Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 206 (2022) 121–131

Procedia Computer Science

www.elsevier.com/locate/procedia

International Society for Research on Internet Interventions 11th Scientific Meeting

Technology-Assisted Motivational Interviewing: Developing a Scalable Framework for Promoting Engagement with Tobacco Cessation Using NLP and Machine Learning

Ahson Saiyed,[a*] John Layton,[b] Brian Borsari,[c,d] Jing Cheng,[b] Tatyana Kanzaveli,[a] Maksim Tsvetovat,[a] Jason Satterfield[b]

[a]*Open Health Network*
1049 C EL Monte Ave #12
Mt. View, CA 94040

[b]*Division of General Internal Medicine, University of California San Francisco*
1701 Divisadero St., Suite 500
San Francisco, CA 94115

[c]*Mental Health Service (116B), San Francisco VA Medical Center*
4150 Clement St.,
San Francisco, CA 94121

[d]*Department of Psychiatry and Behavioral Sciences, University of California, San Francisco*
401 Parnassus Ave.,
San Francisco, CA 94143

---

* Corresponding author.
E-mail address: ahson@ocn.io

"Peer-review under responsibility of the scientific committee of the International Society for Research on Internet Interventions 11th Scientific Meeting"
10.1016/j.procs.2022.09.091

Abstract

Motivational interviewing (MI) improves readiness for smoking cessation but can be time-intensive, require substantial expertise, and patients must still be linked with evidence-based cessation programs sensitive to local resources and patient preferences. Technology-assisted MI may provide a more efficient way to promote readiness and facilitate behavior change.

This study developed the Technology Assisted Motivational Interviewing Coach (TAMI), a digital conversational agent that incorporates machine learning models to deliver MI for tobacco cessation and create tailored quit plans. This manuscript describes and evaluates the architecture and nested machine learning models within TAMI leveraged during the pilot clinical trial.

Keywords: smoking cessation, motivational interviewing, digital health, chatbot, machine learning models

1. Introduction

Smoking continues to be the single greatest cause of premature morbidity and mortality with over 480,000 US deaths per year, disproportionately affecting marginalized and vulnerable populations [1-2]. Of the 34 million Americans who continue to smoke, 70% are interested in quitting but only 20% express the intention to quit in the next 30 days [3-5]. For those who attempt to quit, only 31% will use an evidence-based cessation intervention (EBSCI) causing most quit attempts to end in failure [4]. Despite notable innovations in cessation delivery models, utilization of evidence-based smoking cessation interventions (EBSCI) remains low and cessation rates are disappointing. The field needs more efficient and effective, strategies to address 2 primary problems: 1) enhancing patient motivation to quit and 2) building a tailored, actionable, evidence-based quit plan sensitive to local resources.

Motivational interviewing is a communication style in which the therapist uses MI-consistent (MICO) therapist skills (e.g., reflections, affirmations) to evoke and strengthen the patient's own argument for change, labeled change talk, while acknowledging his or her reasons for staying the same (sustain talk) [6]. MI has been demonstrated to improve readiness for smoking cessation [3] [7-8]; however, MI can be time intensive, require substantial expertise attained through extensive training and coaching, and patients must still be linked with evidence-based cessation programs sensitive to local resources and patient preferences. A digital intervention using MICO skills to evoke and respond to participant change language could provide an effective and efficient way to explore patient ambivalence, improve readiness, and elicit EBSCI preferences so that a more effective quit plan can be made.

Although MI digital tools have been used for smoking cessation [9-12] and alcohol and drug treatment [13-19], they have been disappointing in quality and lack technical sophistication [20-22]. A systematic review of 41 digital MI studies found that while acceptability was high and outcomes tended to be positive, the average MI comprehensiveness score [23-24] was a low 4.78 out of 15 suggesting a need for higher quality tools [25]. Our team recently developed a chatbot ("TAMI") that delivers inexpensive MI to enhance motivation for smoking cessation then assists patients with selecting EBSCI's and creating a tailored quit plan [26-30]. Smoking cessation intervention preferences are captured throughout the conversation and compiled into an emailed pdf summary to provide patients access to their preferred treatments. Client usage analytics and conversational history are delivered in the form of a dashboard, providing insight to metrics to gauge program success and the need for human intervention.

This manuscript focuses on the technical development and architecture of the resulting digital tool and the underlying machine learning models it contains. Due to its technical and clinical complexity, we primarily focus on the development and evaluation of the motivational interviewing module (MI module). In order to contextualize the functional nuances of the MI module within the greater system in which it operates, we first describe the overall chatbot structure and function followed by a description of the base technical architecture of the system as a whole. Lastly, we present the unique natural language understanding (NLU) architecture of the MI session including the development and evaluation of the language classifiers and chatbot response system.

2. Overview of Chatbot Architecture and Design

2.1 Chatbot Structure and Function

The TAMI chatbot was built to perform three key tasks conceptualized as three distinct modules: onboarding, motivational interviewing, and referral to treatment (outboarding). These three modules were selected based on the gold standard of cessation treatment, the "5 A's" [31], where our onboarding encompasses the "Ask, Advise, and Assess" stages then "Assists" the patient by either sending them to the MI module (for patients not ready to quit) or to the outboarding module (for those ready to quit) where patients can develop a tailored quit plan and set a quit date. Currently, TAMI sends automated reminders and tips before and after the patient's quit date but no formal follow-up appointment is "Arranged." In recognition of the time and effort often required to build intrinsic motivation for smoking cessation, we designed TAMI with the expectation that patients would return to the chatbot for multiple conversations over time - a decision that was clinically valid but also more technically complex.

