Semantic Message Representations for Behavioral Prediction in a Diabetes Prevention System

by

Mohammad Hasanisaznaghi

A thesis accepted and approved in partial fulfillment of the
requirements for the degree of
Master of Science
in Computer Science

Dissertation Committee:
Prof. Thanh H. Nguyen, Chair
Prof. Jee Choi, Core Member

University of Oregon
Spring 2026

© 2026 Mohammad Hasanisaznaghi
All rights reserved.

THESIS ABSTRACT

Mohammad Hasanisaznaghi

Master of Science in Computer Science

Title: Semantic Message Representations for Behavioral Prediction in a Diabetes Prevention System

Mobile health programs increasingly use machine learning to predict how participants respond to behavioral interventions. However, how message content is represented limits what these models can learn. This thesis studies a diabetes prevention system that sends weekly health messages to participants in rural India and predicts their responses. The original system encodes each message as a binary identifier, treating every piece of content as an unrelated category. This work replaces those identifiers with semantic embeddings from two BERT-based models, general-domain BERT and Sentence-BERT, evaluated across recurrent, convolutional, and attention-based architectures. Prediction accuracy remains modest across all conditions, and richer models do not beat simpler ones, showing that representation quality matters more than model complexity at this data scale. Beyond accuracy, the embeddings recover category structure that binary encoding misses and reveal that a group of messages aligns more closely with a different behavioral category than the one they were originally assigned.

ACKNOWLEDGMENTS

I would like to sincerely thank my supportive advisor, Prof. Thanh H. Nguyen, for her patience, guidance, and support throughout this thesis. Her mentorship has helped me become a better researcher during my graduate studies, and this thesis would not have been possible without her.

I would also like to thank Dr. Sarah Kinsey, whose prior work on this project gave me a strong foundation to build on. I am also grateful to Prof. Jee Choi for serving on my committee.

Lastly, I want to thank my family and friends for their endless support and for believing in me.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES 7

LIST OF TABLES 8

I. INTRODUCTION 9

II. BACKGROUND 13

2.1. Healthcare Behavioral Prediction 13

2.2. Text Embeddings 14

2.3. Neural Architectures for Sequential Data 16

III. METHODOLOGY 19

3.1. Current System Overview 19

3.2. Statistical Analysis of System 23

3.2.1. Category Distribution Analysis 23

3.2.2. Weekly Engagement Patterns 25

3.2.3. Limitations of Binary Feature Encoding 26

3.3. Semantic Embedding Integration 29

3.3.1. Motivation 29

3.3.2. Pre-computing Embeddings 29

3.3.3. Integrating Embeddings into Feature Vectors 30

3.3.4. Experimental Setup 31

IV. RESULTS AND EVALUATION 33

4.1. Baseline Results 33

4.2. BERT Uncased Results 34
4.3. SBERT Results 35
4.4. Message Embedding Structure 36
4.5. Embedding Evaluation 38
4.6. Limitations and Future Work 40

V. CONCLUSION 41

REFERENCES CITED 42

LIST OF FIGURES

Figure Page

3.1. Personalized Message-based Intervention System Overview, adapted from [1] 19
3.2. Distribution of questions based on behavioral category 24
3.3. Response rates across behavioral categories 24
3.4. Distribution of response quality levels by category (higher level means more engagement) 25
3.5. Weekly response rates throughout the 24-week study time 26
3.6. Unstructured feature space of binary ID encodings 27
3.7. BasicLSTM hidden states during peak engagement (weeks 8 - 15) 28
3.8. BasicLSTM hidden states during the engagement decline phase (weeks 16 - 23) 28

4.1. Comparison of all 57 health messages under binary ID encoding (left), BERT Uncased embeddings (middle), and SBERT embeddings (right) 37
4.2. LOO-CV classification accuracy for binary ID encoding, BERT Uncased, and SBERT (left), and per behavioral category (right). 40

LIST OF TABLES

Table Page

4.1. Train and test accuracy under binary ID encoding. 33

4.2. Train and test accuracy under Bert Uncased. 34

4.3. Train and test accuracy under SBERT. 35

CHAPTER I

INTRODUCTION

Reaching vulnerable populations with effective diabetes prevention programs remains a challenge for public health systems worldwide. Rural populations often lack access to regular health education, face significant barriers to healthcare facilities, and receive little ongoing support for behavior change [1]. These difficulties have motivated researchers toward mobile-based health programs that reach people through platforms they already use in their daily lives.

The system at the center of this study delivers personalized diabetes prevention messages through WhatsApp to participants in rural areas. It sends weekly health messages, collects their responses about behavioral changes, and uses this feedback to select better messages for future weeks. The system was tested in a 24-week study with over 1,000 participants in rural India and showed improvements in participants' diabetes knowledge, exercise habits, and dietary choices [1].

A key limitation of the existing system is how it handles health education content, since each message and question is assigned a unique identifier using binary encoding and the model treats every piece of content as a completely separate category. This means a message like "Try walking 30 minutes daily" and one like "Consider taking stairs instead of elevators" receive numerically unrelated codes with no encoded connection between them, even though both target the same behavioral domain. Without any representation of semantic relationships between messages, the prediction model cannot generalize across similar content and misses patterns in how related messages influence participant behavior.

Bidirectional Encoder Representations from Transformers (BERT) encodes text by modeling contextual relationships between words, producing representations that reflect surrounding context rather than assigning each word a fixed meaning

[2]. These are rich token-level representations, but BERT was not designed to produce sentence embeddings where semantically similar content is reliably close together. Sentence-BERT (SBERT) addresses this directly by training on sentence pairs with an objective that pulls similar sentences closer in the embedding space [3]. Therefore, messages like "try walking 30 minutes daily" and "consider taking the stairs instead of the elevator" are placed near each other even without sharing any content words.

Research in healthcare natural language processing has shown that representing medical content through its meaning, instead of relying on arbitrary codes, generally leads to better prediction of health-related outcomes [4, 5]. By recognizing semantic similarity between messages, a model can generalize across related content instead of treating each item as entirely new. The diabetes prevention system's binary encoding works against this; each message and question is assigned a unique symbol with no information about its actual content. Replacing these identifiers with semantic text representations is a direct way to close this gap.

This work proposes two specific enhancements to address current limitations. First, it replaces the existing binary ID encoding of intervention messages and questions with semantic text representations by comparing two BERT-based approaches. BERT (uncased) serves as a general-domain baseline, learning contextual word meanings from large text corpora. Additionally, SBERT extends this approach by capturing sentence-level semantic similarity, placing messages with similar meanings closer together in the embedding space. These models were selected because the diabetes prevention messages are written in plain, everyday language which makes general-purpose architectures more appropriate than variants trained on specialized clinical or biomedical text. Secondly, this work explores multiple neural network architectures to analyze whether more expressive models can better utilize these se-

mantic representations compared to the existing baseline. Both enhancements focus on improving behavioral prediction accuracy while remaining fully compatible with the system's current message selection and delivery components.

This thesis implements four additional architectures to test whether richer models can make better use of semantic representations. The Bidirectional GRU (BiGRU) treats the participant's weekly history as a sequence and reads it in both directions, thus each prediction draws on the full context of past interactions rather than just what came immediately before [6, 7]. The Temporal Convolutional Network (TCN) takes a different approach, using stacked dilated convolutions to capture behavioral patterns at multiple time scales within the sequence [8]. The Attention-Weighted MLP applies a single attention layer over the input features to identify what is most relevant for a given prediction, while the Transformer Encoder applies full multi-head self-attention across the weekly sequence to determine which past time steps carry the most signal [9].

