A Transformer-based Model for Older Adult Behavior Change Detection

Fateme Akbari
Information Systems
McMaster University
Hamilton, ON, Canada
akbarif@mcmaster.ca

Kamran Sartipi
Department of Computer Science
East Carolina University
Greenville, NC, USA
sartipik16@ecu.edu

***Abstract*—The advancements in smart home technologies have created new opportunities for precise monitoring of the older adults' daily activities to provide timely care, predict health problems, and promote independent living at home. Many recent studies have investigated abnormality detection in Activities of Daily Living (ADL), and some used deep learning methods to handle the problem. In this paper, we leverage Bi-Directional Encoder Representations from Transformers (BERT), as a state-of-the-art method in machine learning, to analyze older adults' ADL sequences. Due to the fine-tuning capability of transformers, they are a good fit for supervised tasks when a large labeled training dataset is not available. Their architecture also allows for Parallel Computation, which is important for (near) real-time abnormality detection. To the best of our knowledge, this is the first effort to represent the older adult's daily behavior as sequences of ADLs with the goal of applying transformers to the behavior change detection problem. We designed four experiments to illustrate the capability of Transformers in detecting individuals' behavior abnormalities. We conducted a case study on a two-resident ADL dataset to evaluate the model in four experiments. Our results show that a BERT-based classifier can effectively detect behavior abnormalities from sequences of ADLs. Also, transfer learning proved to be helpful when it comes to fine-tuning a pre-trained model for a new resident.**

Keywords: ADL, Smart Home, Sensor Network, Transformer, BERT, Sequence Classification, Anomaly Detection, Change Detection.

I. Introduction

The advances in sensor technology has facilitated the enhancement of smart home applications [1]. Smart homes allow for monitoring of older adults' activities in order to make them as independent as possible in their everyday lives and, ultimately, to reduce unnecessary hospitalization or readmissions [2]. For the detection of ADLs, accurate sensor data is required. Obtrusive and inconspicuous sensors are the two types of sensors that can be distinguished. Wearable sensors, sometimes known as obtrusive sensors, track precise location and can be used for health monitoring. Farivar et al. [3] looked into the use of wearable sensing devices among older adults. The difficulty of dealing with these technologies is a disincentive to older adults' adoption decisions, according to the results of their online survey and interviews. The findings also revealed that an older adult's subjective well-being, which is a self-reported measure of well-being, adversely moderates the effect of cognitive age on their use intention, indicating that when an older adult's subjective well-being is poor, cognitive age increases their intention to use the gadget [3]. These findings highlight the importance of employing sensors to monitor the everyday activities of older people who require ongoing care.

Unobtrusive sensors, on the other hand, can identify an individual's interaction with household objects as well as with other subjects and measure the changes in the environment. Unobtrusive sensors are less restrictive and do not need any client-side coordination because data can be collected without the resident's involvement once they are deployed (no need to be worn). Video-based activity monitoring methods have been adopted by some researchers [4] with the justification that they are less obtrusive and do not interfere with daily activities. However, these methods are likely to pose security issues. Also, it is time-consuming to analyze videos making this technology unsuitable for systems that require real-time analysis. Overall, in order to have the most accurate picture of older persons' activities, an integrated sensor network platform using both wearable and non-wearable sensors is required.

There are currently a number of smart home projects that have been developed worldwide [5]–[10]. Data collected from such testbeds has been used in two types of sensory data analysis, lower sensory level and higher activity level. The lower-level analysis is the classification of activities based on sensory readings, while the derived activity labels are processed for further decision-making support in higher-level behavior analysis. Supervised or unsupervised learning is used to learn or discover activity annotations in lower-level analysis, depending on the availability of labeled data. When activity labels are established, these activity labels are reused for further study of anomaly detection or for prompting system reminders in higher level analysis [11].

Detecting change in older adult behavior will give healthcare providers the ability to constantly monitor their health condition and provide medical instructions or prescriptions for the patient. It can also be used to supplement practitioners' knowledge of palliative care by detecting symptom escalation and functional deterioration in real time and aiding proactive therapy. In principle, sensor technologies can substitute for

This work was partially funded by

some caregiving time, which is normally used to monitor the older adult. The decreased time is expected to minimize caregiver stress and improve quality of life for both family caregivers and older adults. In addition, because activities of daily living (ADL) are good indicators of health status, activity monitoring could be used over the long term to detect gradual deterioration in the health status of older adults, which could threaten their ability to live independently.

In this paper, we leverage Transformers in analyzing older adults' ADL sequences. Transformers generate superior results in Natural Language Processing, where they are originally presented and applied [12]. Transformers are used for supervised classification tasks such as sentiment analysis as well as sequence to sequence processing such as in language translation. They are also leveraged in other domains such as system log analysis [13]–[15] and IoT [16]. The followings are the properties of the Transformers that lead us to assume they are a good fit for older adult ADL analysis:

Pre-training: Transformers can be pre-trained on large datasets, then fine-tuned for a given task on another (perhaps smaller) dataset. Thus, Transformers are a good fit for supervised tasks when we don't have access to a large, labelled training dataset.

Parallel Computation: this functionality allows us to use transformers in (near) real-time abnormality detection. While RNN models operate sequentially, Transformers process the entire input at once allowing for parallel processing.