Across all three distinct modules, a structured series of dialogue transcripts for the chatbot were created by smoking cessation experts to guide the overall conversation. The dialogue transcript for the onboarding module features a brief greeting, followed by an assessment of smoking status and readiness to quit smoking. The dialogue transcript for the outboarding module walks the patient through a review of EBSCIs and prompts the user to create a tailored quit plan that can then be exported in PDF form for future reference. Below, we describe the architecture that drives these user-bot interactions followed by an in-depth review and discussion of the MI module.

2.2 Base Architecture

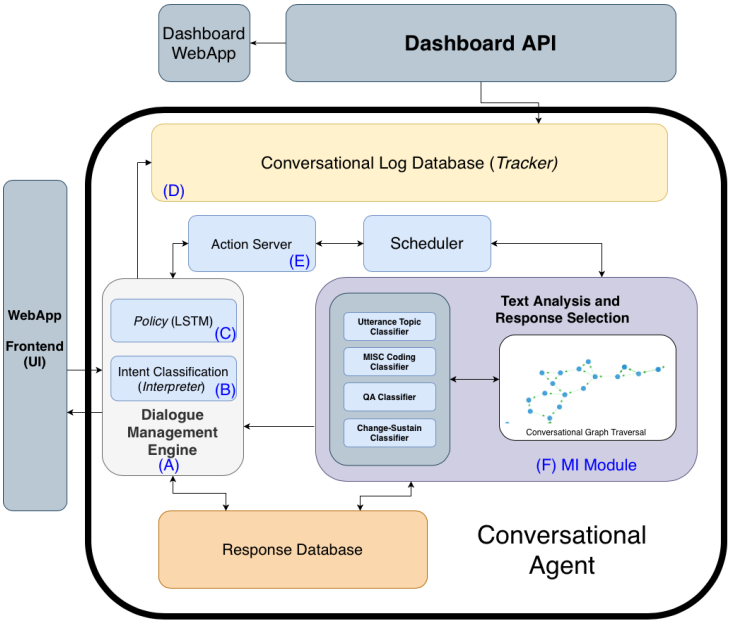

Fig 1. TAMI bot architecture

In order to build an end-to-end MI conversational agent, we relied on open-source software for dialogue management (Fig.1-A). The open-source Python package, Rasa, frames dialogue management as a classification problem [32]. Incoming messages, received through our web application, first interact with a configured Rasa Core server, which contains an intent classification model (Rasa Core *Interpreter*) (Fig.1-B) attempting to understand the meaning of incoming messages and map them to learned intents from training on the expert-designed dialogue transcripts [32]. After characterizing the incoming utterance, Rasa Core's *Policy* (Fig.1-C) module predicts the appropriate bot action to take by passing the predicted intent along with user interaction history into a Long-Short Term Memory Neural Network (LSTM) model trained on sequences of example user-bot conversational volleys from the dialogue transcripts. The LSTM then identifies which module the user is currently in and what action to take considering their position in that module (Fig.2-A) [32].

User utterances during the onboarding and outboarding modules are largely driven by button and natural text input and follow a simple, linear conversational path. User button and text responses in these modules are stored as slot values in the Rasa Core *Tracker* object [32]. The *Tracker* object maintains the conversation state, and is responsible for logging conversational events, and reproducing the conversational state for any given user (Fig. 1-D). Data from the *Tracker* object is extracted and compiled by the action server (E) into a PDF quit plan summary at the end of the outboarding stage. A dashboard displaying user engagement and conversational logs is also created from data in the *Tracker object*.

Rasa Core and its components, *Interpreter*, *Policy* and *Tracker*, were designed to be modular and to be used independently from Rasa NLU [32]. We combine Rasa Core with our custom NLU (see Section 3 below) and response generation service. Unlike the onboarding and outboarding modules, during the MI module, the LSTM delegates response selection to our custom natural language understanding (NLU) and response retrieval method. Our custom NLU solution for the MI module (Fig.1-C, Fig.2-B) begins by evaluating incoming user utterances using transformer-based classifiers that are essential to the delivery of motivational interviewing. Details about the training methodology for these classifiers and how machine learning model output is leveraged for discriminative response selection are further elaborated in the section below.

3. Natural Language Understanding for Motivational Interviewing

3.1 Rationale for Our Approach

One of the greatest obstacles in creating a conversational interface that is accepted by the users is conversation flow. Human conversation relies on the depth and richness of shared cultural experiences, and the ability of the human mind to rapidly shift topics and cogently converse on all related and unrelated manners. There have been multiple moderately successful attempts to create a generic chatbot that can talk about numerous topics (e.g., informed by Wikipedia); however, these bots frequently fail when directed to follow a certain therapeutic structure or topic. In order to effectively administer MI as a digital health intervention through a conversational agent, it is necessary to both analyze and understand the incoming user-generated text and to respond appropriately with a therapeutic focus and conversational topic relevance. With the use of custom NLU models for recognizing change-related language and smoking-related conversational topics, our approach is better able to identify and respond to the most clinically-relevant user utterances throughout the course of the conversation.

3.2 Overview of our NLU Models for MI

In our approach, we frame holistically characterizing a user's current willingness to change and delivering an MI-consistent response as an ensemble text classification problem (Fig.2 -B) whose output is fed into an information retrieval system (Fig. 2-C). As a result, the quality of our text classification methodology directly impacts our ability to understand user behavior and deduce a high-quality response (Fig. 2-D).