Model performance is evaluated across five random seeds and results are averaged to reduce the effect of any single data split. Performance is reported at epoch 50, the point at which test accuracy stabilizes before overfitting begins. Accuracy is measured separately across three behavioral categories (health knowledge, physical activity, and dietary behavior) and accounts for participants who did not respond to every weekly message.

Beyond this specific system, this work also provides practical observations from applying semantic embeddings to a behavioral health prediction task. The comparison between BERT (uncased) and SBERT helps illustrate when general-purpose versus sentence-optimized representations may be useful in practice, depending on the trade-off between simplicity and performance. The architectural experiments also help show how different model families behave in realistic digital health set-

tings, where training data is often limited and imbalanced.

CHAPTER II

BACKGROUND

2.1 Healthcare Behavioral Prediction

Predicting patient behavior has grown into one of the more important challenges in digital health, and early systems in this space focused mainly on clinical outcomes like disease progression and treatment response, with less attention to how patients actually engage with health recommendations over time. Traditional healthcare prediction approaches relied heavily on demographic and clinical data and treated patient behaviors as fixed traits rather than processes that shift over time. These early systems used classifiers such as logistic regression and decision trees that could identify risk factors but were not designed to capture how behaviors evolve across repeated interactions. Reviews of machine learning for health prediction have pointed to this same gap, noting that incorporating longitudinal data to capture temporal dynamics remains a central challenge for the field [10].

Supervised learning approaches brought meaningful progress by demonstrating strong performance in predicting medication adherence, identifying patients at risk of dropout, and modeling behavioral change patterns [11], and they drew on richer data sources including electronic health records, wearable devices, and patient surveys to capture patterns that earlier methods could not detect. However, most of these models still treat behavioral responses as independent events and miss the fact that a participant's decision to engage today directly influences their likelihood of engaging tomorrow, and as digital health programs generate more sequential data this limitation has become harder to ignore.

The current diabetes prevention system illustrates this challenge because it requires sustained changes across multiple areas, including diet, exercise, and health

monitoring, not just a response to a single acute event. Participants often show complex patterns where early enthusiasm decreases over time and individual circumstances shape very different behavioral paths, and this makes it difficult for models that treat each observation independently to capture what is actually happening.

2.2 Text Embeddings

Binary identification encoding is the most basic approach, where each category is assigned a unique numerical identifier and healthcare responses like "never," "sometimes," and "frequently" might be mapped to numbers 1, 2, and 3. It is computationally simple and easy to implement, however it treats all categories as equally different from one another and carries no information about semantic relationships or conceptual similarity between them.

Word2Vec and similar distributional models brought a meaningful step forward by showing that word relationships could be captured by learning vector representations from large text corpora [12], and these models placed semantically similar words closer together in the vector space so that "exercise" and "physical activity" would end up near each other even if they never appeared together in the same sentence. The limitation was that each word still received a single fixed representation regardless of context, therefore a word like "positive" would look the same whether a patient described a "positive attitude" toward treatment or tested "positive" for a condition, and this became a real problem in healthcare settings where the same word can mean very different things depending on the context.

BERT addressed this by introducing bidirectional context understanding through the transformer architecture [2], and unlike earlier approaches that read text from

left to right, it considers context from both directions at the same time and produces richer representations that reflect surrounding meaning. When processing a message like "Try to exercise more regularly," BERT draws on all surrounding words simultaneously and captures the meaning of each term in context, and this makes it more suitable for encoding health intervention content than simpler schemes that treat each word or identifier in isolation.

Medical domain adaptations of BERT have shown strong performance in healthcare applications compared to general-purpose models. ClinicalBERT, trained on clinical notes and medical literature, shows improved understanding of medical terminology and clinical context [5], and Med-BERT, trained on electronic health records, has also achieved strong improvements in disease prediction tasks [4]. These results indicate that domain-specific pretraining can help in clinical settings, though the benefit depends on how closely the target domain matches the pretraining data. Since the diabetes prevention messages in this study are written in plain conversational language, not clinical terminology, general-purpose models are a more appropriate starting point than specialized clinical variants.

SBERT extends the original BERT model by fine-tuning it specifically for sentence-level semantic similarity [3], and while BERT produces strong contextual representations for individual words and tokens, comparing two full sentences using BERT directly is computationally expensive and does not always reliably capture overall sentence meaning. SBERT addresses this by training on sentence pairs using a Siamese network structure, producing fixed-size sentence embeddings where semantically similar sentences are placed closer together in the vector space, and this makes it a more natural fit for tasks that involve comparing complete messages instead of individual words.

2.3 Neural Architectures for Sequential Data

Sequential data modeling requires specialized neural architectures that can capture temporal dependencies and relationships over time. Standard feedforward networks process each input independently, while behavioral data has an inherently sequential structure. A participant's response in a given week is influenced by their history of previous interactions, making it important for models to account for this context.

Recurrent Neural Networks introduced the foundational approach to sequential modeling by maintaining hidden states that carry information across time steps [13]. RNNs process sequences one element at a time, updating their internal memory based on current inputs and previous states. However, basic RNNs suffer from the vanishing gradient problem, where information from early time steps gets lost as sequences grow longer [14]. This limitation proves particularly problematic in healthcare applications where intervention effects often appear over extended periods.

Long Short-Term Memory (LSTM) networks solved the vanishing gradient problem through sophisticated gating mechanisms that control information flow [15]. LSTMs use three types of gates: forget gates that remove irrelevant information, input gates that decide what new information to store, and output gates that control what information gets passed forward. This design allows LSTMs to maintain relevant information across long sequences while discarding unnecessary details.

Gated Recurrent Units (GRU) emerged as a simplified alternative to LSTMs, combining multiple gates into fewer components while maintaining effective long-term modeling capabilities [7]. GRUs use fewer parameters than LSTMs and re-

quire less computational resources, making them attractive for healthcare applications with limited processing power. Comparative studies show that GRUs often perform similarly to LSTMs while training faster and using less memory [16].

Attention mechanisms introduced a major shift in sequential modeling by allowing networks to focus on the most relevant parts of input sequences, which reduces the need to compress all information into fixed-size representations [17]. Instead of relying solely on the final hidden state, attention mechanisms can examine all previous time steps and decide which ones are most important for current predictions. This approach proves particularly valuable in healthcare behavioral prediction where understanding the full context of patient interactions improves accuracy.

Transformer architectures have become a common choice for many sequential modeling tasks, showing that attention alone can outperform recurrent networks in a wide range of settings [9]. Unlike recurrent models, transformers process full sequences at once, which makes training faster and helps the model capture relationships across distant time points. Through self-attention, each element in the sequence can directly use information from the others, giving the model broader context when predicting participant responses over time.

Convolutional approaches offer an alternative way to model sequential data without relying on recurrent hidden states or attention over all time steps. Temporal Convolutional Networks apply dilated causal convolutions across the sequence, where dilation rates double with each layer, allowing the network to capture patterns at progressively longer time scales [8]. Residual connections between layers help preserve information throughout the depth of the network. Because the convolutions are applied in parallel across the sequence rather than step by step, TCNs can be more computationally efficient than recurrent models, and empirical evalua-

tions have shown them to be competitive with LSTMs and GRUs across a range of sequence modeling tasks.

