# Behavioral Data Categorization for Transformers-based Models in Digital Health

De Albuquerque Siebra, Clauirton; Almeida Matias, Igor Alexandre; Wac, Katarzyna

# Behavioral Data Categorization for Transformers-based Models in Digital Health

Clauirton Siebra
Quality of Life Technologies Lab
University of Geneva
Geneva, Switzerland
clauirton.dealbuquerque@unige.ch

Igor Matias
Quality of Life Technologies Lab
University of Geneva
Geneva, Switzerland
igor.matias@unige.ch

Katarzyna Wac
Quality of Life Technologies Lab
University of Geneva
Geneva, Switzerland
katarzyna.wac@unige.ch

2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI) | 978-1-6654-8791-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/BHI56158.2022.9926938

***Abstract*— Transformers are recent deep learning (DL) models used to capture the dependence between parts of sequential data. While their potential was already demonstrated in the natural language processing (NLP) domain, emerging research shows transformers can also be an adequate modeling approach to relate longitudinal multi-featured continuous behavioral data to future health outcomes. As transformers-based predictions are based on a domain lexicon, the use of categories, commonly used in specialized areas to cluster values, is the likely way to compose lexica. However, the number of categories may influence the transformer prediction accuracy, mainly when the categorization process creates imbalanced datasets, or the search space is very restricted to generate optimal feasible solutions. This paper analyzes the relationship between models' accuracy and the sparsity of behavioral data categories that compose the lexicon. This analysis relies on a case example that uses mQoL-Transformer to model the influence of physical activity behavior on sleep health. Results show that the number of categories shall be treated as a further transformer's hyperparameter, which can balance the literature-based categorization and optimization aspects. Thus, DL processes could also obtain similar accuracies compared to traditional approaches, such as long short-term memory, when used to process short behavioral data sequences.**

***Keywords—deep learning, transformers, human behavior, recommendations, behavior informatics, digital biomarkers***

## I. INTRODUCTION

Human behavior is a longitudinal multi-featured measure that directly impacts the health of individuals [1]. Due to this longitudinal aspect, computational systems in health care cannot ignore the sequence of users' behaviors when the aim is to predict or anticipate potential health issues resulting from such behaviors. In other words, these computational systems require models that capture the dependencies between temporal parts of sequential data. Deep learning transformer architectures [2] are currently the cutting-edge modeling approach for capturing the long-term dependencies among longitudinal inputs. Several studies [3,4] show their advantages (e.g., avoiding the vanishing gradient problem and support for parallel training) on previous models, such as recurrent neural networks and their variants. However, these studies focus on natural language processing (NLP) problems, such as language translation and next sentence prediction. Unlike previous studies, this paper evaluates our transformer-based architecture using longitudinal multi-featured behavioral data. The intuition of this architecture comes from the Behavior Sequence Transformer (BST) [5], which is a transformer-based architecture employed for recommendations in the market domain (e.g., for buying of clothes, electronic devices). BST considers individuals' shopping history and profiles to recommend their next purchase. Similarly, this architecture may be adapted to consider the behavioral routines (e.g., physical activities) and recommend modifications to these routines to improve a pre-defined health outcome in the short or long term (e.g., sleep duration).

As transformers-based approaches come from the NLP area, their learning process relies on lexica or vocabularies of terms that define domains. For example, Med-BERT [6] and BEHRT [7] are two transformer-based approaches in the health domain that use the International Classification of Diseases (ICD) codes as a lexicon. A lexicon is also adequate for the longitudinal health analysis of data obtained using questionnaires, such as in the English Longitudinal Study of Ageing (ELSA) [8], since they are mostly based on categories. Differently, many digital health applications passively collect numeric, continuous data corresponding to human physiology, like heart rate or daily behaviors like physical activity or sleep. To leverage the power of the transformers, it is required to understand the accuracy-complexity and other potential tradeoffs related to the choice of the categorizations of these features.