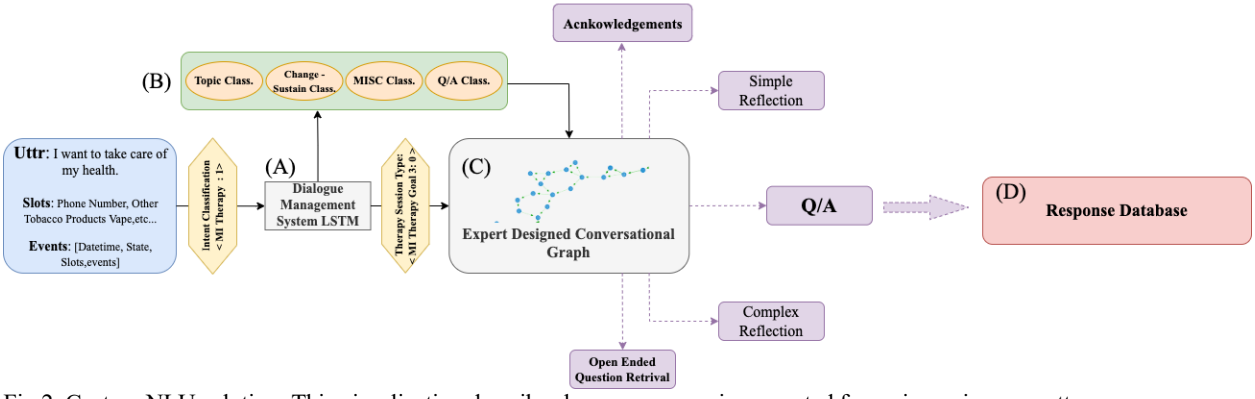

Fig 2. Custom NLU solution; This visualization describes how a response is generated for an incoming user utterance.

Classical machine learning methods for supervised text classification, such as those reliant on handcrafted feature engineering, have been outperformed by deep learning techniques centered around leveraging machine-generated embeddings for representations of text [33]. Fine-tuned variants of large-scale transformer-based Pre-Trained Language Models (PLM) have demonstrated the ability to produce high-quality embeddings and SOTA performance on many downstream NLP tasks, including text classification [33]. More recently, studies have shown that the utility of PLM embeddings can be further enhanced as input to various Graph Neural Networks architectures [34].

Our custom NLU model-based approach evaluates incoming user utterances using the following transformer-based classifiers: change talk-sustain talk classification, utterance-topic classification, question-answering classifier, and motivational interviewing skills code classifier. Detailed descriptions of each classifier along with how they were developed and evaluated can be found in their respective sections below.

3.3 Change Talk- Sustain Talk Classification

Change Talk is any language used by the patient to describe their motivation, ability, or reasons for altering a specific behavior. In contrast, Sustain Talk is language used to indicate a desire to maintain that behavior or to preserve the status quo [6]. MI process research has been eloquently summarized by Magill and Hallgren [35] who argue for the use of MI-consistent (MICO) skills to strategically and simultaneously evoke and strengthen Change Talk while softening Sustain Talk in order to effectively promote behavior change [36-39]. Change Talk-Sustain Talk (CT/ST) classification performance thus plays a central role in a technology-assisted MI system's ability to support end-to-end automated MI-consistent conversations that build intrinsic motivation to change.

3.3.1 Training methodology and evaluation for CT/ST

Datasets containing expert annotated MI interview transcripts across alcohol and smoking-related client sessions were preprocessed and a total of 20,890 coded utterances were collected for model development. The utterances were labeled across 3 classes, “change talk”, “sustain talk”, or “follow neutral” (client language not related to smoking behavior). Oversampling of underrepresented classes was carried out to balance class distribution. Utterances were tokenized and prepared for language model ingestion.

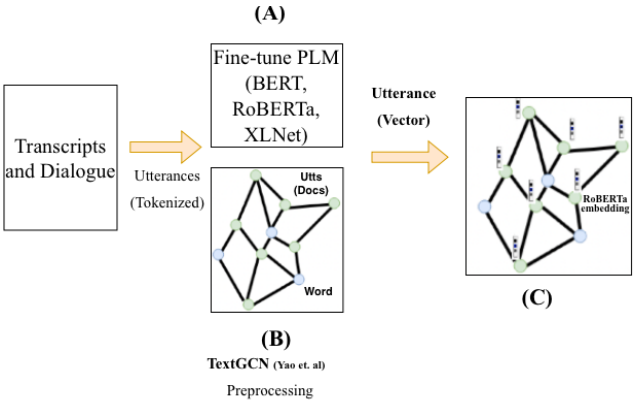

Fig 3. Graph representation of utterances from MI transcripts with embeddings generated from PLMs as node attributes

Utterance level representations for the collected dataset were generated from fine-tuning large-scale pre-trained language models, BERT [40], RoBERTa [41], and XLNet [42] (Fig.3 -A) accessed through HuggingFace [43,44]. Language models were trained with a learning rate of 4e-7 with a batch size of 8 [45-46].

Utterances and words are treated as nodes in a heterogeneous graph, as described in TextGCN (Fig.3-B) [46]. Utterance node attributes are populated with 128-dimensional utterance level representations generated from fine-tuned language models (Fig.3-C). Edges between nodes are assigned and weighted considering the semantic similarity between nodes [45-46].

The learning rate was initialized to 1e-3. The GCN model had 2 GCN layers with 16 attention heads. The λ parameter, which defines the interpolation between BERT and GCN predictions, was found to be most favorable at 0.5.