These architectures each bring different inductive biases to sequential modeling. Recurrent models track dependencies step by step, convolutional models detect patterns across multiple time scales, and attention-based models weigh the relevance of each past time step directly. For a behavioral prediction task like the diabetes prevention system, where participants interact weekly over several months on a relatively small dataset, it is not obvious in advance which of these properties will be most useful. This motivates evaluating several architectural families alongside the semantic embedding changes to understand whether model choice meaningfully affects how well participant responses can be predicted.

CHAPTER III

METHODOLOGY

3.1 Current System Overview

Figure 3.1 provides an overview of the diabetes prevention system.Figure 3.1 provides an overview of the diabetes prevention system. Health workers conduct baseline questionnaires with participants, who then engage with the system weekly through WhatsApp. Participant responses are passed to a reinforcement learning agent that maintains a behavioral profile across five domains and selects the most suitable message and question from the Message Bank and Question Bank for the following week. This feedback loop allows the system to continuously adjust content delivery over the 24-week study period.

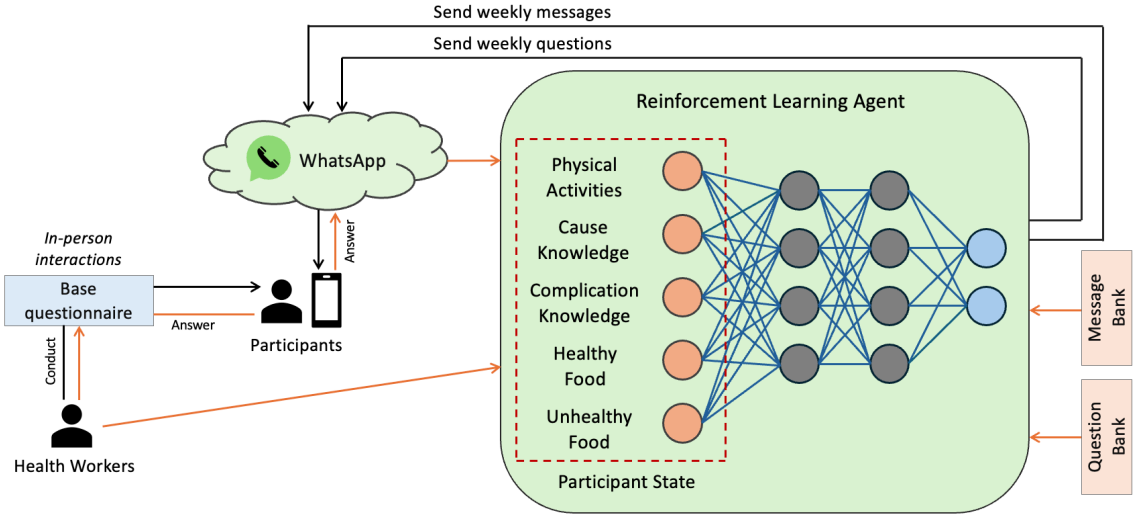


Figure 3.1. Personalized Message-based Intervention System Overview, adapted from [1]

Data Pipeline

The diabetes prevention system uses a four-step pipeline to convert participant interaction data into weekly behavioral response predictions, from data collection through to model training and output.

Step 1: Data collection and integration

Weekly response data spanning weeks 2–31 is loaded from per-week participant files, each containing response values, question identifiers, and message identifiers for that week, and combined with baseline questionnaire data. Participant filtering also happens at this step, keeping only authorized participants and ensuring data quality through validation checks.

Step 2: Feature engineering and encoding

Raw data is transformed into numerical representations suitable for machine learning, and the process handles participant information from baseline questionnaires, encodes previous response patterns using one-hot representations, and converts message and question identifiers into fixed-length binary feature vectors. The binary encoding of message and question identifiers represents the original system's approach, and it is replaced in this study with semantic text embeddings. This step also creates sliding windows of previous interactions that capture behavioral patterns over time and keep feature representation consistent across the three behavioral categories.

Step 3: Model training and testing

Models are trained and validated using random train/test splits across multiple independent seeds and participants are split into training and testing groups while keeping each participant's full sequence intact to preserve the temporal structure in the data. Performance is monitored across behavioral categories using category-

specific metrics throughout training.

Step 4: Prediction and evaluation

Lastly, the model outputs a probability distribution over three response levels for each participant and each week. Entries where no response was recorded are excluded, therefore evaluation reflects only weeks where the participant actively engaged. Performance is measured using overall accuracy and per-category accuracy separately for Consumption, Exercise, and Knowledge.

Data collection and processing flow

Data collection starts by loading participant interaction records from weekly response files, where each file covers one week of the study's 30 calendar weeks (weeks 2 - 31) and contains participant answers to behavioral questions. Initial enrollment timing is identified by detecting each participant's first response week, which accommodates the rolling enrollment design where participants entered the study at different times. Baseline questionnaire data is also loaded to provide participant context from pre-study assessments, and this covers health knowledge, dietary habits, exercise patterns, and demographics, which are transformed into numerical vectors that represent each participant's characteristics at the start of the study.

Data quality is maintained through multiple filtering steps that remove participants not listed in the authorized registry and exclude incomplete response records. For model training specifically, an additional response-rate filter retains only the most engaged participants, helping the model learn from more consistent behavioral patterns and reduce noise from sporadic responses.

Each participant's timeline is also normalized, so their first week becomes week 0 and their final week becomes week 23, creating a standardized 24-week period re-

gardless of when they entered the study, and this allows for meaningful time-based analysis across the full participant group while preserving individual behavioral trajectories.

Participant Demographics and Timeline

The behavioral data comes from a large cohort of participants enrolled in a rural diabetes prevention study conducted in India, and the focus is on individuals from underserved communities with limited access to traditional healthcare resources, which is the target population for this kind of mobile health intervention.

The study timeline spans 24 weeks of active participation per participant, with the overall study period extending across 30 calendar weeks (weeks 2 - 31) to accommodate rolling enrollment, where new participant groups joined at different start dates while each individual still went through the same 24-week program. Both calendar weeks and normalized participant weeks are tracked to support population-level and individual-level analysis.

During each week, participants receive health messages and respond to behavioral questions across three main areas: consumption habits, exercise behaviors, and diabetes knowledge. Participants typically get two questions per week, which adds up to 48 possible responses over the full study period, though not everyone answers consistently and some participants engage more than others. The full 24-week sequence is kept for each participant to preserve the time-based relationships that behavioral modeling depends on. Missing responses are handled by assigning -1 values, and these get converted to [-1,-1,-1] vectors in the one-hot representation rather than filling them in through imputation. A missing response is not just a gap in the data and it can reflect something meaningful about how engaged a participant is, so preserving it helps the model distinguish between different kinds of

missing data and use those patterns when making predictions.