Positional Embedding: when it comes to encoding the tokens in a sequence, this technique plays a critical role in enabling Transformers to embed context in the encoding process.

Bi-Directional Training: this feature allows for the learning of the interrelationships of tokens in a sequence based on both pre- and post-tokens, resulting in a more accurate model.

The main contributions of this research are as follows: (1) We have introduced a new application for Transformers. To the best of our knowledge, this is the first effort to represent daily behavior as sequences of ADLs to apply transformers to the behavior change detection problem. (2) We design four experiments to illustrate the capability of Transformers in detecting behavior abnormalities. (3) We conduct a case study on a two-resident ADL dataset to evaluate the models in the four experiments.

The rest of this paper is laid out as follows. We look at the related work in Section II. We provide some background information about state-of-the-art machine learning methods for sequence classification in Section III. In Section IV, we discuss how to apply BERT transformers to the problem of behavior change detection in older adults. A case study, Section V, on a two-resident ADL dataset is conducted to showcase the capability of the model in predicting behavior abnormalities. Finally, Sections VI and VII include discussion and concluding thoughts, as well as an introduction to the potential future studies.

II. Related Work

Scholars have extensively used machine learning methods to analyze ADLs with the goal of providing on-time care and predicting older adults' health conditions. Many studies have benefited from the availability of datasets on daily activities, including the use of machine learning methods for predicting/detecting anomalous behaviour [17]–[20], the development of reminder and recommender systems in healthcare support, and the supervision of long-term behaviour [21]–[23].

Arifoglu and Langenspepen [12] present an algorithm for learning health changes based on the correlation of context-enriched frequent behaviour patterns and cognitive and physical health deterioration. Although the sequence of activities is taken into account in their work, it is only for short-term behaviour patterns. Moallem et al. [6] presented an anomaly detection method in smart homes based on deep learning. They use binary sensor data to train a predictor model, a recurrent neural network, to predict which sensors will turn on/off and how long the event will last.

Karakostas et al [24] present an anomaly detection approach in which the predicted user activity is represented by a task model. The predicted and actual behaviour are then compared to see if any variance (anomaly) has occurred. The problem with such model-based anomaly detection approaches is that they fail to detect anomalies that have not previously occurred.

Fahad et al [8] propose a method for detecting behaviour anomalies by taking into account two types of abnormality: missing or extra sub-events in an activity and unusual duration of the activity. They trained an H2O model to classify events using labelled activities (normal, anomaly). The main problem with such supervised models is that they must be trained using labelled data, which is time-consuming and difficult to generate.

In [25], a Long-Short Term Memory(LSTM)-based method for detecting anomalies in daily activity sequences is proposed, as well as a comparison of the proposed method with the Hidden Markov Model, which demonstrates comparable results for the LSTM model.

By simplifying the activity prediction problem to a regression learning problem, Ismail et al [26] present a novel solution. They then provide evaluation metrics for the proposed activity predictor. Finally, they demonstrate the applicability of their method by embedding the activity predictor in an activity prompter service, demonstrating the reliability of their approach.

Hochreiter and Schmidhuber [27] propose a context-aware framework for learning and predicting human behavior. Behavior contexts such as weekday and time of day are collected from residents' real-life data to improve the accuracy of activity prediction.

Although behavior abnormality detection has been the focus of many studies by analyzing sensor data, there is a gap in training models that are able to transfer learning from a model, which is trained on a resident's data, to predicting abnormalities in other residents. In this study, we try to close this gap.

III. Background

in this section we provide an overview of state-of-the-art machine learning methods in sequence classification.

A. Recurrent Neural Networks (RNN)

Recurrent Neural Networks including LSTMs and Gated Recurrent Units (GRUs) process sentences word by word. The concept of hidden state is introduced for retaining past information. When processing a word in a sentence, the hidden state of the previous word is required to encode the current word. That is why RNN models cannot be trained in parallel. Also, the forget gate in LSTM architecture is designed to allow for filtering out information that is unrelated in order to incorporate a long-term memory into the model.

To further mitigate the problem of capturing long-term dependencies, some scholars have used Bi-Directional LSTM models, which encode the same sentence from two directions, i.e., from start to end and from end to start. Still, there are issues with these LSTM models that limits their performance in analyzing sequences. LSTMs do not perform well dealing with long sequences in terms of capturing long-term dependencies. The reason is that the probability of keeping the context from a word that is far away from the current word which is being processed decreases exponentially with the distance from it.

B. Transformers

To resolve some of the above problems, researchers have created a technique for paying attention to specific words [12]. Transformers have eliminated the need for recursion by introducing the characteristics described below:

Non-sequential Training: sentences are processed as a whole rather than word by word.

Self-Attention: this is the newly introduced 'unit' used to compute similarity score of each word in a sentence with a current word, and hence it models the dependencies between the tokens of the sequence. Simply put, each word in the sequence pays attention to other words in the same sequence, and therefore, captures the relationships between them.

Positional Embedding: this is another innovation introduced to replace recurrence in RNN techniques. The idea is to use learned weights which encode information related to the position of a token in a long sentence.

