinsic versus intrinsic strategy (ES/IS)* classification task was designed to classify the caregiver strategies into the *extrinsic* and *intrinsic* groups defined in Section 3.2. This setting may enable clinicians or downstream applications to begin distinguishing between more specific types of strategies, allowing for decisions to be made without requiring finer-grained precision.

5. Evaluation

We evaluated the performance of our models using accuracy, precision (P), recall (R), and F_1. In Section 5.1, we compare the performance of our

Model	Acc (%)	Precision	Recall	F_1
Baseline	40.78	0.08	0.20	0.11
LR	57.89	0.69	0.43	0.46
NB	53.95	0.85	0.38	0.38
BERT	**64.47**	**0.73**	**0.53**	**0.56**
Bio-ClinicalBERT	53.94	0.71	0.40	0.39

Table 5: Model performance comparison in the five-class multinomial setting. Acc = accuracy; LR = logistic regression; NB = naïve Bayes.

Model	Acc (%)	P	R	F_1
S/NS	88.15	0.82	0.76	0.79
ES/IS	58.06	0.64	0.58	0.53

Table 6: Model comparison for the pipelined steps. S/NS = strategy vs. no-strategy; ES/IS = extrinsic vs.intrinsic strategies; P = precision; R = Recall.

baseline, LR, NB, BERT, and Bio-ClinicalBERT models. We then discuss our findings from the pipelined approach introduced in Section 4.4, using the highest-performing model from Section 5.1.

5.1. Task Validation

5.1.1. Model Comparison

We present the results for each model condition (baseline, LR, NB, BERT, and Bio-ClinicalBERT) in a five-way multinomial classification setting in Table 5. BERT achieves the highest performance overall, with accuracy=64.47%, P=0.73, R=0.53, and F_1=0.56. It outperformed both our highest performing statistical model (LR, by relative performance increases of accuracy=11.36%, P=5.79%, R=23.25%, and F_1=21.73%) and the other Transformer-based classifier (Bio-ClinicalBERT, by relative performance increases of accuracy=19.52%, P=2.81%, R=32.5%, and F_1=43.58%). The BERT model surpassed our most frequent class baseline by percent increases in accuracy, precision, recall, and F_1 of 23.69%, 65.0%, 20.0%, and 28.0%, respectively.

Since our test set was relatively small, we report our per-class prediction results holistically in Figure 2. We observed that predictions were best for the *environment/context* class, followed by the *sense of self* class. The lowest overlap between predicted and true labels was found in the *activity competence* class.

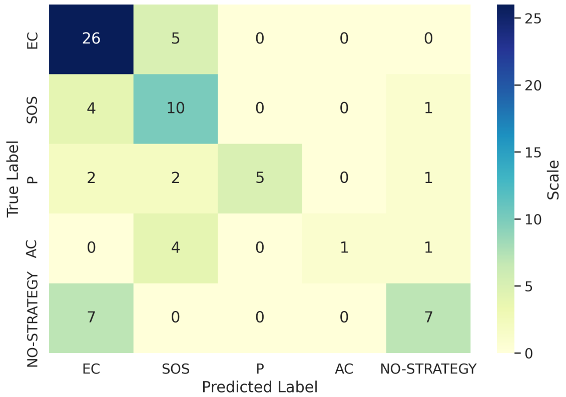

Figure 2: Heatmap with true and predicted labels for the BERT model. EC=environment/context; SOS=sense of self; P=preferences; AC=activity competence.

5.1.2. Pipelined Classification Tasks

We trained and evaluated our highest-performing model, BERT, for the pipelined classification setting introduced in Section 4.4 and present our results in Table 6. We observe that the binary *strategy vs. no-strategy* performance is higher than observed in the multinomial model comparison, with increases in accuracy, precision, recall, and F_1 of 36.73%, 12.32%, 43.39%, and 41.07%. The overall strong accuracy (88.15%) and F_1 (0.79) suggest that our benchmark model is well-equipped to perform binary strategy detection.

For the next stage of the pipeline, we trained and evaluated our model using only caregiver strategies (instances with gold labels of *environment/context*, *sense of self*, *preferences*, and *activity competence*). We observe that our model achieves promising performance at extrinsic strategy vs. intrinsic strategy classification (accuracy=58.06%, P=0.64, R=0.58, and F_1=0.53), although lower than that observed for the strategy vs. no-strategy condition and slightly lower than that observed in the model comparison experiment (Table 5). This may suggest that inclusion of a non-strategy class trivializes the task due to the ease of distinguishing this class from others, whereas conditions without this class remain more complex.