3.2 Statistical Analysis of System

The analysis looks at participant behavior across three behavioral categories over the 24-week study period. The original study enrolled over 1,000 participants across multiple intervention groups, including both AI-driven and control conditions, and the focus here is on the 537 participants from the AI intervention group, who account for 25,776 total question instances (Figure 3.2) and an overall average engagement rate of 32.7%. The three categories used throughout, Consumption, Exercise, and Knowledge, came from the original study's five behavioral domains: healthy food consumption, unhealthy food and tobacco and alcohol consumption, fitness and physical activity, diabetes cause knowledge, and diabetes complication knowledge. This grouping was based on the finding that behavioral responses vary across these types and category-specific models tend to work better than a single unified approach [1]. A chi-square test on response-level distributions across the three categories ($\chi^2 = 14.028$, $p = 0.029$) confirms that participants respond differently depending on the type of behavioral question, and each category produces its own distinct engagement patterns.

3.2.1 Category Distribution Analysis

Knowledge questions make up the largest share with 14,433 instances (56%), followed by Consumption with 10,446 (40.5%) and Exercise with 897 (3.5%). This is mainly because the intervention was built around health knowledge and exercise-related questions were administered less frequently. Figure 3.3 illustrates response rates stayed fairly consistent across all three categories, with Consumption at 35%,

Exercise at 36.7%, and Knowledge at 33.8%. Figure 3.4 also shows the level of engagement through these three categories. Participants' engagement levels were largely consistent regardless of the type of question they were asked.

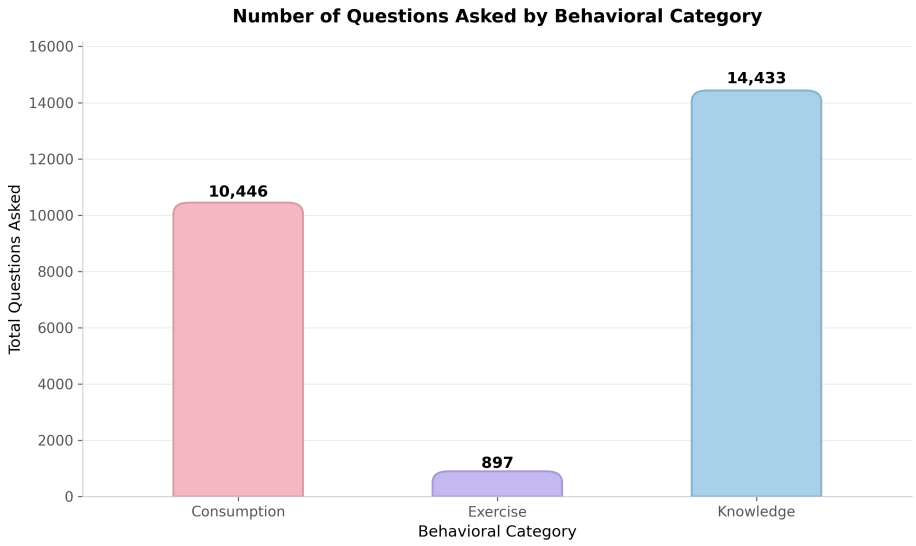


Figure 3.2. Distribution of questions based on behavioral category

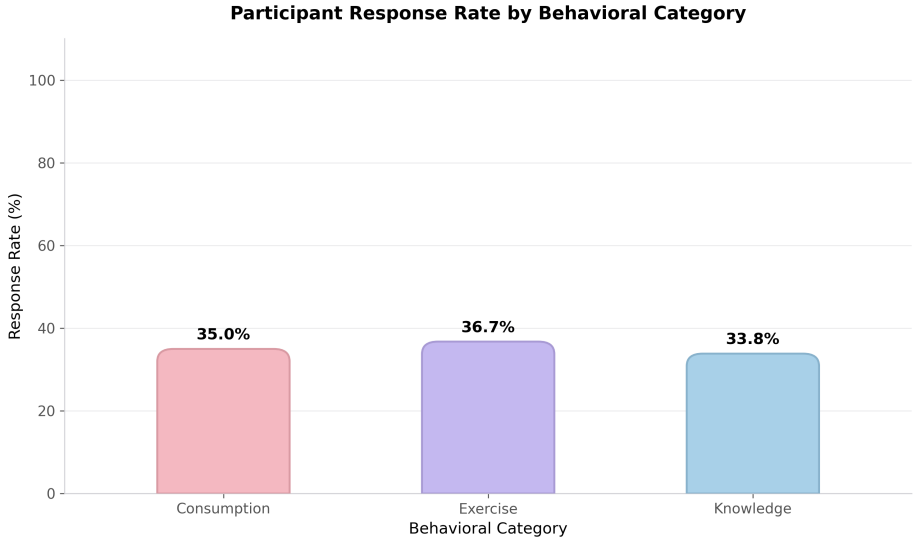


Figure 3.3. Response rates across behavioral categories

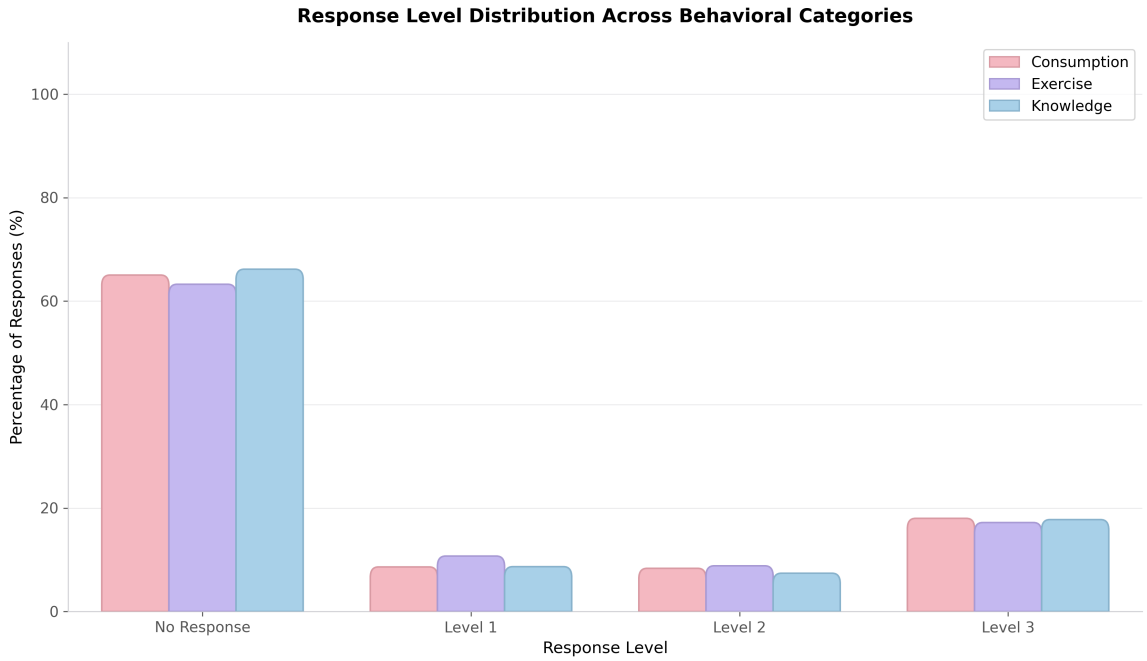


Figure 3.4. Distribution of response quality levels by category (higher level means more engagement)

3.2.2 Weekly Engagement Patterns

Participants' response behavior generally followed three phases across the 24-week period; an early adjustment period, a mid-study peak, and a gradual decline toward the end. Engagement was highest in the middle of the study (weeks 8 - 15), peaking at 48.1% in week 13 and staying mostly elevated compared to the rest of the study time. Response rates gradually declined and stabilized in the following weeks. Additionally, figure 3.5 shows the overall average response rate across all 24 weeks was 32.7%. It was an expected pattern since participants usually need some time to adjust before treating these systems as a regular engagement, and response rates tend to drop off naturally in the later weeks of a longer program because some people might feel overwhelmed or not seeing any benefit from it.

