**Response to the reviewer’s comments**

Dear Reviewer,

We are grateful for your comments and are hereby responding to you:

**Reviewer Comments**(Please note that your submission was reviewed by at least three reviewers and an SPC member. Numbers 1, 2, 3., etc. below represent different reviewers):1: The paper presents a convoluted deep learning based approach for automated segmentation of e-Coaching MI sessions. The methods used are well-explained. However, the paper lacks organization and not coherent. The introduction could focus on the big picture of MI segmentation and how the proposed methods can address the current limitations. The paper does not mention why the genY eating dataset was chosen and how the data is structured. Given that annotation is laborious and time-consuming, information on how the data was annotated and a snippet of data would better explain the underlying dataset. For the same reason, an investigation of the number of instances of label data required for accurately predicting the new/same segment could improve the proposed method.

***Response:*** *In the revised manuscript, the introduction is re-written to focus on the big picture of MI segmentation and how the proposed methods can address the current limitations.* *The revised manuscript eliminates the confusion about the fact that why the genY eating dataset was chosen and how the data is structured because this dataset was a recent email-based MI intervention that relies on personalized e-Coaching to encourage increased fruit and vegetable intake among young adults, aged 21–30. We carefully revised our manuscript to clearly describe our dataset in the Data collection section. We also included a summary table to clearly explain the experimental data. The revised manuscript also explains that how the data was segmented and annotated with 115 distinct MY-SCOPE behavior codes (Figure 1 and Table 1).*

F1-measure, precision and recall are good metrics for evaluating imbalanced datasets. However, AUC is agnostic to class imbalance and quite similar to accuracy. For these reasons, we see a very high AUC value (all over 0.9) for all classifiers and features combination. AUPR should be used instead (both as precision-recall curves as well as AUPR values) as it takes the class imbalance into account and helps to identify the power of the classifier in correctly labelling the minority class. The paper does not mention how many iterations for each fold and how many 5-fold sets for cross-validation were conducted.

***Response:*** *The revised manuscript utilized AUPR values instead of AUC. Unfortunately, precision-recall curves are not included in the final manuscript due to the page limitation. The revised manuscript addressed the 5-fold cross-validation issue in the Evaluation metrics section on page 5. The results are reported based on 5 fold cross-validation and weighted macro-averaging over the folds, in which each fold was used as a testing set and remaining 4 folds was utilized as a training set. In page 5, the revised manuscript also mentioned that the early-stopping strategy is used with 50 epochs for the training of our models.*

2: Good idea to use deep learning for text segmentation. You mentioned that a limitation of this study is that data is collected from a single medical institute but formatting, style and email segment can be different in other settings. It could be better to describe how different these data are.

***Response:*** *We carefully revised our manuscript to clearly describe our dataset in the Data collection section. We also added a summary table (Table 1) to clearly explain the structure of the data with Figure 1.*

3: Summary: Applying prior algorithms using NLP features to segment e-mail based behavioral interventions.1) The classification problem of a new segment or not in the email is identifying a rare event (a new segment start), amongst many existing segment words. The authors classify on each individual word and punctuation. It appears that the algorithm could potentially be splitting up segments by sentences where it could classify many correctly and only a few wrong (multi-sentence segments). A further discussion on the number of multi-sentences and how easily those are detected would speak to this. Further, if it is the case that it tends to identify single sentence segments well, then maybe consider splitting the text into sentences and classify on each sentence instead of words. Also a minor detail, how does the algorithm start automatically: with a 0 at the first word?

