Machine Learning Methods for Discourse Segmentation of Communications in E-Mail Based Behavioral Interventions

Mehedi Hasan, BS[1*], Alexander Kotov, PhD[1*], Sylvie Naar, PhD[2], Gwen L. Alexander, PhD[3], April Idalski Carcone, PhD[4]
[1]Department of Computer Science, Wayne State University, Detroit, Michigan
[2]Center for Translational Behavioral Research, Department of Behavioral Sciences and Social Medicine, Florida State University, Tallahassee, Florida
[3]Department of Public Health Sciences, Henry Ford Health System, Detroit, Michigan
[4]Department of Family Medicine and Public Health Sciences, School of Medicine, Wayne State University, Detroit, Michigan

Abstract *Communication science approaches to developing effective behavior interventions, such as motivational interviewing (MI), are limited by traditionally manual qualitative coding of communication exchanges, which is a very resource-intensive and time-consuming process. This study focuses on the analysis of e-Coaching sessions, behavior interventions that are delivered via email and grounded in the principles of MI. A critical step towards automated annotation of e-Coaching communication exchanges is segmentation of emails into textual fragments that correspond to MI behaviors. In this work, we formulate this task as a classification problem and propose lexical, punctuation and topic features to address it. We experimented both with traditional machine learning methods, such as Support Vector Machine (SVM), Naive Bayes and K-Nearest Neighbor (KNN) classifiers and recurrent neural networks (RNNs). Results indicate that SVM outperformed KNN and RNNs achieving 0.990 macro F1-score overall, and 0.848 and 0.995 macro F1-score for detecting "new segment" and "same segment" classes, respectively.*

Introduction

The emergence of e-Health technologies opened up new ways to deliver a variety of behavioral interventions to any demographic group of patients in any geographical location. Motivational interviewing (MI), an evidence-based communication technique to increase intrinsic motivation and self-efficacy for behavior change[1–3], is one type of these interventions. MI sessions are generally aimed at eliciting "change talk", or statements of intrinsic motivation about patients' own desire, ability, reasons, need for and commitment to behavior change, which have been established by previous research[4] as a reliable mediator of health behavior change. However, communication science approaches to understanding the efficacy of MI are inherently limited by traditional qualitative coding methods.

Qualitative coding of motivational interviews with pre-defined codes has been traditionally performed manually by trained annotators, which is a tedious and resource-intensive process that involves several iterations of reading, comprehension and interpretation of interview transcripts. Rapidly developing computational technologies, specifically, machine learning methods, offer a unique opportunity to accelerate this process. In particular, machine learning methods have been successfully applied to a variety of analytical tasks involving textual data, such as classification[5] and sentiment analysis[6]. In our previous work, we examined the utility of machine learning methods for automated annotation[7,8] and analysis[9] of in-person MI sessions. Specifically, we demonstrated that machine learning methods can be utilized for annotation of MI transcripts according to a simple communication code scheme with the accuracy comparable to human coders[7]. Experimental data utilized in these studies, however, were prepared by transcribing audio conversations, which were clearly segmented into utterances by a counselor, a patient, and, in some cases, a caregiver.

In this study, we focus on the analysis of e-Coaching sessions, behavior interventions that are delivered via email and grounded in the principles of motivational interviewing. Specifically, the e-Coaches involved in this study used emails to communicate motivation-enhancing messages that encourage healthy eating among GenY adolescents. e-Coaching data is comprised of email responses, which are free-text documents, unlike more traditional dyadic clinical interviews that are naturally segmented into utterances due to their conversational nature.

The unstructured nature of e-Coaching exchanges poses a unique set of challenges for their qualitative analysis. A significant barrier to fully automating the behavior coding process of e-Coaching emails is their segmentation into tex-

*Authors provided equal contribution.

tual fragments that correspond to distinct communication behaviors. Automating this task is a unique and challenging problem due to the following major reasons:

1. Emails are unstructured text that contains informal information exchange in a non-traditional format.
2. Discourse segments in e-Coaching emails do not necessarily correspond to sentences or collection of sentences. One sentence can be segmented into multiple MI behavior fragments. On the other hand, an MI behavior may comprise several sentences.