Binary and multi-class text classification was conducted by passing the pre-trained language model embeddings at the utterance level through a feed-forward neural network. Binary node classification and multi-class node classification were performed across Graph Convolutional and Graph Attention Neural architectures. Model performance was evaluated using a Monte Carlo cross-validation scheme (n=10) across an 80:10:10 train, validation, and test split. Performance metrics such as accuracy and F1-score are averaged across iterations. Precision-recall curves were used for hyper-parameter optimization.

3.4 Utterance Topic Classification

Human-administered therapy allows open-domain conversation while maintaining focus on therapeutic goals. SOTA open-domain conversational agents still struggle with handling unseen topics and ambiguous context, as initial errors in the identification of user domain causes subsequent errors to cascade throughout the system and ultimately degrade user experience [46]. Relevant and high-quality utterance topic classification is crucial in providing meaningful end-to-end automated MI-consistent conversation consisting of multiple volleys.

Our framework allows for open-domain conversation to be funneled into a smoking-cessation therapeutic context while attempting to maintain coherent dialogue and engaging long-form conversation. While solely keyword-based classification has limitations, our approach leverages both keyword and semantic information in user utterances for topic classification.

In order for TAMI to respond to a variety of smoking-related conversational topics, comments and posts on smoking-related subreddits on Reddit were scraped and labeled by volunteers to create a training dataset. This dataset was used to fine-tune a PLM and create a topic classification model.

3.4.1 Training methodology and evaluation

14,106 unique utterances were extracted across 81 subreddits and preprocessed to create a dataset for training a topic classification model with 81 topics. This dataset was programmatically processed to remove non-essential text and characters using a variety of string-matching techniques. A second dataset with 7,702 utterances with 77 classes was hand-annotated and curated by volunteers to remove noise from the original scraped dataset and improve real-world performance. Similar to the Change and Sustain text classification methodology, utterance level representations were generated from fine-tuned language models and used in downstream text classification [48,49].

Model performance was evaluated using a Monte Carlo cross-validation scheme (n=10) across an 80:10:10 train, validation, and test split. Performance metrics such as accuracy and F1-score are averaged across iterations. Precision-recall curves were used for hyper-parameter optimization.

3.5 MISC Coding Classification

Motivational Interviewing Skill Code (MISC) is an iterative coding system developed over the past 20 years to measure adherence to MI and is used to describe interactions between therapist and client during a session [47]. Historically, in clinical and research settings, MISC codes of therapist and client language have been used to describe and evaluate MI.

With a high-performing MISC classification system, the NLU decision engine can understand the MISC category of incoming utterances, predict the appropriate MISC category of responses, and take steps to directly incorporate MI theory into its response selection logic. The MISC coding classifier is an XLnet model fine-tuned on MI transcripts labeled for 38 classes across patient and therapist responses (e.g. ADP: Advice with permission; ADW: Advise without permission; AF: Affirm; C: Commitment; QU: Question; QUC: Question Closed). Of note, our MISC classification system did not include change and sustain talk which had its own independent classification model. The Next-MISC prediction model was attempted by training an LSTM, fed sequences of MISC coding volleys of variable length collected from patient-therapist transcripts. Currently, the MISC category is a node attribute on the expert-curated conversational graph and leveraged during response selection.

3.6 Question-Answering Model

Embedded within the custom NLU service is a Question and Answering (QA) model that detects whether incoming utterances and their conversational context are semantically similar to frequently asked questions common to MI-oriented smoking cessation counseling or anticipated bot-user interactions. These questions and their responses span a range of topics and were curated by experts in the field. The necessity of this QA Model was determined by early user behaviors that affected conversational flow and user experience.

We evaluated three strategies to provide coverage for 215 Question-Answer pairs across topics: general questions about smoking, general questions about TAMI, questions about over-the-counter cessation aids and medications, questions about planning to quit, how to reach out for help, and miscellaneous. The first was an extractive model, attempting to produce the appropriate answers given question-context pairs with answer start and end date. The second was a text classification-based

approach, which fine-tuned a RoBERTA-large model, with each question labeled across 91 classes. The third approach leveraged string matching to identify questions nested within incoming utterances.

In order to confidently be implemented in practice, the correctness and relevance of responses generated from the QA model needed to be further assessed. The string matching approach offered the fewest false positives and, crucially in a clinical setting, allowed control over responses. The text classification approach, while initially promising due to the breadth of coverage, had frequent out-of-distribution high confidence predictions preventing real-world implementation. In future iterations, we hope to train a QA model with a larger corpus and with data collected throughout the trial to improve relevance and limit out-of-distribution high confidence predictions.

4. Putting it All Together: Chatbot Response Generation and Selection

Following the creation and validation of the CT/ST and topic classifiers, we created an informational retrieval system capable of selecting and delivering motivational interviewing consistent (MICO) responses. Our approach adhered to core MI principles by 1) evoking and amplifying change talk from the user, 2) promoting compassion, respect, and user autonomy in line with “MI spirit”, and 3) employing MICO conversational skills (open-ended questions, affirmations, and reflections) to further assist the user in resolving their ambivalence about smoking cessation. As such, output from the NLU classifier was utilized to support these aims and select an appropriate expert-generated response.

According to MI theory, evoking and strengthening change talk through the use of MICO skills precedes subsequent changes in behavior. Our information retrieval system was designed to capitalize upon this phenomenon by parsing user input for CT and ST about smoking (opposed to other behaviors such as alcohol use or weight loss), identifying the topic associated with utterance fragments scoring highest for smoking CT, and delivering a response that encourages the user to elaborate upon their previous change statement or think about their smoking in a new light. Consistent with the MI processes of focusing and evocation [6], leveraging the NLU model outputs, we rank-ordered the list of available responses by coherence and relevance to smoking. In this way, the machine learning models directed the MI-module to continue the conversation about smoking along a path most likely to lead to more talk of changing smoking and prevented the conversation from going off-topic.