***Response:*** *Discourse segments in e-Coaching emails do not necessarily correspond to sentences or paragraphs. One sentence can be divided into multiple segments corresponding to multiple MI behaviors. On the other hand, a segment or MI behavior may comprise several sentences. We appreciate your idea that you suggest splitting emails into sentences first. Actually, we came up with the same idea and thought that when we split sentences into smaller segments and merge sentences into large segments the algorithm performs better. But multiple segments in one sentence is not a rare case, especially in the compound and complex sentences. Due to the time limitation, we can not implement your suggested method right now but we will try to implement this approach in our future studies. The algorithm does not always annotate the first word as 0 or same segment class. It depends, for example, in the following email first word (fantastic) annotated as 1 or new segment class because it represents an affirmative behavior in MY-SCOPE:*

*Fantastic*

*In that survey you choose a lower number for your confidence that you can eat more fruit...*

2) In the comparison of algorithms (table 1 and 2), it appears that MLP has the best precision. A discussion on why this might be the case would provide more context for the algorithms chosen. Further, table 3 with the AUC values, for BRNN and CRNN, the POS tags lowered the AUC. A discussion on this would provide more context as well.

***Response:*** *Now, the revised manuscript utilized area under the precision-recall curve (AUPR) instead of AUC as a request of one of the reviewers. We believe that AUPR takes the class imbalance into account and helps to identify the power of the classifier in correctly labeling the minority class. We believe that the BRNN and CRNN performed poorly with POS features because POS tagging is a supervised learning solution that uses features like the previous and next word. Since we already considered neighbor words by utilizing bi-directional RNN, it failed to achieve good results with redundant information. We observed that MLP achieved the highest precision which may be related to the fact that MLP poorly learned “new segment” and misclassified new segment words to same segments in 30%-40% of the time.*3) In terms of evaluation, the standard method is to compare the IAA to the F1-Measure. A discussion on the IAA and how it compares would provide an idea of how hard the problem is. The F1-measure is already quite high as well.

***Response:*** *We appreciate your suggestion that IAA will be a good metric to estimate how hard the segmentation problem is. Actually, this study is the first effort to evaluate the automatic segmentation of e-coaching emails. Each e-coaching session was segmented and annotated by one expert from research staff and student assistants in the Department of Family Medicine and Public Health Sciences at Wayne State University School of Medicine. Therefore, at this moment, we don’t have any IAA results to report here. We also believe that F1-Measure is a well-accepted performance metric in machine learning community for classification problems, especially in imbalanced datasets.*

4: This paper reports a comparison study to evaluate the empirical effectiveness of deep learning architectures in addressing the problem of discourse segmentation in the context of e-mail based behavioral interventions. The main contribution of this paper makes to the AMIA community is insights on discovering unstructured text data in the emails for communication and intervention, which can enrich the available data for further discoveries. This contribution is applicable in real-world as there exists huge amount of unstructure data besides EHR.

Suitability for AMIA: The paper is well written with key points made easy to find through clear pipeline illustration. The work is a good fit for the AMIA audience particularly for attendees interested in deep learning, NLP real-world application.Context for work: There are multiple points that the authors need to address, the data needs to be described more clearly, a summary table would be great. The font for the captions of figures need modification, Fig.3 pipeline is not clear enough.

***Response:****We carefully revised our manuscript to clearly describe our dataset in the Data collection section. We modified the figure 3 in the revised manuscript. We also added a summary table (Table 1) to clearly explain the structure of the data with Figure 1. In addition to the above modifications, we also carefully revised the Convolutional Recurrent Neural Network section to clarify the pipeline depicted in Figure 3.*Suitability of methods: The methods are clear, thoroughly described, and well-matched to answer the stated research questions, especially it is robust to unbalanced data. The detailed implementation of the methodology is good for reproduce. Considering deep learning is mostly used as a black-box, the interpretation of it on the medical application is not that clear.Analysis and importance of results: Although the analysis involved complex steps, readers without a technical background should find this paper quite accessible because of the authors’ clarity. The author successfully convey that CRNN is the better one for the problem they are investigating, however, as the author mentioned, the generalization of the remains unclear and the fields which require the technique are narrow. In all, this can be a good work considering the clear pipeline and clear statement of the evaluation.