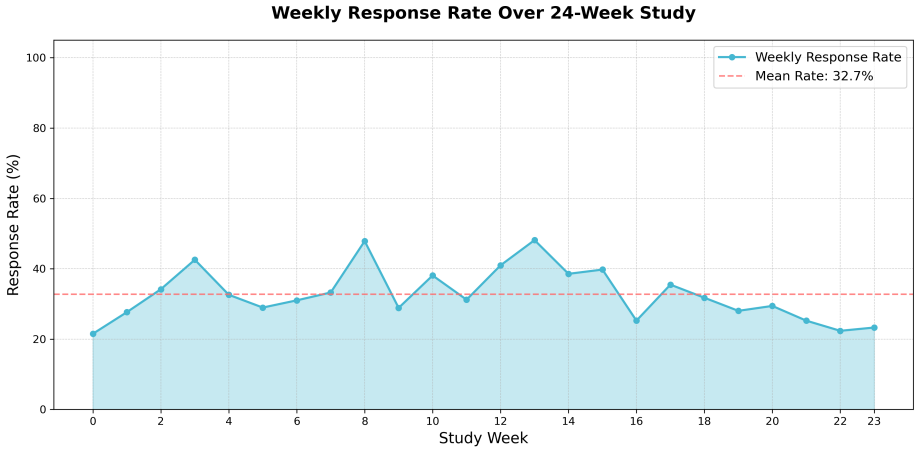


Figure 3.5. Weekly response rates throughout the 24-week study time

3.2.3 Limitations of Binary Feature Encoding

Figure 3.6 shows the raw feature space of the health messages by mapping their high-dimensional binary encoding into a two-dimensional representation. The three behavioral categories (Knowledge, Consumption, and Exercise) do not form distinct clusters and are instead mixed together in a single distribution. The dominant Knowledge category is scattered among points from the Consumption and Exercise categories, and there is no clear separation between them. Raw binary features alone are not enough to distinguish between different types of the messages, and a better feature representation is needed.

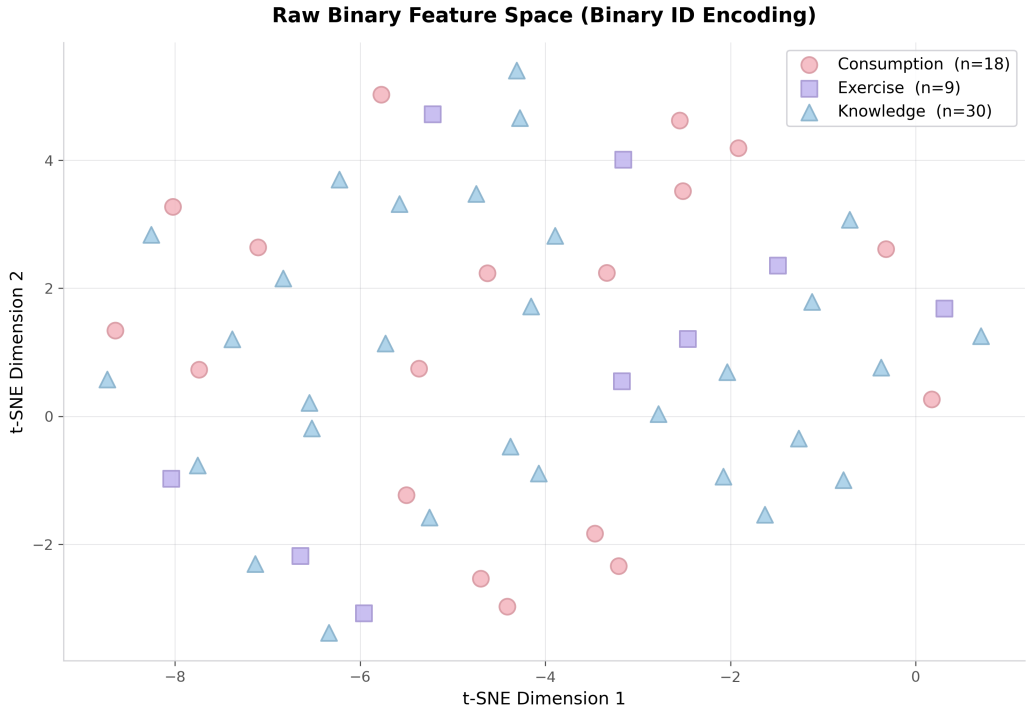


Figure 3.6. Unstructured feature space of binary ID encodings

To show the need for a better approach, a BasicLSTM model was trained on binary ID features to test whether it could learn to separate different response levels. Hidden states were extracted after training and projected into two dimensions to see whether the model had picked up any internal structure. As shown in Figure 3.7, when participant responses were most consistent, response levels show no clear clustering and remain fully mixed, and the model could not distinguish between response types even under these conditions.

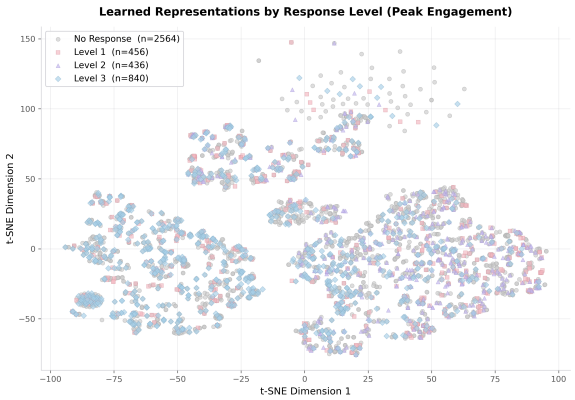


Figure 3.7. BasicLSTM hidden states during peak engagement (weeks 8 - 15)

Figure 3.8 repeats a similar analysis on the engagement decline phase (weeks 16 - 23), where data is more scattered and less consistent. The result is mostly the same, and the non-response points dominate this plot which reflects on the drop in participation engagements.

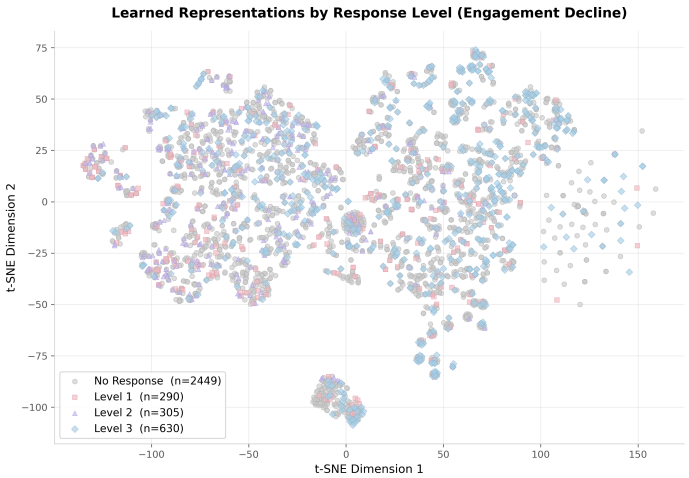


Figure 3.8. BasicLSTM hidden states during the engagement decline phase (weeks 16 - 23)

3.3 Semantic Embedding Integration

3.3.1 Motivation

The original diabetes prevention system encodes health message identifiers using binary numerical representations, which carry no semantic information about message content. This prevents the behavioral prediction model from recognizing relationships between messages that share similar themes, limiting its ability to generalize across semantically related content.