4.1. Response Matrices and Conversational Graphs

Having applied the NLU models to identify the most clinically relevant utterance fragment, we then had to enable the MI-module to respond to it using MICO skills of open-ended questions, affirmations, and reflections. However, it was not feasible to build a natural language generative model capable of creating and delivering these MICO skills independently. Instead, we employed response matrices, indexed by response type (i.e., MICO skill), topic classification, and valence of smoking CT/ST. We designed the information retrieval system to query the matrices for the topic and CT/ST valence associated with the most clinically relevant utterance fragment and deliver a pre-written MI expert response tailored for those inputs. Each response was designed to be consistent with MI-spirit and technical skills. Several different MI expert responses were written for each matrix cell, allowing for continued MI-consistent conversation without repetition.

The last challenging step was to create a decision engine that would pair the most clinically relevant utterance fragment to the most appropriate response-type matrix–whether open-ended question, reflection, or affirmation would be used. Our first approach to this problem consisted of a fourth classifier that predicted, given a sequence of MISC coding labels across conversational history, the next appropriate counselor response type. Upon evaluation of this conversational approach during alpha testing, it became clear that, while functional, the conversational flow lacked linearity, structure, and depth. With this in mind, we created a hybrid approach that employed MI expert-designed conversational graphs. Each graph was designed with a thematic and clinically relevant aim and a predetermined series of key questions designed to progress the conversation through that theme. These conversational graphs also allowed MI experts to query the matrices at various predetermined graph locations. In this way, we were able to maintain the structure of focused MI conversation while flexibly allowing for variance in a user's provision of CT or ST and their desired topic of conversation.

The final approach taken to handle user input and produce an MI consistent response is described in Figure 2. Incoming utterances are scored across the different classifiers and the predicted output is stored (Fig.2-B). The expert-designed conversational graph is then queried based on model output parameters and users are selected to follow a conversational aim, described by a path of unvisited nodes along the expert-curated conversational graph (Fig.2-C). The conversational path then outlines the next appropriate response type given the previous utterance model classification. The corresponding response type matrix is then queried considering the relevant model parameters and the returned response is delivered to the user (Fig.2-D).

4.2 Evaluation Methodology for Response Generation

Chatbot response generation performance will be assessed for preliminary feasibility and acceptability using data collected from an ongoing pilot clinical trial, during which users complete one guided initial onboarding conversation at the beginning of the study then receive regular reminders to speak with the chatbot on their own over the course of six months. Transcripts of participant conversations extracted from the conversational log database (Fig.1-D) will be reviewed in aggregate by our team and assessed for MI comprehensiveness using a 15-item scale used with other digital MI studies [25]. Other user analytics to be retrieved, include the number of conversations with TAMI and the number of participants that completed a quit plan. At three-month and six-month time points, study participants will complete follow-up surveys that provide smoking-related outcomes, such as cessation or initiation of treatment, as well as participant satisfaction ratings of TAMI from one to five stars.

5. Results

Utterance classification of CT-ST by PLM embeddings alone was outperformed by GCN architecture considering an utterance-word heterogeneous graph with PLM document level embeddings as node features (see Table 1). In comparison to expert coders, the automated change/sustain classifier achieved an 84% accuracy rate and an F1-score of 0.88 as a binary node classification task (see Table 2). Due to the variety of language and overrepresentation of "Follow/Neutral" class utterances, multi-class node classification had an accuracy of 0.74 and an F1-score of 0.75. Of all PLM approaches, RoBERTa produced the highest performing embeddings for CT-ST text classification.

Topic classification leveraging fine-tuned XLNet performance on real-world scraped data was a lower 65% (Table 3), triggering the need for more manual annotation and label correction. Topic classification on a subsequent much smaller hand-curated dataset with heavy pruning of scraped data performed better, however, breadth of coverage and larger contextual information were sacrificed for optimizing performance metrics.

Table 1. Change Talk, Sustain Talk or Follow Neutral Classifier (n=3)

Model	Accuracy	F1
BERT	.65	.668
RoBERTa	.70	.715
XLNet	.63	.65
RoBERTaGCN	.74	.75
XLNetGCN	.73	.73
RoBERTaGAT	.70	.68

Table 2. Change Talk, Sustain Talk (n=2)

Model	Accuracy	F1
RoBERTaGCN	.8397	.8838
RoBERTaGAT	~.77	.81

Table 3. Topic Classification

	(n=81), 14k documents		(n=77), 7k documents	
Model	Accuracy	F1	Accuracy	F1
XLNet	.65	.63	.81	.79

Preliminary data analysis of the TAMI pilot trial at 3 months shows a strong MI comprehensiveness rating score of 13/15 [25]. Of the n=34 smokers currently using TAMI, 44% have completed a quit plan, 26% have initiated treatment and 15% report successful cessation at a 3-month follow-up. Test subjects have rated the chatbot with 3/5 stars indicating a need for an improved user experience.

6. Conclusion

The three chatbot modules - onboarding, MI, and outboarding - successfully completed the common clinical tasks associated with smoking cessation screening, counseling, and referral. As anticipated, the data gathering in the onboarding stage and the guided creation of a tailored quit plan presented few technical challenges and were well received by users. For the more challenging MI module, the machine learning classifiers (CT/ST and topic) demonstrated strong performance during initial testing. In our preliminary work, we have found that TAMI can successfully stay on the relevant conversational topic while using MICO skills to explore a user's ambivalence about quitting tobacco. Repeated assessments for readiness to quit allow the user to transition to the final outboarding module (creating a quit plan) at the appropriate time.