As opposed to RNN models, Transformers do not rely on past hidden states to capture dependencies with previous words. Instead, they process a sentence as a whole, which does not need the hidden state and forget gate mechanism. Therefore, there is no risk of losing (or 'forgetting') past information [28]. In other words, at each step the algorithm has direct access to all the other steps (self-attention), which leaves practically no room for information loss. On top of that, we can look at both future and past elements at the same time, which also brings the benefit of Bi-Directional RNNs, without need for double computation. And of course, all this happens in parallel (non-recurrent), which makes training much faster. Moreover, the Transformers use multi-head attention which is a technique to provide the opportunity to capture the relationship among different words from different (or multiple) aspects. Finally, positional embedding enables the model to differentiate between the same word appearing in different positions (or contexts).

It should also be noted that Transformers can only capture dependencies within pre-defined and fixed-size input words for training. That is, if a maximum sentence size is set to 50, the model will not be able to capture dependencies between the first word of a sentence and those that occur beyond the 50 words, which may be in the next paragraph.

C. BERT

Bidirectional Encoder Representations for Transformers (BERT) are standard building blocks for training task-specific Natural Language Processing (NLP) models [29]. BERT models, which are pre-trained on web-domain huge text corpus, are the focus of many task-specific NLP problems [30]–[32]. Pre-trained BERT models have been proven to be effective when they are fine-tuned for specific tasks using domain-specific training data [33].

In a BERT model, the input consists of text spans, such as sentences separated by special tokens [SEP]. Using Byte-Pair Encoding (BPE) [34] a small set of sub-words that can compactly form all words in the given corpus are first identified using a greedy algorithm. The text corpus and vocabulary may preserve the character's case (cased) or convert all characters to lowercase (uncased). The input token sequence is first processed by a lexical encoder, which combines a token embedding, a (token) position embedding and a segment embedding (i.e., which text span the token belongs to) by element-wise summation. This embedding layer is then passed to multiple layers of Transformer modules [12]. In each Transformer layer (self-attention head), contextual representations of tokens are generated. This is done by calculating a non-linear transformation of tokens' representations via multiplying the embedded input by three matrices, Query, Key, and Value. The matrices get trained through the learning mechanism of the model. The final layer outputs contextual representations for all tokens, which combines information from the entire text span.

A Masked Language Model (MLM) task and/or a Next Sentence Prediction (NSP) task are used for BERT pre-training. The MLM randomly replaces a subset of tokens by a special token, [MASK], and asks the language model to predict them. The training objective is to minimize the cross-entropy loss between the original tokens and the predicted ones. NSP is used for pre-training the model by training it to predict the sentence that follows each sentence in the training corpus.

IV. Approach

In this section, we introduce our approach for behavior change detection using BERT encoders. First, we explain how we model an individual's daily behavior based on sensor data in a way that BERT layers can generate the output. Then, we elaborate on the BERT model suggested for predicting potential abnormalities in older adult behavior.

A. Behavior Representation

Older adult indoor behavior needs to be modeled for relatively unconstrained environments. We assume that activities are carried out in a certain order in our model. As a result, for the sake of simplicity, we ignore concurrent activities. When a behaviour is viewed as a series of discrete tokens/events (for example, sleeping, eating, watching TV, and cooking), two essential quantities emerge: i) *acts*: the acts that make up a behaviour; and ii) *Order*: the order in which the activities are performed.

In our study, we use the concept of tokenizing behaviour in the same manner that Natural Language Processing (NLP) experts have looked at documents as vectors of their constituent words through VSM (Vector Space Model). VSM [35] is a technique that efficiently captures the content of a sequence, but its methods disregard word order entirely. However, natural activity-orderings, not only activity content, characterise behaviour. As a result, a model is needed that explicitly captures activity order. In our representation model, we split the data into days. After that, each day is depicted as a series of events. Each sequence appears to have a different length than the others. Because the daily sequences are fed into the BERT model in the following stage and the BERT model only accepts input sequences of the same length, a maximum sequence length must be set and padding tokens are added to the end of each sequence to ensure that they are all the same length. In this study, we use an ordered sequence of events to model the human behaviour B:

$$B = e_1, e_2, ..., e_i, ..., e_W \tag{1}$$

where ei denotes an event. To define events, we borrow from the NLP corpus vocabulary concept. We consider two ADL features to distinguish events: type t_i and duration d_i:

$$\begin{aligned} &e_i = Concatenate(t_i, d_i); \\ &where\ \ t_i \in \{activity\ \ types\}\ and \\ &d_i \in \{Short, Medium, Long\} \end{aligned} \tag{2}$$

Because the model works with categorical data, we discretize the values for activity duration. We believe that, while it does not affect the model's correctness, it simplifies it by reducing the state space. First, we normalise the duration of each activity type separately (using only the training data) as the range of time in different activity types differs, as shown in Fig. 1. The duration values are then converted to category values using an equal width discretization approach.

B. BERT for Behavior Change Detection

The BERT model can now accept ADL sequences as input because daily behaviour is now shaped as ADL sequences, comparable to text corpus.

As illustrated in Figure 1, the ADL sequences are fed into the BERT model, which first tokenizes them into constituent tokens. Tokenized input is then transformed into tensors of numbers. In order for the tensors to represent tokens' context as well as tokens' content, the embedding process is different from conventional embedding methods such as Word2Vec [36]. BERT embedding uses positional embedding to incorporate the context of a token into the vector. As a result, a brief nap in the middle of the day is coded differently than a short nap at night, because the position of the token is taken into account when embedding. While the token in both cases is "Sleep", they will be embedded differently.