6. Discussion

Existing pediatric rehabilitation tools already solicit for information about strategies and may employ manual approaches to identify suitable goal attainment strategies (Kaelin et al., 2021a), but the increase of users submitting strategies to these tools complicates the task of sorting and selecting relevant strategies and increases caregiver burden (Kaelin et al., 2021a). CareCorpus enables and encourages research on automated strategy classification, filling an existing void. We empirically demonstrate that high-performing Transformer-based **language models can be fine-tuned to capably classify strategies using CareCorpus**, with results for the five-way classification task reaching an accuracy of 64.47% and F_1=0.56. These results are comparable to and slightly higher than prior participation-related classification tasks using videorecordings for engagement detection among children with ASD, which reached an unweighted accuracy=53.90% (Chorianopoulou et al., 2017).

Beyond the clinical need for this work, **CareCorpus creates new opportunity to study rich, open-ended, domain-specialized language in a lower-resourced setting.** Participation and its related constructs are widely recognized as being complex in nature (Imms et al., 2017; World Health Organization, 2001), as supported by our own investigation. For example, BERT outperformed Bio-ClinicalBERT in our experiments, despite the latter's clinical focus. Analyses on Bio-ClinicalBERT's predictions revealed low performance for the *preferences*, *activity competence*, and *no-strategy* classes (see Figure 3), which may be explained by stylistic differences between health-related data obtained from patients staying in critical care units and caregiver-submitted strategies (Johnson et al., 2016). Specifically, data pertaining to the classes *preference* and *no-strategy* are less common in clinical notes (Gaudry et al., 2017; Choong et al., 2018) and are potentially more related to general-domain language. Bio-ClinicalBERT's lower performance on *activity competence* is more surprising from that perspective, but reinforces findings from prior classification tasks for mobility and self-care/domestic life (Newman-Griffis et al., 2021). *Preferences*, *activity competence*, and *no strategy* were all also less frequent in the dataset, and mispredictions for these classes were generally labeled as members of more frequent classes.

The least challenging of our proposed tasks was the strategy detection task. Our own annotators' agreement was also high for this task, further supporting this. Interestingly, macro-averaged F_1 in the initial model comparison experiments (i.e., five-way classification task) was slightly higher than for the *ES/IS* classification (the second stage of the proposed pipelined tasks). This may suggest that the non-strategy class is particularly easy to identify—the per-class F_1 for *non-strategy* is slightly above the average F_1 across other classes (see Figure 2 for additional breakdown of predictions). Thus, inclusion of this class may have resulted in deceptively high macro-averaged scores in the model comparison experiments.

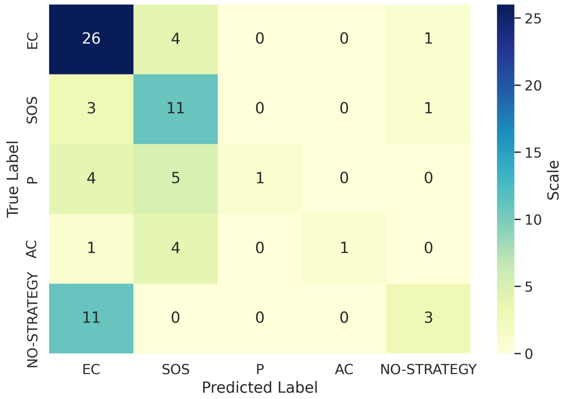

Figure 3: Heatmap with true and predicted labels for the Bio-ClinicalBERT model. EC=environment/context; SOS=sense of self; P=preferences; AC=activity competence.

Overall, results support feasibility of all proposed caregiver strategy classification tasks and general validity of the dataset. From a clinical perspective, components of the pipelined paradigm, such as a model that successfully detects and filters non-strategies, might be sufficient for automatically prompting caregivers towards submitting valid strategies when using electronic pediatric rehabilitation tools. High-performing finer-grained classification models are more suitable for building large-scale end-to-end systems. Our results and subsequent analyses reveal opportunities to improve model performance for both settings. CareCorpus could be leveraged by others, including in the ways outlined here, to optimally support caregivers when searching for solution-focused strategies to help develop pediatric rehabilitation care plans. It could also offer value as an experimental sandbox for other similarly challenging, low-resourced, domain-specific language tasks.