6.1 Challenges

CT/ST Classifier:

Although approximately 21,000 utterances were used to train the CT/ST classification model, model results would have improved with the use of more training data. Existing training data for the CT/ST classifier required sampling strategies to address the class imbalance. Labeling across classes, at times, was heavily context dependent, and considered information outside

the range of the sample utterance such as conversational tone or previous volleys. To improve machine learning model performance in the future, text classification performance leveraging embeddings generated from more recent and larger language models such as GPT-3, GPT-J, T5, and Megatron-Turing NLG should be evaluated.

Topic Classifier:

While data scraped from social media websites such as Reddit do have the potential to provide a large body of topic training data, our findings suggest that such data may be too noisy to be used for this purpose. For example, training data extracted from the "r/stopsmoking" subreddit with a post title that includes the topic keyword "cravings" may provide valuable training data for topic classification; however, other commenters may discuss topics such as nicotine replacement therapy (NRT), which will confound the model's ability to distinguish between the "cravings" topic and the "NRT" topic. Our team spent a significant amount of time cleaning the Reddit topic dataset by manually removing confounding topics. Topic data extracted from Reddit may also differ in form and style from topics featured in real MI-style conversations for smoking cessation.

Utterance topic classification was found to be analogous to text topic classification but had key distinctions which, if addressed, could improve real-world performance across models: 1) Utterances received through a conversational agent tend to be shorter than comments or web scraped content and transcribed speech [47], and 2) Large MI-specific transcribed datasets or datasets of in-person therapeutic delivery would provide further insight into relevant domains of conversation [47].

QA and MISC Classifiers:

Despite the promising breadth of coverage, the QA transformer-based classifier was excluded from real-world implementation due to out-of-distribution high confidence predictions that would often interrupt conversational flow and negatively impact engagement. We found greater success by providing QA results through keyword and keyphrase string-matching. A lack of annotated MISC transcripts and dialogue led to insufficient samples across the ~38 MISC coding classes, and ultimately insufficient model performance and exclusion from the trial. While MISC coding was considered during conversational graph construction and is leveraged by the information retrieval system, automated decision making would be improved with a higher quality MISC coding classifier.

6.2 Limitations

A few limitations of our paper should be noted. First, feasibility and acceptability data were obtained from a small number of participants (N=34) in an ongoing trial. Participant engagement and satisfaction with the TAMI chatbot may vary greatly in a larger sample. Second, model performance was evaluated using holdout test sets from MI and Reddit training data. Performance of the models during the course of the trial may vary if the characteristics of user utterances differ from training utterances. Following completion of the trial, all user conversation transcripts will be coded by MI experts for change talk, sustain talk, follow neutral, and topic. These expert codes will then be compared with chatbot model classifications for accuracy and F1 score. And, lastly, MI fidelity, a measure of adherence to MI, is currently unknown but will similarly be assessed by MI experts following trial completion. Assessment of MI comprehensiveness [25], while useful, should also be coupled with an assessment of MI fidelity in order to fully report MI quality and thus the chatbot's clinical performance.

6.3 Summary and Next Steps

Machine learning and NLP models, in combination with an information retrieval approach, can be successfully used to deliver motivational interviewing and guide patients to evidence-based interventions for smoking cessation. TAMI is currently being further evaluated in a pilot RCT to determine whether its use can facilitate smoking cessation above and beyond current available smoking cessation resources (e.g., quit lines). Undoubtedly, opportunities for iterative improvements in chatbot functions and flow will emerge and will need to be developed and evaluated.

We view this formative work developing TAMI as the first step into a rich area of digital health research. There are many areas of future interest including how TAMI may facilitate user engagement, better evoke and resolve personal ambivalence, and eventually develop individualized strategies to change one's behaviors. For example, it remains unknown how to best facilitate emotional attachment to TAMI in order to promote engagement and repeated use. TAMI's clinical effectiveness could be enhanced by pairing its use with a live telehealth quit coach or peer counselor. In this way, TAMI may initiate a conversation about change that can be enhanced and improved through contact with a live provider who can also facilitate connections with cessation treatment programs and resources (e.g., getting free NRT or pharmacotherapy). Moreover, if TAMI demonstrates efficacy in smoking cessation, TAMI's architectural framework and "training" methods are readily adaptable to other substance use behaviors (e. g., alcohol use) as well as numerous other "target behaviors" in which individuals may experience ambivalence (e.g., medication adherence, exercise).

Acknowledgements

This project was funded by the California Tobacco-Related Disease Research Program (Grant #T30IP0972) award to Dr. Satterfield. Dr. Borsari's contribution is the result of work supported with resources and the use of facilities at the San Francisco VA Medical Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Department of Veterans Affairs or the United States Government.