The embedded sequences pass through the encoding layer with attention heads to encode the entire sequence in the next layer. When processing each token, the self-attention head looks for interrelationships between the token and other tokens. For each word, the Query, Key, and Value vectors are calculated by multiplying the embedded vector by three matrices that we trained during the training process. Then, all of the other tokens in the input sequence will be scored against this token to represent the degree of association between the token and other tokens in the sequence.

We use the BERT-Base-uncased architecture for our implementations, which has 12 attention heads. Finally, a standard classifier, such as a logistic regression classifier, receives the encoded sequence and produces the sequence label, which specifies whether the ADL sequence is a potential anomaly or not.

Transfer Learning is a key advantage of using BERT models in sequence analysis. The idea is that BERT models can be pre-trained on a dataset consisting of general-domain corpus of text. The pre-trained BERT can then be fine-tuned on a certain domain-specific data for a specific task. Existing research suggests that transfer learning is effective in BERT models [29]. However, a recent study questions the effectiveness of transfer learning when it comes to using pre-trained BERT models for domains with a high percentage of exclusive vocabulary such as the biomedical domain [33].

Transfer learning, on the other hand, can be useful for training the BERT model using datasets obtained from multiple residents' ADLs and then fine-tuning the BERT for a specific resident in a shorter period of time than training the model from scratch. As a result, for a new resident, the BERT model does not need to be trained from scratch. This feature of the model considerably improves the model's generality. In the following section, data are presented on running our model in two different settings and comparing the outcomes, keeping these two characteristics of transfer learning in mind: (1) Training the BERT model from scratch and fine-tuning on each resident's data. (2) Using pre-trained BERT models and fine-tuning on each resident's ADL sequence data.

V. Case Study

In this section, we discuss results from our case study to show the efficiency of our proposed method in detecting behavior changes in older adults. First, we introduce the dataset. Then, we discuss the evaluation metrics and evaluation process. Finally we present the results of our work.

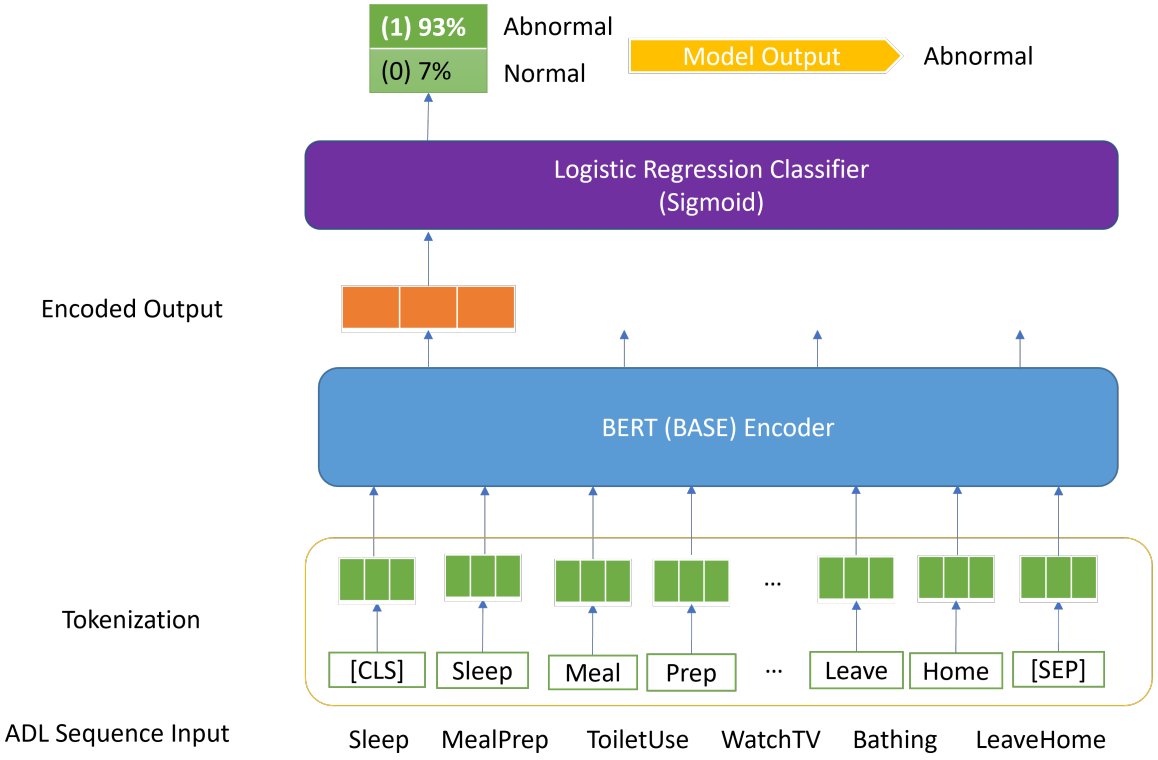

Fig. 1. The Architecture of the BERT Model for Behavior Change Detection

A. Dataset

To evaluate the proposed BERT classifier for behavior change detection, we chose the CASAS-Twor2010 dataset [10] which consists of normal daily activities that two residents, R1 and R2, performed in the WSU smart apartment testbed during the 2009-2010 academic year. A few examples from this dataset are shown in Table I. In this dataset, thirteen types of indoor activities were recorded, such as: *Bathing, Bed-Toilet-Transition, Eating, Enter-Home, Housekeeping, Leave-Home, Meal-Preparation, Personal-Hygiene, Sleep, Sleeping-Not-in-Bed, Wandering-in-Room, Watch-TV*, and *Work*. These activities were recorded using motion sensors, door sensors and temperature sensors. As shown in Table I, start and end times for each activity were recorded, making it possible to calculate the duration of the activity. Also, the time ordering of activities was captured. As there is no overlap in the times of activities performed, we can conclude that concurrent activities were not considered. The CASAS-Twor2010 dataset has 2,804,813 records which represent a total of 3744 activities comprising 1903 activities of resident R1 and 1841 activities of resident R2.