7. Conclusion

In this work, we introduce CareCorpus, a new dataset created by an interdisciplinary research team to support the smart classification of caregiver strategies into established pediatric rehabilitation strategy types. The dataset includes 780 caregiver-submitted strategies collected using a working pediatric rehabilitation tool, each of which are dual-annotated for strategy type by individuals with expertise in participation-focused pediatric rehabilitation. Inter-annotator agreements, reflecting the complexity of constructs related to participation (Imms et al., 2017), together with our validation experiments support the feasiblity of this dataset for learning.

Our findings also provide a framework for future dataset augmentation using various methods to strengthen model performance, which is currently underway. We compare a variety of strategy classification models, achieving a macro-averaged F_1=0.56 when classifying across all strategy categories. This provides evidence of classification feasibility while also highlighting the challenging nature of this task. We also introduce two pipelined classification tasks for use in downstream clinical applications, including a strategy detection task (*S/NS*) and an extrinsic versus intrinsic strategy classification task (*ES/IS*, and empirically validate them. We make our dataset publicly available. Ultimately, this work offers a foundation for follow-up translational efforts geared towards assisting caregivers in developing personalized care plans for their child's pediatric rehabilitation services.

Limitations

In this work, we introduced a publicly available dataset of real-world caregiver strategies categorized into established rehabilitation classes and a *no-strategy* class. Our work is limited by three factors. First, although CareCorpus is larger and more clinically diverse than many datasets in the healthcare domain (Albrecht et al., 2020), it is still relatively small for training deep learning models. Second, the dataset is imbalanced (EC=325 strategies, SOS=151 strategies, NS=138 strategies, P=104 strategies, and AC=62 strategies). Balancing our class distribution may result in improvements for each class, and in turn greater overall model performance. Third, CareCorpus includes only English-language strategies. English is a high-resource language, and it is unclear whether the proposed techniques would yield similarly high performance in less-resourced languages. Thus, the methods and findings within are limited in their scope and may not generalize beyond the English language. In the long term, it is our hope to extend this work to other languages.

Ethics Statement

Like in other healthcare tasks (Kaelin et al., 2022b), real-world data is needed for the development of smart applications for pediatric rehabilitation service delivery. However, protecting the rights and privacy of research participants is critical (Abouelmehdi et al., 2018; Soceanu et al., 2015). Multi-institutional ethics approval was obtained prior to participant recruitment and remains active for this study. Per IRB protocol, we ensured that each participant provided consent for study participation and was informed about their rights to withdraw their participation at any time. Participants were reimbursed with a $10 gift card. We also de-identified our dataset by eliminating any usernames or other identifying data, and replacing any names written directly within instances with generic name tokens (i.e., *[name]*). Data is available in the Inter-university Consortium for Political and Social Research (ICPSR) portal (Kaelin et al., 2023b). The intended use for the dataset is to analyze solution-focused caregiver strategies such as by exploring automated classification to support rehabilitation service provision. Part of this dataset can be linked to another dataset with restricted access. ICPSR requests IRB approval for researchers to access this additional linked dataset to ensure research done with this dataset aligns with ethical regulations and principles.

8. Acknowledgements

Data for this study were collected with funding from the National Institutes of Health (NIH grant number 1R03HD084909-01A1; PI: M. Khetani) and the American Occupational Therapy Foundation (AOTF grant number AOTFIR20KHETANI; PI: M. Khetani). Analyses of these data were supported by the National Science Foundation (grant abstract number W8XEAJDKMXH3; PIs: N. Parde, M. Khetani, J. Dooling-Litfin), the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR grant number 90SFGE0032-01-00; PI: V. Kaelin), and the University of Illinois Chicago. The content is the responsibility of the authors and does not necessarily represent the official views of these funding agencies. We thank Vivian Villegas, Dianna Bosak, and Zurisadai Salgado from the Children's Participation in Environment Research Lab (CPERL) at University of Illinois Chicago for assisting with data annotation. We also thank Vivian Villegas from CPERL and Shahla Farzana from the Natural Language Processing Lab (NLP Lab) at University of Illinois Chicago, for their feedback on a prior version of this manuscript.