This paper contributes with a discussion on how the number of categories used during the training process impacts the model accuracy, indicating that such a number could be later treated as a learning hyperparameter. Therefore, the remainder of this paper is organized as follows: Section II summarizes the proposed architecture and its main components. Section III presents the case example, its dataset, the data augmentation strategy, the evaluation, and results. Section IV concludes this study with the main remarks and research directions.

## II. MQOL-TRANSFOERMER ARCHITECTURE

### A. Conceptualization: Behavior Sequence Transformer

The training stage of our learning approach uses the longitudinal behavioral data of several individuals to generate a BST model. The inputs are sequences of data assessed from *n* individuals $i_x$ (Fig. 1). These sequences have size *t* (e.g., *t* =7 days) and they are composed of tuples. Each tuple has, as first element, a multi-featured assessment $a_i$ composed of several features $[a_i^1, a_i^2, a_i^3, ..]$ (e.g., number of daily steps, average heart rate). The second element $\beta_i$ is a physiological/physical (e.g., obesity), psychological (e.g., stress), or social (e.g., level of interaction via social media) feature that presents some

This project (Onto-mQoL) has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grand agreement H2020-MSCA-IF-2020-101024693, AAL-2019-6-120-CP (Guardian), Swiss NSF project (157003), swissuniversities P-13 AGE-INT, and UCPH Data+ AI@CARE.

correlation with $a_i$. The prediction variable is the last element β (or $\beta_t$) of the sequence and it represents the target behavioral variable to be potentially changed. The prediction of $\beta_t$ considers the assessment $a_t$, and all the sequence ($a_1$, $\beta_1$) to ($a_{t-1}$, $\beta_{t-1}$).

After the training stage, the BST model works like an evaluator of simulated behaviors. Therefore, it receives as input the sequence ($a_1$, $\beta_1$) to ($a_{t-1}$, $\beta_{t-1}$) of an individual $i_x$, together with a simulated behavior $S_t$, to return a resultant prediction $\beta_t^P$. This process can be conducted with different behaviors $S_t$. Thus, someone can generate different instances of $S_t$ and verify which one presents the best effect to improve $\beta_t$. As the assessments are multi-featured and all these features also change and unfold over time, we needed to adapt the BST architecture, as detailed in the next section.

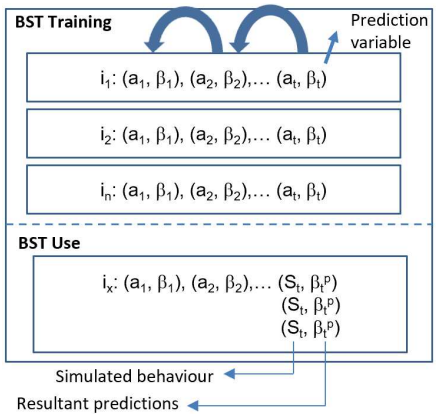


Fig. 1. BST application schema for behavioural analysis and prediction

## B. Specification

Four modules compose our architecture (Fig. 2), which are the set of static embedding layers (SEL), the set of behavior embedding layers (BEL) that also contains the positional embedding layer (PEL), the transformer component, and the set of fully connected layers.

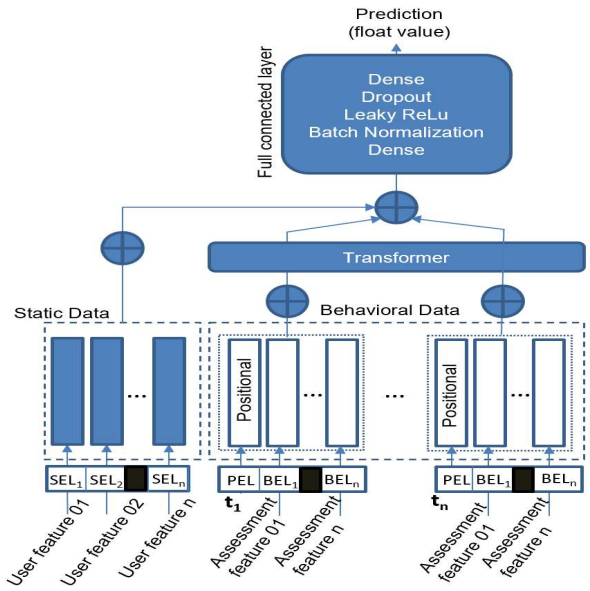


Fig. 2. mQoL-transformer: architecture proposed (adapted from [5])