As the original data are not labeled, in order to use it for training of the supervised change detection model, we injected samples of behavior abnormalities by rearranging ADLs and manipulating activity duration. For example, while in the original ADL sequences eating occurs after meal preparation, we reversed the ADLs' order to inject partially misordered sequences. We also created some random abnormalities by shuffling the ADLs. We also intentionally made the abnormal records frequent (i.e., oversampling) in order to avoid the imbalanced data issues.

B. Evaluation Metrics

In this section, we indicate the metrics we used for evaluating the results of the BERT model for detecting behavior changes. Accuracy is a widely-used metric for measuring the accuracy of a prediction. It is computed by the sum of true predictions divided by the total predictions (See Formula 3(a)). While the model's ability to distinguish positive and negative classes can be measured by accuracy, it is not merely enough to measure the efficiency of a predictor model.

The first issue with the accuracy metric is that it gives equal importance to all classes. In problems that predicting one class is of more importance than the other's, such as anomaly detection, it is required to use other evaluation metrics such as recall and precision.

In Precision, the focus is on the positive class predictions (i.e., true positive *TP* and false positive *FP*) as shown in Formula 3(b). If the model predicts negative class poorly (i.e., false negative *FN*), it would not be caught by the Precision result. Also, if the data is imbalanced, Precision would not be sufficient for evaluation. Recall, which can be calculated from Formula 3(c), takes into account the false negatives, which are very important in fraud detection, anomaly detection, etc.

TABLE I
EXAMPLE DATA FROM CASAS-TWOR2010 DATASET.

Date	Time	Sensor ID	Sensor State	Activity
24-8-09	00:15:25	M034	ON	R2_Sleep begin
24-8-09	00:15:27	M047	OFF	
.	.	.	.	
24-8-09	00:16:27	M047	ON	R1_Sleep end
24-8-09	00:16:29	M048	ON	R1_Wandering_in_room begin
.	.	.	.	
24-8-09	00:23:44	M048	OFF	
24-8-09	00:23:52	M048	ON	R1_Wandering_in_room end
24-8-09	00:23:53	M047	ON	R1_Sleep begin
24-8-09	00:23:53	M046	ON	
.	.	.	.	
24-8-09	06:32:46	P001	507	R1_Sleep end
24-8-09	06:32:46	D005	CLOSE	R1_Personal_Hygiene begin
24-8-09	06:32:47	M038	OFF	
.	.	.	.	
24-8-09	06:37:48	M040	OFF	R1_Personal_Hygiene end
24-8-09	06:38:22	P001	579	R1_Bathing begin
24-8-09	06:39:08	T004	20.5	
.	.	.	.	
24-8-09	06:51:00	M040	OFF	
24-8-09	06:51:02	P001	5053	R1_Bathing end
24-8-09	06:51:04	M038	OFF	R1_Personal_Hygiene begin
.	.	.	.	
24-8-09	06:54:37	D005	OPEN	R1_Personal_Hygiene end
24-8-09	07:07:50	M034	OFF	R2_Sleep end
.	.	.	.	
24-8-09	07:07:57	M038	ON	R2_Personal_Hygiene begin
.	.	.	.	
24-8-09	07:08:45	M019	ON	R1_Meal_Preparation end
.	.	.	.	
24-8-09	07:08:48	M024	ON	R1_Leave_Home begin
.	.	.	.	
24-8-09	07:08:58	M024	OFF	R1_Leave_Home end
.	.	.	.	
24-8-09	07:10:43	M037	ON	R2_Personal_Hygiene end

Finally, the F1 measure is a combined metric which can be computed according to Formula 3(d).

$$
\begin{aligned}
Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} && (a)\\
Precision &= \frac{TP}{TP+FP} && (b)\\
Recall &= \frac{TP}{TP+FN} && (c)\\
F1 &= \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*TP}{2*TP+FP+FN} && (d)
\end{aligned}
\tag{3}
$$

Another useful metric for evaluating binary classifiers is AUC that measures the area under the Receiver Operating Characteristic (ROC) curve. AUC indicates how well the model can distinguish between classes, whereas ROC is a probability curve. The True Positive Rate (TPR) is displayed against the False Positive Rate (FPR) on the ROC curve, with TPR on the y-axis and FPR on the x-axis. Classifiers that give curves closer to the top-left corner indicate a better performance. The classifier becomes less accurate as the curve approaches the 45-degree diagonal. The AUC of an excellent model is near 1, indicating that it has a high level of separability.

Unlike accuracy, AUC is independent from the decision threshold. This feature makes it an excellent statistic for evaluating the model in an unbiased way.