Semantic embeddings address this by representing message content in a continuous vector space where similar messages occupy nearby positions. Two embedding approaches are evaluated in this work: BERT uncased and SBERT. Both replace binary ID encoding during the feature construction phase and produce 768-dimensional vectors, but they differ in how those vectors are generated and what linguistic properties they capture. A dictionary with separate messages and questions entries indexed by their respective IDs is used as the structure for both files, and acts as a lookup table during feature construction.

3.3.2 Pre-computing Embeddings

Embeddings are generated once through a standalone script before any model training begins. This offline approach eliminates repeated model inference during training, as all message and question embeddings are computed and stored in advance as JSON files. Embeddings are generated from the English-language text of health messages and behavioral questions recorded in the study dataset files.

BERT uncased uses the pre-trained bert-base-uncased model [2] from Hugging Face's Transformers library without any fine-tuning. For each of the 57 health

messages and 32 behavioral questions, the text is tokenized using BERT's WordPiece tokenizer with a maximum sequence length of 128 tokens, passed through the frozen model, and a 768-dimensional embedding is produced through mean pooling over the final hidden layer activations, and their results are stored in JSON file. Mean pooling aggregates information across all token representations rather than relying on the [CLS] token alone, capturing the overall semantic content of the full input sequence.

BERT Embedding Generation with Mean Pooling

Input: health message or question text
Output: 768-dim semantic embedding vector

```
inputs ← tokenizer(text)
outputs ← BERT(inputs)
                                        ▷ Mean pooling over all token representations
embedding ← mean(outputs.last_hidden_state, dim = 1)
return embedding
```

SBERT uses the all-mpnet-base-v2 model [3] from the sentence-transformers library. Unlike the BERT pipeline above, SBERT encodes text through a single encode() call that handles tokenization, inference, and pooling internally. The model was specifically trained for sentence-level semantic similarity tasks, producing embeddings where semantically related sentences are geometrically closer in the vector space than regular BERT embeddings.

3.3.3 Integrating Embeddings into Feature Vectors

Embeddings are integrated into the behavioral prediction pipeline through a lookup and concatenation process during feature vector construction, and the BehaviorData class loads the pre-computed embeddings dictionary at initialization so all semantic representations are immediately available without requiring model loading during data preprocessing. For each behavioral sample in the `encode_row()`

method, the system looks up the 768-dimensional embedding for each message ID and question ID. An ID of zero, showing no message or question for that slot which maps to a zero vector of the same dimensionality. For all valid IDs, the embedding is retrieved from the pre-computed lookup table. A binary-encoding fallback exists in the code as a defensive measure but is never triggered in the reported experiments, since embeddings are generated in advance for all 57 messages and 32 questions. Each message ID that previously contributed six binary features now contributes 768 semantic dimensions, and each question ID expands similarly, allowing the models to use semantic relationships between health messages.

Message and Question ID Embedding Integration

Input: message_id, question_id, pre-computed embeddings
Output: feature vector combining message and question embeddings

if message_id = 0 **then**
    msg_emb ← zeros(768) ▷ No message for this slot
else if message_id ∈ embeddings **then**
    msg_emb ← embeddings[message_id]
else
    msg_emb ← binary_encode(message_id) ▷ Defensive fallback, not used in reported experiments
end if
if question_id = 0 **then**
    q_emb ← zeros(768) ▷ No question for this slot
else if question_id ∈ embeddings **then**
    q_emb ← embeddings[question_id]
else
    q_emb ← binary_encode(question_id) ▷ Defensive fallback, not used in reported experiments
end if
return [msg_emb, q_emb]

3.3.4 Experimental Setup

Three feature encoding conditions are compared: binary ID encoding, BERT uncased embeddings, and SBERT embeddings. The models are evaluated across all

three conditions to isolate the effect of embedding choice on an established architecture. All experiments use five random seeds, a learning rate of 0.001, a response percentage of 0.5, and a three-week history window. Train/test splits and seeds are held constant across feature conditions so that observed differences reflect the quality of the feature representation, not experimental variance.

CHAPTER IV

RESULTS AND EVALUATION

The experiments cover behavioral response prediction across three feature encoding conditions: 1. Binary ID encoding 2. BERT Uncased embeddings 3. SBERT embeddings. All models were evaluated at epoch 50, where test accuracy tends to stabilise across architectures before overfitting starts to affect the results. Results are averaged over 5 random seeds to reduce the influence of any single data partition.

4.1 Baseline Results

The train and test accuracy results for all models evaluated under binary ID encoding are shown in Table 4.1, which is a reference point for evaluating what semantic embeddings add in the sections that follow.

Model	Train Accuracy	Test Accuracy
LSTM	$60.8 \pm 0.4\%$	$57.1 \pm 2.3\%$
LogisticRegressor	$57.1 \pm 0.6\%$	$56.4 \pm 1.8\%$
BasicNN	$58.9 \pm 0.2\%$	$57.2 \pm 1.7\%$
BidirectionalGRU	$61.3 \pm 0.1\%$	$57.1 \pm 2.5\%$
AttentionWeightedMLP	$56.5 \pm 0.1\%$	$54.8 \pm 2.4\%$
TemporalConvolutionalNetwork	$62.2 \pm 1.6\%$	$56.7 \pm 2.3\%$
TransformerEncoder	$61.3 \pm 0.5\%$	$57.0 \pm 0.8\%$

Table 4.1. Train and test accuracy under binary ID encoding.

All models reach test accuracy in the range of 54-57% under binary ID encoding, with train-test gaps of approximately one to six points. Train accuracy itself only reaches 56-62%, and binary features clearly do not give the models enough signal to fit the training data well, let alone generalize from it.

4.2 BERT Uncased Results

Table 4.2 shows results for all seven models evaluated using BERT Uncased embeddings, where binary IDs are replaced with 768-dimensional vectors that carry the semantic content of each health message.

Model	Train Accuracy	Test Accuracy
LSTM	$79.4 \pm 0.2\%$	$57.1 \pm 2.6\%$
LogisticRegressor	$60.5 \pm 0.3\%$	$58.7 \pm 1.8\%$
BasicNN	$60.8 \pm 0.5\%$	$58.8 \pm 1.8\%$
BidirectionalGRU	$63.8 \pm 0.9\%$	$56.8 \pm 2.0\%$
AttentionWeightedMLP	$61.0 \pm 1.0\%$	$56.5 \pm 2.2\%$
TCN	$61.8 \pm 1.1\%$	$57.4 \pm 2.1\%$
TransformerEncoder	$59.2 \pm 1.4\%$	$57.8 \pm 1.9\%$

Table 4.2. Train and test accuracy under Bert Uncased.

The most notable result in Table 4.2 is how the LSTM behaves under BERT encoding. Its training accuracy reaches 79.4% while test accuracy stays at 57.1%, which is a quite big difference around 22 points and substantially larger than what is seen under binary ID encoding. This clearly shows that the richer BERT feature space gives the LSTM enough capacity to overfit the training data without learning patterns that generalize to new examples. The other models show more modest train-test gaps of three to seven points and seem to generalize more effectively to unseen data.