Figure 1 illustrates a segmentation of an e-Coaching email exchange, in which the first sentence is segmented into 2 MI behavior fragments, while the fourth and fifth segments correspond to one and three sentences, respectively. Segmentation of e-Coaching emails corresponds to a special type of discourse analysis[10] aimed at better understanding the effective e-Coaching communication strategies and revealing the unique socio-psychological characteristics of a patient.

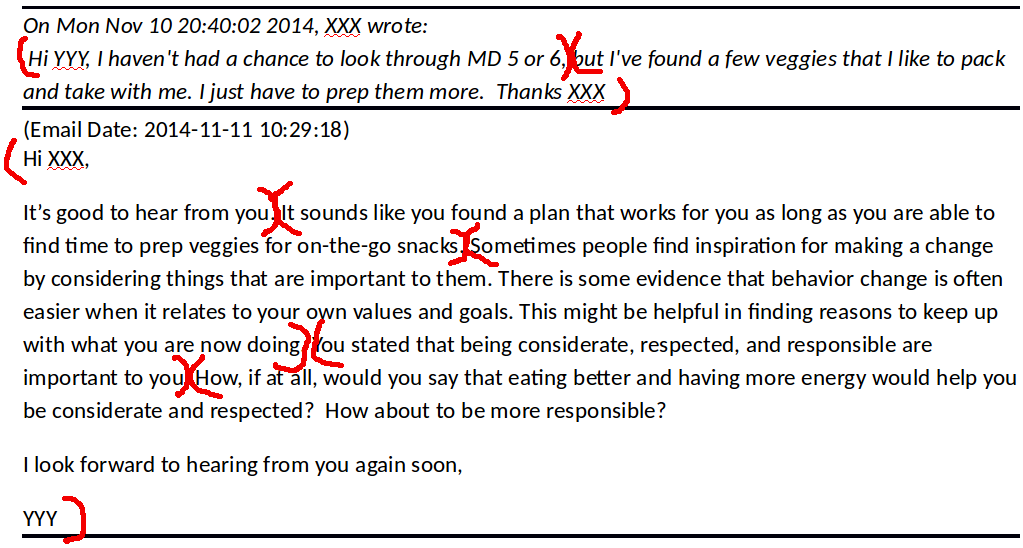

Figure 1: Example of e-Coaching emails segmented into fragments that correspond to MI behaviors of an e-Coach and a patient.

The goal of this research study is to assess the applicability of machine learning methods for automated segmentation of e-Coaching emails into textual fragments corresponding to individual behaviors, which is the first step of the coding process of e-Coaching communications. In particular, we introduced lexical, topic and punctuation features and experimented with both traditional supervised machine learning methods, such as Support Vector Machine (SVM), Naive Bayes (NB) and K-Nearest Neighbor (KNN) classifiers, and deep learning methods, such as Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU), to find the best performing method and feature combination.

Relevant previous work in the biomedical domain primarily focused on segmentation of text into sections and headers[11–14] or sentence boundary detection[15–17]. Apostolova et al.[11] applied SVM along with word-vector cosine similarity metric combined with several heuristics to segment clinical reports into sections, such as demographics, history, procedure, finding and impression. After identification of each line in the document, Tepper et al.[13] trained Maximum Entropy models for section classification. Denny et al.[12] proposed a SecTag algorithm, which combined natural language processing techniques, terminology-based rules, and Naive Bayes classifier to identify the sections and headers that achieved 99% recall with 95.6% precision. On the other hand, SVM based on prosodic and part of speech features[16] and recurrent convolutional neural networks using word embeddings[15] were utilized for detecting sentence boundaries. Segmentation of e-Coaching emails is different from traditional shallow discourse analysis of conversa-

tions[18] in that the focus is on segmentation, rather than on determining the types of transitions between the utterances or assigning utterances to speakers.