References

[1] World Health Organization. Tobacco. Posted July 26, 2021. Accessed August 5, 2021. https://www.who.int/en/news-room/fact-sheets/detail/tobacco
[2] Cornelius ME,Wang TW, Jamal A, Loretan CG, Neff LJ. Tobacco product use among adults—United States, 2019. MMWR Morb Mortal Wkly Rep. 2020; 69(46):1736-1742. doi:10.15585/mmwr.mm6946a4
[3] Rigotti NA, Kruse GR, Livingstone-Banks J, Hartmann-Boyce J. Treatment of Tobacco Smoking: A Review. JAMA. 2022;327(6):566-577. doi:10.1001/jama.2022.0395
[4] Babb S, Malarcher A, Schauer G, Asman K, Jamal A. Quitting smoking among adults—United States, 2000-2015. MMWR Morb Mortal Wkly Rep. 2017;65(52):1457-1464. doi:10.15585/mmwr.mm6552a1
[5] Smoking Cessation: A Report of the Surgeon General—Executive Summary. US Dept of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health; 2020.
[6] Miller, William R., and Stephen Rollnick. Motivational Interviewing: Helping People Change. Guilford Publ., 2012.
[7] West R, Raw M, McNeill A, et al. Health-care interventions to promote and assist tobacco cessation: a review of efficacy, effectiveness and affordability for use in national guideline development. Addiction. 2015;110(9):1388-1403. doi:10.1111/add.12998
[8] Lai DTC, Cahill K, Qin Y, Tang JL. Motivational interviewing for smoking cessation. Cochrane Database of Systematic Reviews 2010, Issue 1. Art. No.: CD006936. DOI: 10.1002/14651858.CD006936.pub2.
[9] Almusharraf F, Rose J, Selby P. Engaging Unmotivated Smokers to Move Toward Quitting: Design of Motivational Interviewing-Based Chatbot Through Iterative Interactions. J Med Internet Res. 2020;22(11):e20251. Published 2020 Nov 3. doi:10.2196/20251
[10] Abroms LC, Lee Westmaas J, Bontemps-Jones J, Ramani R, Mellerson J. A content analysis of popular smartphone apps for smoking cessation. Am J Prev Med 2013; 45(6):732-6.
[11] Cobb NK, Jacobs MA, Saul J, Wileyto EP, Graham AL. Diffusion of an evidence-based smoking cessation intervention through Facebook: a randomised controlled trial study protocol. BMJ Open 2014; 4(1):e004089.
[12] Sadasivam RS, Delaughter K, Crenshaw K, et al. Development of an interactive, Web-delivered system to increase provider-patient engagement in smoking cessation. J Med Internet Res 2011; 13(4):e87.
[13] Carroll, K.M., Ball, S.A., Martino, S. Nich, C., Gordon, M.A., Portnoy, G.A. & Rounsaville, B.J. (2008). Computer-assisted delivery of cognitive behavioral therapy for addiction: A randomized trial of CBT4CBT. Am J Psychiatry. 2008 Jul;165(7):881-8. doi: 10.1176/appi.ajp.2008.07111835.
[14] Carroll, K.M, Kiluk, B.D., Nich, C., Gordon, M.A., Portnoy, G.A., Martino, D.R., & Ball, S.A. (2014). Computer-Assisted Delivery of Cognitive-Behavioral Therapy: Efficacy and durability of CBT4CBT among cocaine-dependent individuals maintained on methadone. Am J Psychiatry, 171, 436-444.
[15] Campbell ANC, Nunez EV, Matthews AG, et al. Internet-Delivered Treatment for Substance Abuse: A Multisite Randomized Controlled Trial. Am J Psychiatry 2014; 171:683–690.
[16] Dulin PL, Gonzalez VM, Campbell K. Results of a pilot test of a self-administered smartphone-based treatment system for alcohol use disorders: usability and early outcomes. Subst Abus. 2014;35(2):168-75. doi: 10.1080/08897077.2013.821437.
[17] Norberg MM, Rooke SE, Albertella L, Copeland J, Kavanagh DJ, et al. (2013) The First mHealth App for Managing Cannabis Use: Gauging its Potential Helpfulness. Journal of Addictive Behaviors, Therapy and Rehabilitation, DOI:10.4172/2324-9005.S1-001.
[18] Gonzalez VM, Dulin,PL. Comparison of a smartphone app for alcohol use disorders with an Internet-based intervention plus bibliotherapy: A pilot study. J Consult Clin Psychol. 2015 Apr;83(2):335-45. doi: 10.1037/a0038620.
[19] Gustafson DH, McTavish FM, Chih MY, Atwood AK, Johnson RA, Boyle MG, Levy MS, Driscoll H, Chisholm SM, Dillenburg L, Isham A, Shah D. A smartphone application to support recovery from alcoholism: a randomized clinical trial. JAMA Psychiatry. 2014;71(5):566-72.
[20] Milne-Ives M, de Cock C, Lim E, et al. The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review. J Med Internet Res. 2020;22(10):e20346. Published 2020 Oct 22. doi:10.2196/20346
[21] Tudor Car L, Dhinagaran DA, Kyaw BM, et al. Conversational Agents in Health Care: Scoping Review and Conceptual Analysis. J Med Internet Res. 2020;22(8):e17158. Published 2020 Aug 7. doi:10.2196/17158
[22] Schachner T, Keller R, V Wangenheim F. Artificial Intelligence-Based Conversational Agents for Chronic Conditions: Systematic Literature Review. J Med Internet Res. 2020;22(9):e20701. Published 2020 Sep 14. doi:10.2196/20701
[23] Borsari, B., Apodaca, T. R., Jackson, K. M., Mastroleo, N. R., Magill, M., Barnett, N. P., & Carey, K. B. (2015). In-session processes of brief motivational interventions in two trials with mandated college students. Journal of consulting and clinical psychology, 83(1), 56-67.
[24] Moyers TB, Martin T, Manuel JK, Miller WR, & Ernst D. Revised Global Scales: Motivational Interviewing Treatment Integrity 3.1.1 (MITI 3.1.1) University of New Mexico Center on Alcoholism, Substance Abuse and Addictions (CASAA).
[25] Shingleton RM, Palfai TP. Technology-delivered adaptations of motivational interviewing for health-related behaviors: A systematic review of the current research. Patient education and counseling. 2016;99(1):17-35. [26] Layton, J. F., Delacruz, J., Saiyed, A., Tsvetovat, M., Borsari, B., & Satterfield, J. (2021). Technology-Assisted Motivational Interviewing: Building a Smart Chatbot for Smoking Cessation. Annals of Behavioral Medicine (Vol. 55, pp. S440-S440).