The static embedding layers (SEL) module contains an embedding layer for each categorical feature that does not (or rarely) evolve over time (e.g., birth date). Behavioral longitudinal data ($a_{i..t}$) are the part of the individual's state assessment that changes over time. Heart rate and level of physical activities are some examples. Another component of this module is the positional embedding layer (PEL), which adds a positional vector to each set of BELs assessed at the same time. This step is essential since transformers process all the inputs in parallel, differently from RNN or Long short-term memory (LSTM) approaches, where inputs are fed in sequence. The transformer layer mainly relies on the multihead attention component, which enables attending to specific positions of the input sequence that are in fact important to compute a representation of the input sequence. Details about this layer are given in [5]. The final module contains a set of fully connected layers to further learn the interactions among dense features. A decimal value is generated as the resultant prediction.

# III. Case Example

## A. Behavioral Dataset and Prediction Scenario

This case example relies on data from 30 healthy adult volunteers that used Withings smartwatches for up to nine months. The following data were used along our experiments: Daily number of steps (integer); Distance travelled per day (meters); Time spent in soft physical activities - SPA (secs); Time spent in moderate physical activities - MPA (secs); Time spent in intense physical activities - IPA (secs); Daily heart rate (mean of beats per minute); and Sleep duration of a night (secs). We organized these daily samples in frames of seven days. Thus, the idea is to use the data of seven days to predict the sleep duration of the last night of this sequence (seventh day).

## B. Data Augmentation

Our experiment only employed daily samples that had complete values. This means they do not present missing data. Thus, 612 days were removed from the original set of 7786 days. These numbers give a missing data rate of 7.86% (the data were missing at random). We used the sliding window strategy (window size of seven days and timestep of one day) to augment the number of samples, as exemplified in the following schema (1). In this schema, the left-hand side of the arrow represents the input data, while the right-hand side represents the target data.

$$\begin{aligned} &x_1, x_2, x_3, x_4, x_5, x_6, x_7, \rightarrow sleep\ duration(x_7) \\ &x_2, x_3, x_4, x_5, x_6, x_7, x_8, \rightarrow sleep\ duration(x_8) \\ &x_3, x_4, x_5, x_6, x_7, x_8, x_9 \rightarrow sleep\ duration(x_9) \\ &\ldots \\ &x_{n-6}, x_{n-5}, x_{n-4}, x_{n-3}, x_{n-2}, x_{n-1}, x_n, \rightarrow sleep\ duration(x_n) \end{aligned} \quad (1)$$

After applying this strategy for each user, we obtained a total of 4019 longitudinal structures of seven day/samples each (mean of 134 structures per a study participant and $\sigma = 116.9$).

## C. Evaluation Process

We used the Long Short-Term Memory (LSTM) neural network to create a prediction baseline. Transformer-based approaches do not suffer from the vanishing gradient problem (i.e., long-term dependencies) when they process long sequences. This feature is one of their main advantages. In other words, transformers can remember old connections, which is

critical for digital health's longitudinal data that intend to analyze years of data. However, our experiments use short sequences (seven days), and consequently, their resultant accuracies tend to only be similar to well-known approaches such as LSTM.