Recently, an online clinical intervention called MENU GenY[19] (Making Effective Nutrition Choices for Generation Y) was proposed and evaluated. MENU GenY is a technology-based public health intervention that relies on personalized e-coaching to encourage increased fruit and vegetable intake among young adults, aged 21-30. The goal of MENU GenY was to develop a better coding dictionary among GenY to improve eating habits. However, segmentation of clinical conversation in the context of electronically delivered interventions, in particular, segmentation of clinical interaction text into groups of MI behaviors, is still performed manually, which slows down qualitative analysis of these interventions. This study introduces a novel computational approach to address this problem and the authors are unaware of any other work that focused on the same problem.

Methods

Data collection

The experimental dataset for this work was constructed from 49 e-coaching sessions, which include a total of 3,138 segmented and annotated MI behaviors. Each session represents an MI intervention delivered via email. During pre-processing, we removed all non-ASCII characters and applied stemming to normalize morphological variants of related concepts, such as "eating", "eats", and "eat". We formulate the segmentation task as a binary classification problem, as illustrated in Figure 2 and experimented both with traditional machine learning methods, such as Support Vector Machine (SVM), Naive Bayes (NB) and K-Nearest Neighbor (KNN) classifiers and recurrent neural networks (RNNs), such as Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU). For NB, SVM and KNN models, e-Coaching email exchanges are partitioned into adjacent word pairs. Each pair is classified into either "new segment" or "same segment" class. The gold standard is manually segmented where adjacent word pairs are classified into "new segment" class. If all adjacent word pairs of a block of text are classified into "same segment" class, the entire block is treated as one textual fragment corresponding to a single MI behavior. In total, we obtained 95,421 word pairs, which include 3,138 "new segment" and 92,283 "same segment" instances. For RNNs, a block of text was taken as input sequence, such that one-hot encodings of each word or punctuation marks in a block were used as input into an RNN and binary labels (1 or 0) corresponding to "new segment" and "same segment" classification decision were considered as the output of RNN at each step. In the gold standard, words within the same segment were assigned the label of 0 and the last word or punctuation mark of a segment were assigned the label of 1.

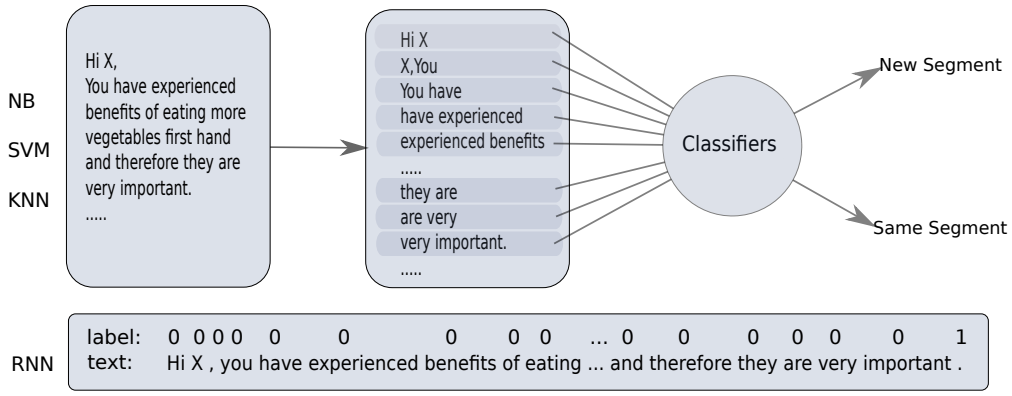

Figure 2: Transformation of text segmentation task into text classification task.