A key deficiency of the aforementioned metrics is that they do not consider the model confidence in predicting classes since these metrics do not reflect if a model predicts a true class (*TP* or *TN*) with a high probability or a marginal probability. This is why the model uses Cross Entropy or Log Loss for training (See Formula 4, where p is the prediction probability and y is the class label, 0 or 1). Predictions that are closer to the class label receive a lower Cross Entropy loss while the accuracy is a binary true/false for a specific sample.

$$CrossEntropy = -(y\log(p) + (1-y)\log(1-p)) \tag{4}$$

C. Results:

In this section, we present the results of running the BERT model for detecting behavior changes of two residents. First, we train the predictor model with 90 percent of ADL sequences of resident R1. We then evaluate the model using the remaining 10 percent of the unseen ADL sequences of resident R1 (Experiment 1). We also repeat this experiment for resident R2 (Experiment 2).

TABLE II
RUN PARAMETERS

Parameter	Value
Training Epochs	10
Loss function	Cross Entropy
Learning Rate	2e-5
Warm up Proportion	0.1
Drop-out rate	0.1
Max Sequence Length	128

TABLE III
COMPARISON OF EXPERIMENTS.

	Evaluation Metrics				
Experiments	**Accuracy**	**Precision**	**Recall**	**F1**	**AUC**
E1 : Classifier for R1 ADLs trained on R1 ADLs	0.87	0.89	0.84	0.86	0.88
E2 : Classifier for R2 ADLs trained on R2 ADLs	0.82	0.88	0.75	0.81	0.83
E3 : Classifier for R2 ADLs trained on R1 ADLs	0.81	0.64	0.90	0.75	0.79
E4 : Classifier for R2 ADLs trained on R1 ADLs and fine-tuned on R2 ADLs	0.84	0.69	0.90	0.78	0.82

In two separate experiments, we tested the model for predicting behavior abnormalities of resident R2 without fine-tuning (Experiment 3) and with fine-tuning (Experiment 4) on resident R2 data. In experiment 3, we used a pre-trained model which was trained on ADL sequences of resident R1 to predict abnormalities in resident R2. In experiment 4, we fine-tuned a pre-trained model using ADL sequences of resident R2, where the pre-trained model is trained on ADL sequences of resident R1. The intuition is that while different residents have their unique routines of life, there are commonalities that can be transferred from one model to the other (the Transfer Learning feature in Transformers). We expect the former experiment to show less accurate predictions. The reason is that different residents are supposed to have their unique routine of life which makes it unlikely to precisely predict their behavior change without fine-tuning the model on their specific data.

We also determined the number of training epochs by monitoring training loss and evaluation loss in order to avoid under- or over-fitting. The goal was to train an accurate model with training data (low training loss) that also shows promising performance on the evaluation data (low evaluation loss). The parameters we set for running the BERT-based classifier are listed in Table II.

Table III shows the results of the four experiments for 10 training epochs. The high values of accuracy, AUC, precision, and recall illustrate the capability of the BERT-based classifier model in predicting behavior abnormalities for both residents (E1 and E2). We acknowledge that training the model from scratch for each new resident is inefficient and possibly impossible. Therefore, experiments E3 and E4 were created to investigate the transfer learning characteristic of Transformers in this specific problem. In experiment E3, we assume that we do not have access to resident R2's ADL data. As a result, we train the model using data from resident R1 and test it using data from resident R2. The model predicts abnormalities well (high accuracy and recall), but it has a significant False Positive rate (noticeably poorer precision than E2), which means the classifier incorrectly labels some normal patterns as abnormal. In experiment E4, we use resident R2's ADL data to fine-tune the trained model from experiment E3. The results show a slight increase in all metrics, which we interpret as the capability of the model to transfer learned knowledge from one resident's behavior to predicting the behavior anomalies of others.

Some sample outputs from the classifier are shown in Table IV. The first three samples are correctly predicted as abnormal with relatively high probabilities. The reason is that they have clues of abnormal behavior such as long personal hygiene at night or leaving home without returning, which are not usual behavior of the resident. The next two samples (4 and 5) are also correctly predicted as normal with high probabilities. The last sequence is not classified correctly.

Our findings suggest that the BERT-based classifier is capable of detecting behavior abnormalities in ADL sequences. Transfer learning has also proven to be useful in fine-tuning a pre-trained model for a new resident. These results acknowledge the applicability of Transformer models to the behavior change detection problem through analyzing the ADL sequences. It is also a significant finding that transfer learning feature of Transformers is effective in training the models for new residents without requiring a huge amount of data collection and labeling for the new resident.

VI. DISCUSSION

As pre-trained BERT models are trained on general-domain texts, mainly from the web, some researchers [33] recommend training from scratch for the domain-specific data with training tasks, i.e., Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). Still, fine-tuning is considered as an important feature of Transformers that can be conducted for the in-hand task. We suggest comparing two different training approaches for future studies: i) training the model from scratch with domain-specific data (through MLM and NSP tasks) and fine tuning it for the specific task in-hand; and ii) using a pre-trained BERT model, which is trained on general-domain text, and then fine-tuning the model with domain-specific data. The reason for suggesting these two approaches is to determine which approach gives superior results for this domain-specific data. Although, the proposed representation of ADL sequences shares the same words/tokens with natural language, the logic is different. That is, the order of tokens in an ADL sequence does not follow the underlying logic of natural language. Also, we suggest testing different Transformer models and architectures, such as Roberta [37], GPT-3 [38], or BERT-Large, for building behavior change detection models.