The next experiments use our BST-based approach (Section 2) to generate models using different lexicon sizes for the six input features. The first experiment uses different number of categories for each of the six features that compose the assessment. The categorization was based on discussions of the literature. For example, *Inactive*, *Average*, *Active*, and *Very Active* are categories used to cluster step counts [9]. Therefore, the six features have the following number of categories: steps (4), distance (5), SPA (3), MPA (3), IPA (3), and HR (4). The remainder experiments use the population-based percentile to define the categorization limits of the features. For example, when eight categories are used (Table I), the limits are defined by the 12.5$^{th}$, 25$^{th}$, 37.5$^{th}$, 50$^{th}$, 62.5$^{th}$, 75$^{th}$, 87.5$^{th}$, and 100$^{th}$ percentiles. Table I shows the respective values for such percentiles considering each of the six features. The same idea is used to other numbers of categories.

TABLE I. LIMITS FOR CATEGORIZATION OF INTPUT FEATURES USING EIGHT CATEGORIES (CAT)

| Cat | Steps (p/ day) | Distance (m) | SPA (sec) | MPA (sec) | IPA (sec)[a] | Mean HR (p/ min) |
|---|---|---|---|---|---|---|
| C1 | 2832 | 164 | 6600 | 60 | 0 | 68 |
| C2 | 4163 | 3237 | 8401 | 420 | 600 | 71 |
| C3 | 5279 | 4189 | 9781 | 1260 | 2520 | 73 |
| C4 | 6313 | 5089 | 11280 | 1980 | 72802 | 77 |
| C5 | 7519 | 6082 | 12781 | 2580 | - | 80 |
| C6 | 9145 | 7527 | 15000 | 3480 | - | 85 |
| C7 | 12013 | 10089 | 18300 | 4979 | - | 90 |
| C8 | 31590 | 26322 | 58740 | 20760 | - | 122 |

[a] About 62% of users did not conduct intense physical activities (C1). Thus, this feature is imbalanced in terms of categories (C1 has 62% of samples) for this and other scenarios.

The categorization of sleep duration (prediction variable) defines four numeric categories and their intervals in minutes: 1 (0 to 395), 2 (396 to 446), 3(447 to 499), and (500 to 902). These categories are maintained along all experiments to facilitate the comparison regarding the prediction error of experiments that use different numbers of input categories. Moreover, as specified in our architecture (Fig. 2), the output of the last layer is a decimal number. Thus, the error is calculated by the difference between this output and the numeric values that represent the real sleep duration category. This strategy supports a more granular analysis of the results. After the percentiles-based categorization, we randomly split the data into training and validation sets using a rate of 70/30%. The training process was performed using a batch size of eight during 50 epochs or until the learning saturates. The model compiler was configured using the Adagrad optimizer with a learning rate of 0.01%, mean squared error as loss function, and the root mean squared error (RMSE) as validation metric. We captured the loss curves for training and validation to observe the evolution of the accuracy (average and standard deviation) and possible overfitting.

### *D. Results*

The LSTM approach returned an average RMSE value of 0.3528 with a standard deviation of 0.0075. We used these values as the baseline, i.e., the best possible result for prediction.

The next graph (Fig. 3) shows the results for the experiment using literature-based categories. The validation average RMSE of this experiment is 2.09 times higher than the baseline. Moreover, the loss curves present a smooth decreasing behavior, which means the learning process has a small gain over the epochs. The columns chart (Fig. 3) also shows the imbalanced levels of categories (columns) for each feature. Differently, the distribution of the other experiments (e.g., Figs. 4 and 5) are balanced and each of their categories has around the number of samples divided by the number of categories.

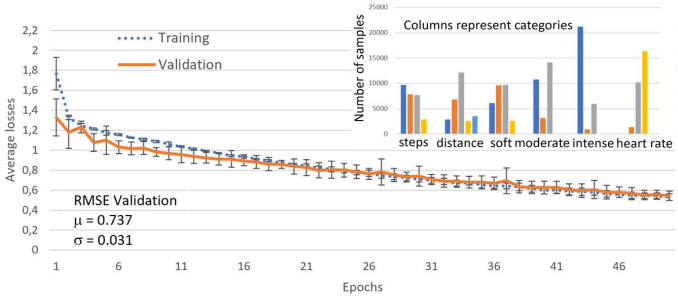


Fig. 3. Training and validation loss curves (literature-based categories)

The 8-categories experiment (Fig. 4) shows that the increase of categories and data balance, provided by the percentile-based distribution, improved the RMSE to 0.43, which is 1.21 higher than the baseline. Moreover, the learning curve is sharper than in the previous scenario and saturates about the epoch 31.