Features

We utilized three types of features in conjunction with SVM, NB and KNN methods: lexical, punctuation and topic features. Lexical features represent each word in a pair as a binary vector, in which 1 corresponds to the word in

question and 0 to all other words. Since topic models[8,20,21] have been shown to be effective semantic abstractions of individual words, we derived topic features from Labeled LDA[22], a topic model for annotated corpora. We considered each textual segment in the training set as a document and the MI code assigned to the segment as a label. First, we derived a label-specific multinomial $p(w|l)$ for each label l from the word distributions $p(w|l,z)$ for topic z specific to label l by marginalizing over topics:

$$p(w|l) = \sum_{z=1}^{K} p(w|z,l) \tag{1}$$

where K is the optimal number of topics experimentally determined to minimize perplexity. After that, distributions over labels, $p(l|w)$, for each word were obtained from label-specific multinomials $p(w|l)$ using Bayesian inversion:

$$p(l|w) = \frac{p(w|l)p(l)}{p(w)} \tag{2}$$

where $p(l)$ is the prior probabilities of label l in the training set and $p(w)$ is the prior probability of word w in the language model estimated from the training set using maximum likelihood. We use the distribution $p(l|w)$ as the topic feature vector, which shows how indicative each word w is for a particular behavior code l. Topic features allow to abstract away from individual words and detect segment boundaries by indirectly capturing transitions between behavior codes when only observing a pair of words that are specific to different behavior codes. To find the optimal topic feature, we also derived the same topic feature $p(l|w)$ by using two recently proposed probabilistic generative latent variable models: Discriminative Labeled Latent Dirichlet Allocation (DL-LDA) and Latent Class Allocation (LCA)[8]. Similar as L-LDA, DL-LDA-based topic feature is extracted by first deriving a label-specific language model $p(w|l)$ for each label l from the word distributions $p(w|l,z)$ for topic z specific to label l by using the Eq. 1 and 2. On the other hand, LCA directly provides label-specific multinomials $p(w|l)$, which can be used to derive topic feature by using Eq. 2. Punctuation marks, which correspond to one of the symbols {'.', ',', '!', '?', ':', ';', '-'} between a pair of words, are also employed as a feature, since punctuation marks designate the boundary of a sentence, clause, and phrase and often also correspond to segment boundary.

Classifiers

We experimented both with traditional classifiers, including Naive Bayes (NB)[23], Support Vector Machine (SVM)[24], K-Nearest Neighbor (KNN)[25] and two variants of Recurrent Neural Networks (RNNs)[26]: Long Short Term Memory (LSTM)[27] and Gated Recurrent Unit (GRU)[28].

Naive Bayes: this model estimates the prior probability of classes along with conditional probabilities of each feature given the class. Then, the posterior probability is computed to predict the class for each sample by applying the Bayes rule with the assumption that features are conditionally independent. We used Multinomial Naive Bayes in experiments.

Support Vector Machine: state-of-the-art supervised classification method proven to be effective for text categorization[29] for its ability to cope with high dimensional input feature space. SVM finds the hyperplane that maximizes the separation between the closest "new segment" and "same segment" training examples. We used L1-regularized linear SVM with squared error minimization in experiments.

K-Nearest Neighbor: this method considers each training sample as a point in the input feature space. For a new test sample, Euclidean distance is calculated to find the k-nearest classified neighbors from the training set. Finally, the test sample is classified into a majority class of its k-nearest neighbors. We experimentally determined that best performance was achieved when the number of neighbors is 3.

Recurrent Neural Networks: RNN is a neural network architecture designed to capture sequential patterns. When we predict the "new segment" point using RNNs, particular combinations of words in a sequence will indicate the change of MI behavior. Long Short Term Memory (LSTM) networks[27] are a special type of RNN that are capable of

handling variable size input sequence and have an internal memory that can be reset. GRU[28] is a variant of LSTM mathematically represented by the following formula:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{3}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{4}$$

$$\tilde{h}_t = tanh(W_h x_t + r_t \odot U_h h_{t-1} + b_h) \tag{5}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \tag{6}$$