TABLE IV
SAMPLES OF BERT-BASED ADL CLASSIFIER OUTPUT.

		True Label	Prediction		
ID	**Input ADL Sequence**		**Class "0" Probability**	**Class "1" Probability**	**Predicted Label**
(1)	PersonalHygieneLongNight WorkShortNight SleepShortNight PersonalHygieneShortMidNight WorkShortMidNight LeaveHomeShortMidNight	1	0.01	0.99	1
(2)	PersonalHygieneMediumMidNight WorkShortMidNight LeaveHomeShortMidNight	1	0.33	0.66	1
(3)	SleepShortNight PersonalHygieneMediumMorning BathingMediumMorning MealPreparationShortMorning LeaveHomeShortMorning	1	0.31	0.69	1
(4)	EnterHomeShortNight WorkShortNight PersonalHygieneShortNight SleepShortMidNight BedToiletTransitionShortMidNight PersonalHygieneShortMorning LeaveHomeShortMorning EnterHomeshortMorning	0	0.9989	0.0010	0
(5)	SleepShortNight WorkShortNight SleepShortNight BedToiletTransitionShortNight SleepShortNight WorkShortMorning SleepShortMorning WatchTVShortNight	0	0.81	0.19	0
(6)	WorkShortNight LeaveHomeShortNight EnterHomeShortMidNight SleepShortMidNight LeaveHomeShortMidNight	1	0.51	0.49	0

"0" is the Normal Class, and "1" is the Abnormal Class.

Gradual change in older adult daily behavior is common, especially in people with chronic disorders such as cognitive impairment [39]. As Transformers have the capability of learning long-term patterns in sequential data, gradual change in older adult daily behavior should also be explored in future research in order to determine their efficacy and generality in such environments. Moreover, in a lower-level data analysis, ADL impairment, which is found to be associated with chronic health problems in older adults [19], might be detected from sensor data by using Transformer models.

A major limitation of the supervised models is the scarcity of labeled data. By rearranging the ADLs and manipulating the duration of activities we were able to inject artificial irregularities into the original data in this study. To reproduce frequent behaviour aberrant patterns in future experiments, we recommend collecting labeled abnormal ADL sequences from residents with health issues.

VII. CONCLUSION

In this paper, we leveraged BERT Transformers in analyzing older adults' ADL sequences. Due to the fine-tuning capacity of Transformers, they are a good fit for supervised tasks when there is no access to a large, labeled training dataset. Their architecture also allows for Parallel Computation, which works hand in hand with the fine-tuning feature to make (near) real-time abnormality detection possible. The Positional Encoding feature in Transformers is a vital capacity leading to taking into account the context of a token when encoding. Finally, Transformers train in a Bi-Directional manner that allows for learning interrelationships of tokens in a sequence based on both pre- and post-tokens, resulting in a more accurate model.

This study contributes to the behavior abnormality detection literature by introducing a novel representation of daily behavior. This representation, which is borrowed from natural language, allows us to apply NLP sequence classifier models on ADL data in order to perform different supervised and unsupervised tasks including the behavior change detection task, which is presented in this paper. To the best of our knowledge, this is the first time that a BERT-based model is used for training a classifier that classifies daily ADL sequences into normal and abnormal classes. Also, transfer learning proved to be helpful when it comes to fine-tuning a pre-trained model for a new resident.

Behavior change detection in older adults, especially for those with chronic diseases, can help to improve their quality of life through providing on-time care and required precautions. It also can decrease the burden on caregivers by mitigating the need for continuous home care.

REFERENCES

[1] Mohammed Gh Al Zamil, Majdi Rawashdeh, Samer Samarah, M Shamim Hossain, Awny Alnusair, and Sk Md Mizanur Rahman. An annotation technique for in-home smart monitoring environments. *IEEE Access*, 6:1471–1479, 2017.

[2] Dorothy N Monekosso and Paolo Remagnino. Behavior analysis for assisted living. *IEEE Transactions on Automation science and Engineering*, 7(4):879–886, 2010.

[3] Samira Farivar, Mohamed Abouzahra, and Maryam Ghasemaghaei. Wearable device adoption among older adults: A mixed-methods study. *International Journal of Information Management*, 55:102209, 2020.

[4] Hemant Ghayvat, Muhammad Awais, Sharnil Pandya, Hao Ren, Saeed Akbarzadeh, Subhas Chandra Mukhopadhyay, Chen Chen, Prosanta Gope, Arpita Chouhan, and Wei Chen. Smart aging system: uncovering the hidden wellness parameter for well-being monitoring and anomaly detection. *Sensors*, 19(4):766, 2019.

[5] Paula Lago, Claudia Roncancio, and Claudia Jiménez-Guarín. Learning and managing context enriched behavior patterns in smart homes. *Future Generation Computer Systems*, 91:191–205, 2019.

[6] M Moallem, H Hassanpour, and AA Pouyan. Anomaly detection in smart homes using deep learning. *Iranian (Iranica) Journal of Energy & Environment*, 10(2):126–135, 2019.

[7] Marco Manca, Parvaneh Parvin, Fabio Paternò, and Carmen Santoro. Detecting anomalous elderly behaviour in ambient assisted living. In *Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, pages 63–68, 2017.