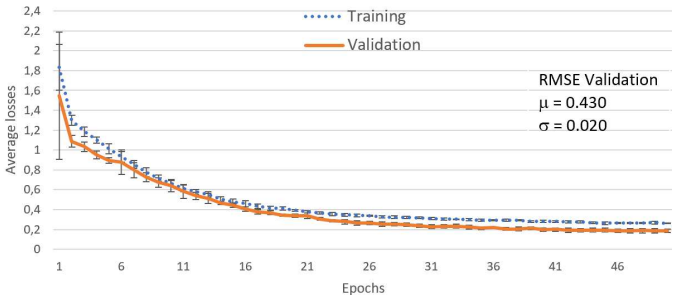


Fig. 4. Training and validation loss curves (eight categories)

The 16-categories experiment (Fig. 5) shows that the RMSE value did not significantly improve, even using the double amounts of categories for each feature. This RMSE value is 0.436 or 1.24 higher than the baseline. This value is almost the same as the previous experiment. However, the behavior of the learning curve, which is sharper than the 8-categories experiment, presents a saturation point about the epoch 23.

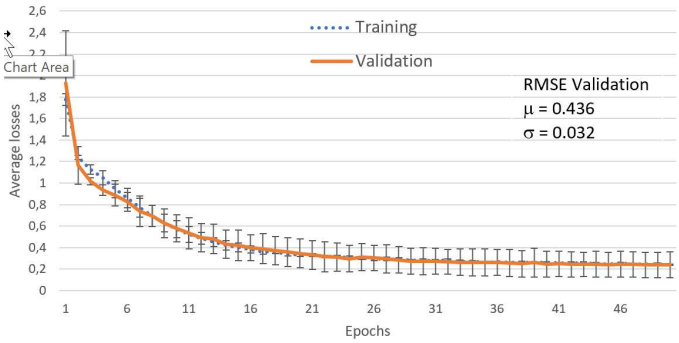


Fig. 5. Training and validation loss curves (sixteen categories)

Table II summarizes these and other experiments using their RMSE, standard deviation and learning saturation epoch values. These experiments are not detailed due to space restrictions. According to this table, the use of ten categories seems to be the optimal value for this problem.

TABLE II. SUMMARY OF EXPERIMENTS

| # cat. | Average RMSE | Std. Deviation | Saturation point (epoch) |
|---|---|---|---|
| Mixed | 0,736603582 | 0,030867636 | 49 |
| 4 | 0,619142313 | 0,027700254 | 46 |
| 6 | 0,426324946 | 0,022646566 | 34 |
| 8 | 0,429564250 | 0,019817069 | 31 |
| 10 | 0,404202831 | 0,033191876 | 30 |
| 12 | 0,416290563 | 0,019160941 | 22 |
| 16 | 0,435720688 | 0,032468449 | 23 |
| 24 | 0,406366000 | 0,063400000 | 21 |

### *E. Discussion*

Categorization is useful in this type of situation because it allows the organization of things, objects, and ideas, simplifying the understanding of the world. Thus, it assists processes such as prediction, inference, decision-making and all kinds of environmental interactions. Our experiments used the percentile strategies to create categories since it ensures that these categories have approximately the same number of samples. Thus, we avoid imbalanced data and, consequently, issues associated with bias in the results. However, real digital health applications tend to use feature-specific categories. For example, the heart rate feature can be categorized into five zones regarding the level of physical activities: recovery/easy, aerobic/base, tempo, lactate threshold, and anaerobic [10]. Thus, the definition of categories must reconcile the literature-specific categorization and strategies to optimize the prediction. For example, we can break or join categories to improve the search space. The data balance must also be considered in such actions.