In Eq. 3-6, σ corresponds to sigmoid function and $\odot$ designates an element-wise product. The update gate z_t and reset gate r_t at time step t are computed by the Eq. (3) and (4), where W_z, W_r, W_h, U_z, U_r, U_h are the weight matrices and b_z, b_h and b_r are bias vectors. The activation h_t of the GRU at time t is a linear combination of previous activation h_{t-1} and the candidate activation $\tilde{h}_t$, which is represented by Eq. (6) and (5). Our GRU model includes one hidden layer, output layer, and input layer. We reset our model state after feeding each input sequence where input was given as one-hot encoding of word vector. Since LSTM and GRU use one-hot vector as input, results for these models are reported only when lexical and punctuation features are used. We experimentally determined that the best performance is achieved when the number of hidden units is 25, batch size is 1, and Adam[30] is used for optimization.

Evaluation metrics

We report standard metrics for experiments (precision, recall, and F1-measure) to evaluate the performance of binary classifiers[31]. However, accuracy is not reported as a performance metric because accuracy is highly sensitive to the prior class probabilities and does not fully describe the actual difficulty of the decision problem for an unbalanced dataset. The results are reported based on 5 folds cross-validation and weighted macro-averaging over the folds. We also estimate the area under the receiving operating characteristics (ROC) curve[32] (AUC) metric due to its effectiveness in measuring the quality of binary classifiers for imbalanced datasets[33].

Results

Our experimental results have two important directions. First, results are reported with respect to "new segment" and "same segment" classes as well as their weighted average in Table 1. Second, classification performance of different machine learning methods are outlined in Tables 2, 3, and 4 when topic features are extracted with different probabilistic generative latent variable models.

Table 1: Performance of NB, SVM, and KNN methods for identification of "new segment" and "same segment" classes as well as their weighted average. The highest value for each performance metric is highlighted in bold.

Method	**new segment**			**same segment**			**overall**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
lexical features only									
NB	0.594	0.662	0.626	0.988	0.985	0.987	0.975	0.974	0.975
SVM	0.762	**0.673**	0.715	**0.989**	0.993	0.991	0.981	0.982	0.982
KNN	**0.808**	0.663	**0.728**	**0.989**	**0.995**	**0.992**	**0.983**	**0.984**	**0.983**
lexical and puctuation features									
LSTM	**0.800**	0.646	0.714	0.993	**0.994**	**0.994**	**0.986**	0.983	0.984
GRU	**0.800**	**0.715**	**0.741**	**0.994**	**0.994**	**0.994**	**0.986**	**0.985**	**0.986**

As follows from Table 1, KNN outperforms all other models in terms of precision and F1-measure achieving 0.808 precision with 0.728 F1-measure for new segment detection. KNN also shows superior performance in all performance metrics for "same segment" class and wighted average over "new segment" and "same segment" classes. NB demonstrates the lowest performance among all models in terms of all performance metrics. On the other hand, SVM has the highest recall of 0.673 when only lexical features are used to identify "new segment". Results indicate that

performance of classifiers is remarkably higher for "same segment" class compared to "new segment" class, which is expected, since 96.71% instances belong to "same segment" categories. For example, KNN achieves 22.40%, 50.07%, and 36.26% higher precision, recall, and F1-measure, respectively, in same segment identification compared to new segment detection. When lexical features are used in combination with punctuation mark, GRU demonstrates the best performance in all cases showing 10.68% higher recall and 3.78% higher F1-measure than LSTM for new segment identification, 0.1% higher precision than LSTM for same segment identification and 0.2% higher recall and 0.2% higher F1-measure than LSTM for overall classification.

Table 2: Performance of NB, SVM, and KNN methods with different topic model-based features for identification of "new segment" class in e-Coaching emails, when all features are used together. The highest value for each performance metric is highlighted in bold.