[8] Labiba Gillani Fahad and Syed Fahad Tahir. Activity recognition and anomaly detection in smart homes. *Neurocomputing*, 423:362–372, 2021.

[9] Caroline Bell, Cara Fausset, Sarah Farmer, Julie Nguyen, Linda Harley, and W Bradley Fain. Examining social media use among older adults. In *Proceedings of the 24th ACM conference on hypertext and social media*, pages 158–163, 2013.

[10] Diane J Cook. Learning setting-generalized activity models for smart spaces. *IEEE intelligent systems*, 2010(99):1, 2010.

[11] Marilyn J Rantz, Greg Alexander, Colleen Galambos, Amy Vogelsmeier, Lori Popejoy, Marcia Flesner, Annette Lueckenotte, Charles Crecelius, Mary Zwygart-Stauffacher, and Richelle J Koopman. Initiative to test a multidisciplinary model with advanced practice nurses to reduce avoidable hospitalizations among nursing facility residents. *Journal of Nursing Care Quality*, 29(1):1–8, 2014.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[13] Sergio Ryan Wibisono and Achmad Imam Kistijantoro. Log anomaly detection using adaptive universal transformer. In *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–6. IEEE, 2019.

[14] Haixuan Guo, Shuhan Yuan, and Xintao Wu. Logbert: Log anomaly detection via bert. *arXiv preprint arXiv:2103.04475*, 2021.

[15] Sasho Nedelkoski, Jasmin Bogatinovski, Alexander Acker, Jorge Cardoso, and Odej Kao. Self-attentive classification-based anomaly detection in unstructured logs. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1196–1201. IEEE, 2020.

[16] Sihao Xu, Wei Zhang, and Fan Zhang. Multi-granular bert: An interpretable model applicable to internet-of-thing devices. In *2020 IEEE International Conference on Energy Internet (ICEI)*, pages 134–139. IEEE, 2020.

[17] Damla Arifoglu and Abdelhamid Bouchachia. Abnormal behaviour detection for dementia sufferers via transfer learning and recursive auto-encoders. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 529–534. IEEE, 2019.

[18] Ahmad Lotfi, Caroline Langensiepen, Sawsan M Mahmoud, and Mohammad Javad Akhlaghinia. Smart homes for the elderly dementia sufferers: identification and prediction of abnormal behaviour. *Journal of ambient intelligence and humanized computing*, 3(3):205–218, 2012.

[19] Daniele Riboni, Claudio Bettini, Gabriele Civitarese, Zaffar Haider Janjua, and Rim Helaoui. Fine-grained recognition of abnormal behaviors for early detection of mild cognitive impairment. In *2015 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 149–154. IEEE, 2015.

[20] Nagender Kumar Suryadevara, Subhas C Mukhopadhyay, Ruili Wang, and RK Rayudu. Forecasting the behavior of an elderly using wireless sensors data in a smart home. *Engineering Applications of Artificial Intelligence*, 26(10):2641–2652, 2013.

[21] Hapugahage Thilak Chaminda, Vitaly Klyuev, and Keitaro Naruse. A smart reminder system for complex human activities. In *2012 14th international conference on advanced communication technology (ICACT)*, pages 235–240. IEEE, 2012.

[22] Yongkoo Han, Manhyung Han, Sungyoung Lee, AM Sarkar, and Young-Koo Lee. A framework for supervising lifestyle diseases using long-term activity monitoring. *Sensors*, 12(5):5363–5379, 2012.

[23] Yan Zhao, Baoqiang Ma, Pengbo Jiang, Debin Zeng, Xuetong Wang, and Shuyu Li. Prediction of alzheimer's disease progression with multi-information generative adversarial network. *IEEE Journal of Biomedical and Health Informatics*, 25(3):711–719, 2020.

[24] Anastasios Karakostas, Alexia Briassouli, Konstantinos Avgerinakis, Ioannis Kompatsiaris, and Magda Tsolaki. The dem@ care experiments and datasets: a technical report. *arXiv preprint arXiv:1701.01142*, 2016.

[25] Kundan Krishna, Deepali Jain, Sanket V Mehta, and Sunav Choudhary. An lstm based system for prediction of human activities with durations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–31, 2018.

[26] Bryan David Minor, Janardhan Rao Doppa, and Diane J Cook. Learning activity predictors from sensor data: Algorithms, evaluation, and applications. *IEEE transactions on knowledge and data engineering*, 29(12):2744–2757, 2017.

[27] Walaa N Ismail, Mohammad Mehedi Hassan, and Hessah A Alsalamah. Context-enriched regular human behavioral pattern detection from body sensors data. *IEEE Access*, 7:33834–33850, 2019.

[28] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[30] Marco Polignano, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro, and Valerio Basile. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR, 2019.

[31] Zhengjie Gao, Ao Feng, Xinyu Song, and Xi Wu. Target-dependent sentiment classification with bert. *IEEE Access*, 7:154290–154299, 2019.

[32] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*, 2019.

[33] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[34] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[35] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[36] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[37] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[38] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[39] Margaret G Stineman, Dawei Xie, Qiang Pan, Jibby E Kurichi, Debra Saliba, and Joel Streim. Activity of daily living staging, chronic health conditions, and perceived lack of home accessibility features for elderly people living in the community. *Journal of the American Geriatrics Society*, 59(3):454–462, 2011.