9. Bibliographical References

Karim Abouelmehdi, Abderrahim Beni-Hessane, and Hayat Khaloufi. 2018. Big healthcare data: Preserving security and privacy. *Journal of Big Data*, 5(1):1–18.

Erin C. Albrecht, Vera C. Kaelin, Briana L. Rigau, Jodi K. Dooling-Litfin, Elizabeth A. Scully, Natalie J. Murphy, Beth M. McManus, and Mary A. Khetani. 2020. Pilot implementation of an electronic patient-reported outcome measure for planning and monitoring participation-focused care in early intervention. *BMC Medical Informatics and Decision Making*, 20(1):1–11.

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical BERT embeddings. *CoRR*, abs/1904.03323.

Jeremy Beaudry, Alyssa Consigli, Colleen Clark, and Keith J. Robinson. 2019. Getting ready for adult healthcare: Designing a chatbot to coach adolescents with special health needs through the transitions of care. *Journal of Pediatric Nursing*, 49:85–91.

Steven Bird. 2006. Nltk: The natural language toolkit. In *Proceedings of the COLING/ACL on Interactive Presentation Sessions*, COLING-ACL '06, page 69–72, Sydney, Australia. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(0):993–1022.

Dianna L Bosak, Jessica M Jarvis, and Mary A Khetani. 2019. Caregiver creation of participation-focused care plans using Participation and Environment Measure Plus (PEM+), an electronic health tool for family-centred care. *Child: Care, Health and Development*, 45(6):791–798.

Thamar J.H. Bovend'Eerdt, Rachel E. Botell, and Derick T. Wade. 2009. Writing smart rehabilitation goals and achieving goal attainment scaling: A practical guide. *Clinical Rehabilitation*, 23(4):352–361.

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5599–5611, Florence, Italy. Association for Computational Linguistics.

Karen Choong, Douglas Fraser, Samah Al-Harbi, Asm Borham, Jill Cameron, Saoirse Cameron, Ji Cheng, Heather Clark, Tim Doherty, Nora Fayed, et al. 2018. Functional recovery in critically ill children, the "weecover" multicenter study. *Pediatric Critical Care Medicine*, 19(2):145–154.

Arodami Chorianopoulou, Efthymios Tzinis, Elias Iosif, Asimenia Papoulidi, Christina Papailiou, and Alexandros Potamianos. 2017. Engagement detection for children with autism spectrum disorder. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5055–5059, New Orleans, USA. IEEE.

Jan Christian Blaise Cruz and Charibeth Cheng. 2020. Establishing baselines for text classification in low-resource languages. *arXiv preprint arXiv:2005.02068*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Shahla Farzana, Mina Valizadeh, and Natalie Parde. 2020. Modeling dialogue in conversational cognitive health screening interviews. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1167–1177, Marseille, France. European Language Resources Association.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: Computer applications*, pages 231–243. Springer.

Kavita Ganesan and Michael Subotin. 2014. A general supervised approach to segmentation of clinical texts. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 33–40, Washington DC, USA. IEEE.

Stéphane Gaudry, Jonathan Messika, Jean-Damien Ricard, Sylvie Guillo, Blandine Pasquet, Emeline Dubief, Tanissia Boukertouta, Didier Dreyfuss, and Florence Tubach. 2017. Patient-important outcomes in randomized controlled trials in critically ill patients: A systematic review. *Annals of Intensive Care*, 7(1):1–11.

Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Bing Liu, Ben Gerber, and Lisa Sharp. 2020. Goal summarization for human-human health coaching dialogues. In *The Thirty-Third International Flairs Conference*, North Miami Beach, USA. AAAI.

Christine Imms, Mats Granlund, Peter H. Wilson, Bert Steenbergen, Peter L. Rosenbaum, and Andrew M. Gordon. 2017. Participation, both a means and an end: A conceptual analysis of processes and outcomes in childhood disability. *Developmental Medicine & Child Neurology*, 59(1):16–25.

Jessica M. Jarvis, Andrea Gurga, Alexa Greif, Heather Lim, Dana Anaby, Rachel Teplicky, and Mary A. Khetani. 2019. Usability of the Participation and Environment Measure Plus (PEM+) for client-centered and participation-focused care planning. *The American Journal* *of Occupational Therapy*, 73(4):7304205130p1–7304205130p8.