According to the results, the number of categories affects the accuracy of the predictions. Thus, its definition could be similar to any other machine learning hyperparameter, such as learning rate, batch size, and the number of epochs. The challenge is to define some function that guides the definition of this number. For example, relating this function to the number of samples or their distribution (statistic side), allied to pre-defined recognized intervals (domain-dependent side), such as in [10]. The experiments also indicated that our approach only obtained similar prediction accuracies compared with the LSTM baseline.

As previously discussed, the power of transformers is demonstrated when they process long sequences. For example, ongoing research of our lab intends to use daily data of about 900 volunteers collected over three years to predict the psychological health of mobile users. In this case, we expect our approach to obtain better results due to the long sequences that it needs to analyze (3x365 = 1095 time points/person). Another strategy that may improve the results is to use different numbers of categories for each feature. Our experiments based on percentiles defined a unique number of categories, and this number was used to categorize the values of all the features. However, a feature such as *heart rate* may not require many categories such as the feature *number of steps*. The challenge is to find characteristics regarding the data (e.g., their distribution) that could support this definition. As the recommendation outcomes indicate categories that users should move to, the average value of these categories may be used as a more concrete target. For example, if a recommendation indicates C2 for steps (Table I), this value is around 3502 steps per day. This action is similar to Defuzzification, which transfers fuzzy inference results into a crisp output. Methods such as Center of Gravity (COG), Mean of Maximum (MOM), and Center Average could be used to find this value and give a concrete notion on recommendations. Another interesting question is about the generalization of this approach. For example, the experiments show that using ten categories seems to return the most accurate results. However, we defined these categories using data from a specific device (Withings). Other devices such as Samsung, Apple or Fitbit Smartwatches probably have different hardware and software to collect data and present deviations compared to Withings. Thus, assessments in such devices may fall into different categories. Future use of data from other devices aims to verify this aspect.

## IV. CONCLUSION

This paper emphasizes the importance of tuning the number of categories as an additional step to improve the accuracy of learning processes that use implementations of behavior sequence transformers. This step applies to the cases where the prediction/task at hand does not define the number of categories. One of the main limitations of this investigation is the size and lack of high dimensionality of the dataset used, which only contains data of 30 users for a few months. Even using a sliding window strategy to augment the data, the use of large datasets is important to confirm our findings. Future studies intend to conduct experiments using long sequences of passive data, which are currently being collected together with the questionnaire-based quality of life data.

## REFERENCES


[1] V. Manea, and K. Wac, "mQoL: mobile quality of life lab: from behavior change to QoL," In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, pp. 642–647, 2018.

[2] A. Vaswani, et al., "Attention is all you need. Advances in neural information processing systems," vol. 30, 2017.

[3] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, No. 01, pp. 6706-6713, 2019.

[4] S. Karita, et al., "A comparative study on transformer vs rnn in speech applications," In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 449-456, 2019.

[5] Q. Chen, H. Zhao, W. Li, P. Huang, and W. Ou, "Behavior sequence transformer for e-commerce recommendation in alibaba," In Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data, pp. 1–4, 2019.

[6] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," NPJ digital medicine, 4(1), 1-13., 2021.

[7] Y. Li, et al., "BeHRt: transformer for electronic Health Records," Sci. Rep. vol. 10, pp. 1–12, 2020.

[8] J. Banks, et al., "English Longitudinal Study of Ageing: Waves 0-9, 1998-2019 [data collection]," 36th Edition. UK Data Service, 2021.

[9] D. M. Bravata, et al., "Using pedometers to increase physical activity and improve health: a systematic review," Jama, 298(19), 2296-2304, 2007.

[10] M. L. Eckard, H. C. Kuwabara, and C. M. Van Camp, "Using heart rate as a physical activity metric," Journal of Applied Behavior Analysis, 52(3), 718-732, 2019.