Among all architectures, BasicNN achieves the highest test accuracy at 58.8%. LogisticRegressor follows closely at 58.7%, while BidirectionalGRU, AttentionWeightedMLP, TCN, and TransformerEncoder all perform in the range of 56% to 58%. The TransformerEncoder also shows the smallest train-test gap, at about 1.4 percentage points, showing less overfitting compared with the other neural architec-

tures.

4.3 SBERT Results

Lastly, table 4.3 demonstrates results under SBERT embeddings, which is a fine-tuned variant of BERT that is specifically designed to produce more semantically coherent sentence representations and was tested here as an alternative to BERT Uncased for encoding the health message content.

Model	Train Accuracy	Test Accuracy
LSTM	$72.3 \pm 0.3\%$	$57.8 \pm 2.5\%$
LogisticRegressor	$60.1 \pm 0.4\%$	$58.4 \pm 1.7\%$
BasicNN	$60.5 \pm 0.5\%$	$58.7 \pm 1.5\%$
BidirectionalGRU	$65.4 \pm 1.1\%$	$58.1 \pm 2.2\%$
AttentionWeightedMLP	$57.7 \pm 1.2\%$	$55.6 \pm 1.1\%$
TCN	$66.1 \pm 1.0\%$	$57.2 \pm 2.0\%$
TransformerEncoder	$62.6 \pm 0.9\%$	$57.3 \pm 2.5\%$

Table 4.3. Train and test accuracy under SBERT.

Results under SBERT follow a broadly similar pattern to what was seen under BERT Uncased, and the overall picture does not change much across architectures. The LSTM still shows a big train-test difference, though it narrows slightly from 22 to 14 points, and this shows that SBERT representations are somewhat less susceptible to memorisation. BasicNN again achieves the highest test accuracy at 58.7%, which is consistent with what it achieved under BERT Uncased. However, BidirectionalGRU improves slightly to 58.1% and AttentionWeightedMLP drops to 55.6%, the lowest test accuracy across all conditions.

No single architecture consistently outperforms BasicNN by a meaningful margin across either embedding condition, and test accuracy for most models falls within

roughly three points of each other regardless of which architecture is used, and this points more to dataset size and task difficulty as the main issues on performance instead of model complexity.

4.4 Message Embedding Structure

Figure 4.1 shows the projection of BERT embeddings, where the three categories form visually distinct regions. Consumption messages cluster toward the upper-right, Exercise messages form a tight group in the upper-center, and Knowledge messages spread across the lower center and left side of the plot. SBERT shows a similar clustering pattern, although the clusters appear in different positions and are more dispersed compared to BERT Uncased.

In the binary ID encoding (left plot), all categories are completely mixed with no visible structure, which is expected since a binary representation of a sequential integer carries no semantic meaning and adjacent IDs may correspond to entirely different behavioral domains. In contrast, both BERT Uncased and SBERT embeddings (middle and right plots) show clearly separated category clusters for the same 57 messages, highlighting the limitations of binary encoding and the structured semantic space produced by embedding-based representations.

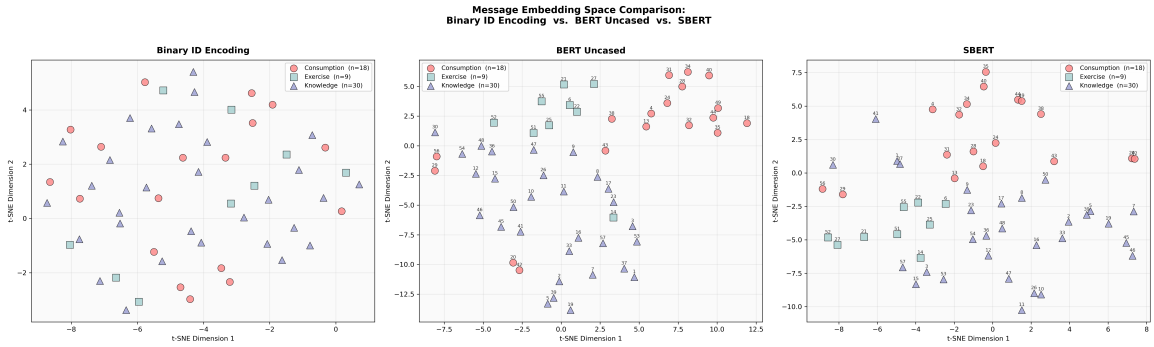


Figure 4.1. Comparison of all 57 health messages under binary ID encoding (left), BERT Uncased embeddings (middle), and SBERT embeddings (right)

Looking at Figure 4.1 again, four messages (20, 29, 42, 56) in BERT Uncased plot are all labeled as Consumption but sit far from the main Consumption cluster and appear near the Knowledge region instead. All four are about tobacco and smoking, with language focused on health consequences and diabetes complications rather than dietary behavior. For instance, message 42 warns "If you have diabetes, smoking cigarettes and chewing tobacco makes diabetes complications (heart disease, nerve damage) more severe." It reads more like a Knowledge message, and BERT assigned it to Knowledge instead.

The issue is not with the embedding method but with some ambiguity in the original categorization, since messages labeled as Consumption due to tobacco avoidance contain no food-related content and this creates a mismatch between the labels and what the embeddings pick up. This conflicting signal likely contributes to the lower and less consistent accuracy seen for the Consumption category, and relabeling these messages as Knowledge or introducing a separate subcategory for health consequence content could improve signal clarity, though this is left for future work. Both BERT and SBERT capture structure that binary encoding misses.

4.5 Embedding Evaluation

To test whether BERT embeddings introduce measurable semantic structure into the message feature space, silhouette scores [18] were computed for all three encoding schemes. The silhouette score for a sample i is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where $a(i)$ is the mean distance to all other points in the same cluster, and $b(i)$ is the mean distance to points in the nearest neighboring cluster. Scores range from -1 to $+1$, where values near $+1$ indicate well-separated clusters and values near 0 indicate no meaningful structure. The three behavioral categories (Consumption, Exercise, Knowledge) were used as cluster labels, and cosine distance was applied as the similarity metric, which is standard practice for high-dimensional embedding spaces. Scores were computed over all 57 health messages.

Binary ID encoding produced a score of -0.087, meaning messages are on average closer to points outside their assigned category than within it. Both BERT Uncased and SBERT produced a score of approximately $+0.126$, representing a consistent directional improvement over binary encoding. The modest absolute value is expected given genuine semantic overlap between categories, most notably the four tobacco-related messages labeled as Consumption that are semantically closer to Knowledge content, as discussed earlier.

A parallel analysis on the 32 behavioral questions produced the same directional pattern. Binary encoding scored -0.052, confirming no meaningful structure. BERT Uncased scored $+0.125$ and SBERT scored $+0.285$, with the higher SBERT score reflecting its stronger sentence-level representations for short question text. These figures should be read with caution given the small number of questions per

category. The consistency across both messages and questions confirms that both embedding models recover semantic structure that binary encoding cannot capture.