Jessica M. Jarvis, Vera C. Kaelin, Dana Anaby, Rachel Teplicky, and Mary A. Khetani. 2020. Electronic participation-focused care planning support for families: A pilot study. *Developmental Medicine & Child Neurology*, 62(8):954–961.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

James Joyce. 2003. Bayes' theorem. In *The Stanford Encyclopedia of Philosophy (Fall 2021 Edition)*. The Metaphysics Research Lab, Philosophy Department, Stanford University.

Daniel Jurafsky and James H. Martin. 2008. *Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing*. Upper Saddle River, NJ: Prentice Hall.

Vera Kaelin, Vivian Villegas, Yi-Fan Chen, Natalie Murphy, Elizabeth Papautsky, Jodi Litfin, Natalie Leland, Varun Maheshwari, Beth McManus, and Mary Khetani. 2022a. Effectiveness and scalability of an electronic patient-reported outcome measure and decision support tool for family-centred and participation-focused early intervention: PROSPECT hybrid type 1 trial protocol. *BMJ open*, 12(1):e051582.

Vera C. Kaelin, Dianna L. Bosak, Vivian C. Villegas, Christine Imms, and Mary A. Khetani. 2021a. Participation-focused strategy use among caregivers of children receiving early intervention. *The American Journal of Occupational Therapy*, 75(1):7501205090p1–7501205090p11.

Vera C. Kaelin, Andrew D. Boyd, Martha M. Werler, Natalie Parde, and Mary A. Khetani. 2023a. Natural language processing to classify caregiver strategies supporting participation among children and youth with craniofacial microsomia and other childhood-onset disabilities. *Journal of Healthcare Informatics Research*, 7:480–500.

Vera C. Kaelin, Mina Valizadeh, Zurisadai Salgado, Natalie Parde, and Mary A. Khetani. 2021b. Artificial intelligence in rehabilitation targeting the participation of children and youth with disabilities: Scoping review. *Journal of Medical Internet Research*, 23(11):e25745.

Vera C. Kaelin, Mina Valizadeh, Zurisadai Salgado, Julia G. Sim, Dana Anaby, Andrew D. Boyd, Natalie Parde, and Mary A. Khetani. 2022b. Capturing and operationalizing participation in pediatric re/habilitation research using artificial intelligence: A scoping review. *Frontiers in Rehabilitation Sciences*, 3:48.

Vera C. Kaelin, Mina Valizadeh, Elizabeth A. Scully, James E. Graham, Beth McManus, Natalie Parde, and Mary A. Khetani. 2023b. *Early Intervention Colorado (EI-CO) Participant Characteristics, Service Use, and Patient-Reported Outcomes, Colorado, 2017-2021*. Inter-university Consortium for Political and Social Research [distributor].

Jashanjot Kaur and P. Kaur Buttar. 2018. A systematic review on stopword removal algorithms. *International Journal on Future Revolution in Computer Science & Communication Engineering*, 4(4):207–210.

Mustafa Khanbhai, Patrick Anyadi, Joshua Symons, Kelsey Flott, Ara Darzi, and Erik Mayer. 2021. Applying natural language processing and machine learning techniques to patient experience feedback: A systematic review. *BMJ Health & Care Informatics*, 28(1):e100262.

Mary A Khetani, Anna B Cliff, Cathy Schelly, Lisa Daunhauer, and Dana Anaby. 2015. Decisional support algorithm for collaborative care planning using the Participation and Environment Measure for Children and Youth (PEM-CY): a mixed methods study. *Physical & Occupational Therapy in Pediatrics*, 35(3):231–252.

Mary A Khetani, Ellen S Cohn, Gael I Orsmond, Mary C Law, and Wendy J Coster. 2013. Parent perspectives of participation in home and community activities when receiving part C early intervention services. *Topics in Early Childhood Special Education*, 32(4):234–245.

Mary A Khetani, Heather K Lim, and Marya E Corden. 2017. Caregiver input to optimize the design of a pediatric care planning guide for rehabilitation: descriptive study. *JMIR rehabilitation and assistive technologies*, 4(2):e7566.

Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng. 2006. Some effective techniques for naive bayes text classification. *IEEE Transactions on Knowledge and Data Engineering*, 18(11):1457–1466.

Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan Devylder, Michel Walter, Sofian Berrouiguet, et al. 2021.

Machine learning and natural language processing in mental health: systematic review. *Journal of Medical Internet Research*, 23(5):e15708.

Su-In Lee, Honglak Lee, Pieter Abbeel, and Andrew Y. Ng. 2006. Efficient l~ 1 regularized logistic regression. In *AAAI*, volume 6, pages 401–408, Boston, USA.

Peter Ljunglöf, Britt Claesson, Ingrid Mattsson Müller, Stina Ericsson, Cajsa Ottesjö, Alexander Berman, and Fredrik Kronlid. 2011. Lekbot: A talking and playing robot for children with disabilities. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 110–119, Edinburgh, UK. Association for Computational Linguistics.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, Samir Kumar Borgohain, and Sivaji Bandyopadhyay. 2021. Low resource language specific pre-processing and features for sentiment analysis task. *Language Resources and Evaluation*, 55(4):947–969.

Denis Newman-Griffis, Jonathan Camacho Maldonado, Pei-Shu Ho, Maryanne Sacco, Rafael Jimenez Silva, Julia Porcino, and Leighton Chan. 2021. Linking free text documentation of functioning and disability to the icf with natural language processing. *Frontiers in Rehabilitation Sciences*, 2:1–17.

Joël Plisson, Nada Lavrac, and Dunja Mladenic. 2004. A rule based approach to word lemmatization. In *Proceedings of IS04*, volume 3, pages 83–86.

Payam Refaeilzadeh, Lei Tang, and Huan Liu. 2009. Cross-validation. *Encyclopedia of Database Systems*, 5:532–538.

José A. Reyes-Ortiz, Beatriz A. González-Beltrán, and Lizbeth Gallardo-López. 2015. Clinical decision support systems: A survey of nlp-based approaches from unstructured data. In *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)*, pages 163–167, Valencia, Spain. IEEE.

Wing-Chee So, Chun-Ho Cheng, Wan-Yi Lam, Ying Huang, Ka-Ching Ng, Hiu-Ching Tung, and Wing Wong. 2020. A robot-based play-drama intervention may improve the joint attention and functional play behaviors of chinese-speaking preschoolers with autism spectrum disorder: A pilot study. *Journal of Autism and Developmental Disorders*, 50(2):467–481.

Alexandru Soceanu, Maksym Vasylenko, Alexandru Egner, and Traian Muntean. 2015. Managing the privacy and security of ehealth data. In *2015 20th International Conference on Control Systems and Computer Science*, pages 439–446, Bucharest, Romania. IEEE.

Mina Valizadeh and Natalie Parde. 2022. The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660, Dublin, Ireland. Association for Computational Linguistics.

Vivian C Villegas, Dianna L Bosak, Zurisadai Salgado, Michelle Phoenix, Natalie Parde, Rachel Teplicky, Mary A Khetani, and High Value Early Intervention Research Group Kuznicki L. Pedrow A. Howell A. 2023. Diversified caregiver input to upgrade the Young Children's Participation and Environment Measure for equitable pediatric re/habilitation practice. *Journal of Patient-Reported Outcomes*, 7(1):87.

Shreyas Wankhede, Ranjit Patil, Sagar Sonawane, and Ashwini Save. 2018. Data preprocessing for efficient sentimental analysis. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 723–726, Coimbatore, India.

World Health Organization. 2001. *International Classification of Functioning, Disability and Health*. World Health Organization, Geneva, Switzerland.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093*.

Aida Zhanatkyzy, Zhansaule Telisheva, Aizada Turarova, Zhanel Zhexenova, and Anara Sandygulova. 2020. Quantitative results of robot-assisted therapy for children with autism, adhd and delayed speech development. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 541–542, Cambridge, UK. Association for Computing Machinery.

Wen Zhang, Taketoshi Yoshida, and Xijin Tang. 2011. A comparative study of tf* idf, lsi and multi-words for text classification. *Expert Systems with Applications*, 38(3):2758–2765.

Ayah Zirikly, Bart Desmet, Denis Newman-Griffis, Elizabeth E Marfeo, Christine McDonough, Howard Goldman, Leighton Chan, et al. 2022. Information extraction framework for disability determination using a mental functioning use-case. *JMIR Medical Informatics*, 10(3):e32245.