[27] Layton, J. F., Saiyed, A., Delacruz, J., Borsari, B., Tsvetovat, M., Haddadi, Y., Jamison, T., Luu, E., Truong, T., Satterfield, J. (January 2021). Development and Validation of Natural Language Processing Models for Technology-Assisted Motivational Interviewing. Invited talk given at the 2021 SRNT Virtual Meeting; San Francisco, CA.
https://cdn.ymaws.com/www.srnt.org/resource/resmgr/conferences/2021_annual_meeting/srnt2021_program.pdf
[28] Satterfield, J., Layton, J. F., Delacruz, J., Saiyed, A., Tsvetovat, M., & Borsari, B. (June 2020). Technology-Assisted Motivational Interviewing: Leveraging Machine Learning to Promote Readiness for Smoking Cessation and Tailored Referrals. Poster accepted at Tobacco Control, Research, and Education: Joining Forces to Address New Challenges; Palm Desert, CA. [Mtg canceled due to COVID-19]
[29] Delacruz, J., Saiyed, A., Layton, J. F., Borsari, B., Satterfield, J., Kanzaveli, T., & Tsvetovat, M., (June 2020). TAMI-BERT Convolutional Neural Networks for Automated Health Behavior & Change Talk Evaluation in Motivational Interviewing for Tobacco Cessation. Poster accepted at Tobacco Control, Research, and Education: Joining Forces to Address New Challenges; Palm Desert, CA. [Mtg canceled due to COVID-19]
[30] Saiyed, A., Tsvetovat, M., Kanzaveli, T., Layton, J., Delacruz, J., Borsari, B., Satterfield, J. (October 2021). Change or Sustain Talk Classification with Language Models and Graph Neural Networks. Poster presented at 10th Annual Health Informatics and Data Science Symposium at Georgetown University; Arlington, VA.
[31] Fiore MC, Jaén CR, Baker TB, Bailey WC, Benowitz NL, Curry SJ, et al. Treating tobacco use and dependence: 2008 update. Washington, DC: US Department of Health and Human Services; 2008. Available from:
www.ahrq.gov/clinic/tobacco/treating_tobacco_use08.pdfAccessed 2022 May 31
[32] Bocklisch, Tom & Faulker, Joey & Pawlowski, Nick & Nichol, Alan. (2017). Rasa: Open Source Language Understanding and Dialogue Management.
[33] Minaee, Shervin, et al. “Deep Learning--Based Text Classification.” ACM Computing Surveys, vol. 54, no. 3, 2021, pp. 1–40., doi:10.1145/3439726.
[34] Wu, Lingfei, et al. “Graph Neural Networks for Natural Language Processing: A Survey.” ArXiv, 2021, pp. 207–221., doi:10.1017/9781108924184.015.
[35] Magill M, Hallgren KA. Mechanisms of Behavior Change in Motivational Interviewing: Do We Understand How MI Works? Current Opinion in Psychology. 2019;30:1-5.
[36] Apodaca TR, Jackson KM, Borsari B, et al. Which individual therapist behaviors elicit client change talk and sustain talk in motivational interviewing? Journal of Substance Abuse Treatment. 2016;61:60-65.
[37] Miller WR, Rollnick S. Motivational Interviewing: Helping People Change. 3rd ed. New York: Guiford Press; 2013.
[38] Gaume J, Bertholet N, Faouzi M, Gmel G, Daeppen JB. Counselor motivational interviewing skills and young adult change talk articulation during brief motivational interventions. Journal of Substance Abuse Treatment. 2010;39(3):272-281.
[39] Gaume J, Gmel G, Faouzi M, Daeppen JB. Counsellor behaviours and patient language during brief motivational interventions: A sequential analysis of speech. Addiction. 2008;103(11):1793-1800.
[40] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “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), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics
[41] Liu, Yinhan, 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 abs/1907.11692 (2019): n. pag.
[42] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: generalized autoregressive pretraining for language understanding. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, Article 517, 5753–5763.
[43] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, et al.. 2020. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.
[44] Rajapakse TC. Simple Transformers. Published 2019. https://github.com/ThilinaRajapakse/simpletransformers
[45] Lin, Yuxiao, et al. “BERTGCN: Transductive Text Classification by Combining GNN and Bert.” Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, doi:10.18653/v1/2021.findings-acl.126.
[46] Yao, Liang, et al. “Graph Convolutional Networks for Text Classification.” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 7370–7377., doi:10.1609/aaai.v33i01.33017370.
[47] Ahmadvand, Ali et al. “ConCET: Entity-Aware Topic Classification for Open-Domain Conversational Agents.” Proceedings of the 28th ACM International Conference on Information and Knowledge Management (2019): n. pag.
[48] De Jonge, J., Schippers, G., & Schaap, C. (2005). The Motivational Interviewing Skill Code: Reliability and a Critical Appraisal. Behavioural and Cognitive Psychotherapy, 33(3), 285-298. doi:10.1017/S1352465804001948
[49] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
[50] McInnes et al., (2018). UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software, 3(29), 861, https://doi.org/10.21105/joss.00861