To further quantify the semantic structure in these embeddings, a logistic regression classifier was used to test whether message embeddings contain enough category information to distinguish between behavioral domains. Classification was performed using leave-one-out cross-validation (LOO-CV), with each iteration training on 56 messages and testing on a single held-out message. Random guessing would score 33.3% across the three categories. Binary ID encoding scored 36.8%, which is close to random performance, while BERT Uncased and SBERT achieved much higher scores of 89.5% and 91.2%, respectively. At the category level, Consumption was the hardest to classify under both models, reflecting the semantic ambiguity of the tobacco-related messages identified earlier. This is supported by three complementary findings: the t-SNE analysis revealed that four tobacco-related Consumption messages sit semantically closer to Knowledge content, the linear separability test showed Consumption had the lowest per-category classification accuracy (78% for both BERT Uncased and SBERT), and the category combines two behaviorally distinct domains, dietary habits and tobacco avoidance, whose response patterns are not necessarily correlated. These factors create conflicting signals during training that no model architecture fully resolves.

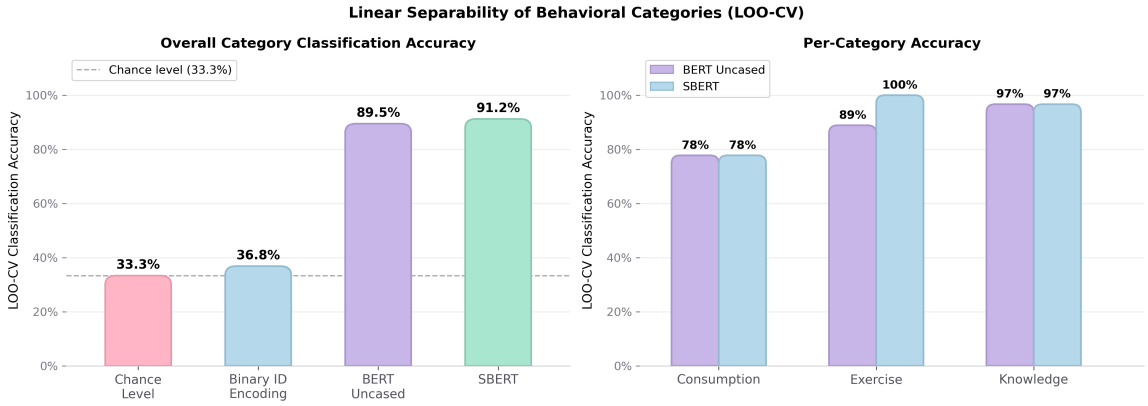


Figure 4.2. LOO-CV classification accuracy for binary ID encoding, BERT Uncased, and SBERT (left), and per behavioral category (right).

4.6 Limitations and Future Work

With only 57 health messages and a heavily imbalanced question distribution where Exercise accounts for just 3.5% of total interactions, the results give a general sense of performance instead of definitive conclusions. Leave-one-out cross-validation was used to handle the small sample size, but it cannot fully compensate for a corpus this limited. No hyperparameter search was conducted and all architectures were evaluated under a fixed epoch 50 cutoff, which might not be the best stopping point for every model. Neither BERT Uncased nor SBERT was fine-tuned on health-domain text, thus the embeddings reflect general language patterns and not diabetes-specific semantics. Specialized models like BioBERT and ClinicalBERT exist for biomedical and clinical text, however, they are trained on technical literature and clinical notes and not the plain everyday language used in this work. A model fine-tuned directly on text for people facing diabetes interventions may not currently exist as a standard resource, and this is something future work could look into.

CHAPTER V

CONCLUSION

Binary encoding functioned more as a numbering system than as a meaningful feature representation, and the results of this work make that distinction concrete. Replacing message identifiers with BERT-based semantic embeddings did not produce dramatic accuracy gains, but it changed what the model could see, and that difference carries more weight than the accuracy numbers alone suggest.

The accuracy improvements were minor, and the data constraints largely explain why. Considering the number of participants and a response rate below 40%, there was not enough signal for higher-capacity architectures to outperform the simpler baselines, and the results reflect that consistently across all conditions. BasicNN and LogisticRegressor both benefited from the richer representations and achieved slight improvements respectively, while the four new architectures fell in the same range as the simpler baselines, showing at this data scale, representation quality matters more than model complexity.

Additionally, the semantic analysis provides stronger evidence for the value of the embeddings. Silhouette scores moved from negative under binary encoding to a consistent positive under both BERT Uncased and SBERT, and leave-one-out classification went from 36.8% to 89.5% and 91.2% respectively, which shows the embeddings recover genuine category structure that binary encoding is structurally incapable of capturing. The four tobacco-related messages that appeared consistently near the Knowledge cluster were an interesting finding, not because they disrupted the model but because they revealed a labeling ambiguity that had always been present in the data and was not visible before.

REFERENCES CITED

[1] S. E. Kinsey et al., *Building a Personalized Messaging System for Health Intervention in Underprivileged Regions Using Reinforcement Learning.* IJCAI, 2023, 6022–6030.

[2] J. Devlin et al., *Bert: Pre-training of deep bidirectional transformers for language understanding*, Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, 4171–4186.

[3] N. Reimers and I. Gurevych, *Sentence-bert: Sentence embeddings using siamese bert-networks*, Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), 2019, 3982–3992.

[4] L. Rasmy et al., *Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction*, npj Digital Medicine 4 (2021), no. 1, 86, DOI: 10.1038/s41746-021-00455-y.

[5] E. Alsentzer et al., *Publicly Available Clinical BERT Embeddings*, Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, 72–78, DOI: 10.18653/v1/W19-1909.

[6] M. Schuster and K. K. Paliwal, *Bidirectional recurrent neural networks*, IEEE transactions on Signal Processing 45 (1997), no. 11, 2673–2681.

[7] K. Cho et al., *Learning phrase representations using RNN encoder-decoder for statistical machine translation*, arXiv preprint arXiv:1406.1078 (2014).

[8] S. Bai, J. Z. Kolter, and V. Koltun, *An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling.* CoRR abs/1803.01271 (2018), arXiv: 1803.01271.

[9] A. Vaswani et al., *Attention is all you need*, Advances in neural information processing systems 30 (2017).

[10] U. Madububambachu, A. Ukpebor, and U. Ihezue, *Machine learning techniques to predict mental health diagnoses: A systematic literature review*, Clinical Practice and Epidemiology in Mental Health: CP & EMH 20 (2024), e17450179315688.

[11] A. Rajkomar et al., *Scalable and accurate deep learning with electronic health records*, NPJ digital medicine 1 (2018), no. 1, 18.

[12] T. Mikolov et al., *Efficient estimation of word representations in vector space*, arXiv preprint arXiv:1301.3781 (2013).

[13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning representations by back-propagating errors*, nature 323 (1986), no. 6088, 533–536.

[14] Y. Bengio, P. Simard, and P. Frasconi, *Learning long-term dependencies with gradient descent is difficult*, IEEE transactions on neural networks 5 (1994), no. 2, 157–166.

[15] S. Hochreiter and J. Schmidhuber, *Long short-term memory*, Neural computation 9 (1997), no. 8, 1735–1780.

[16] J. Chung et al., *Empirical evaluation of gated recurrent neural networks on sequence modeling*, arXiv preprint arXiv:1412.3555 (2014).

[17] D. Bahdanau, K. Cho, and Y. Bengio, *Neural machine translation by jointly learning to align and translate*, arXiv preprint arXiv:1409.0473 (2014).

[18] P. Rousseeuw, *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*, J. Comput. Appl. Math. 20 (1987), 53–65.