Method	**L-LDA-based topic features**			**DL-LDA-based topic features**			**LCA-based topic features**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
NB	0.573	0.680	0.622	0.573	0.677	0.621	0.593	0.662	0.626
SVM	0.800	0.799	0.799	**0.848**	**0.849**	**0.848**	0.836	0.839	0.837
KNN	0.808	0.741	0.773	0.814	0.732	0.771	0.814	0.740	0.775

Table 2 summarizes the performance of NB, SVM, KNN, and RNN models for detecting "new segment" class. We observed that performance is remarkably better for "new segment" class compared to new segment detection, which is expected, since 96.71% instances belong to "same segment" class. GRU demonstrates the same result as KNN for precision and F1-measure, but not for recall. SVM shows moderate performance, achieving 0.981, 0.982, and 0.983 for precision, recall, and F1-measure, respectively, when lexical features are used.

Table 3: Performance of NB, SVM, and KNN methods with different topic model-based features for identification of "same segment" class in e-Coaching emails, when all features are used together. The highest value for each performance metric is highlighted in bold.

Method	**L-LDA-based topic features**			**DL-LDA-based topic features**			**LCA-based topic features**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
NB	0.989	0.983	0.986	0.989	0.983	0.986	0.988	0.985	0.987
SVM	0.993	0.993	0.993	**0.995**	**0.995**	**0.995**	**0.995**	0.994	0.994
KNN	0.991	0.994	0.993	0.991	0.994	0.993	0.991	0.994	0.993

As shown in Table 3, performance of classifiers is remarkably higher for "same segment" class compared to new segment class, which is expected, since 96.71% instances belong to "same segment" categories. KNN achieves 22.40%, 50.07%, and 36.26% higher precision, recall, and F1-measure, respectively, for lexical features; and 20.85%, 33.96%, and 27.47% higher precision, recall, and F1-measure, respectively, for combined features compared to new segment classification. In contrast to new segment detection, RNN demonstrates the highest performance among all models. The impact of punctuation and topic model-based features is consistent with "new segment" classification. Results indicate that F1-measure increases by 0%, and 0.1% for SVM and KNN models and decreases by 0.1% for NB.

Table 4: Performance of NB, SVM, and KNN methods with different topic model-based features for segmentation of e-Coaching emails as a weighted average over "same segment" and "new segment" classification results, when all features are used together. The highest value for each performance metric is highlighted in bold.

Method	**L-LDA-based topic features**			**DL-LDA-based topic features**			**LCA-based topic features**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
NB	0.975	0.973	0.974	0.975	0.973	0.974	0.975	0.974	0.975
SVM	0.987	0.987	0.987	**0.990**	**0.990**	**0.990**	0.989	0.989	0.989
KNN	0.985	0.986	0.985	0.985	0.986	0.985	0.985	0.986	0.986

Table 4 summarizes the weighted average results of the models for segmentation of e-Coaching emails. Overall, SVM has the best performance across all topic model-based features in terms of all performance metrics, achieves 0.990 precision, recall and F1-measure when DL-LDA-based topic features are used. Similar to results in Tables 1, 2 and 3, NB demonstrates the lowest performance among all methods while shows best performance with LCA-based topic features compared to other model-based topic features. Influence of the additional features is also consistent with the results in Tables 2 and 3. Precision increases by 0%, 1.43%, and 0.2%; recall increases by 0%, 1.32%, and 0.2%; and F1-measure increases by 0%, 1.32%, and 0.3% for NB, SVM, and KNN methods, respectively, when lexical feature is used in combination with punctuation and best model-based topic features.

Discussion

This study is the first effort to evaluate the automatic segmentation of e-Coaching emails. Experimental results indicate that SVM is the best model among all machine learning methods considered for this study. SVM achieved 0.990 F1-measure in overall, 0.848 and 0.995 F1-measures for detecting "new segment" and "same segment", respectively. The robust performance of SVM provides the evidence that machine learning models are capable to learn conceptual information from clinical exchange. It also indicates that topic features are more important than lexical features, even when deep learning methods are employed. Although the domain of this study was intentionally quite small, we believe that our study is not limited to the e-Coaching domain, and it can be successfully applied to other domains, in which discourse segmentation is an preliminary step for annotation.

Punctuation mark and topic model-based features made a significant improvement in performance of all machine learning methods. Among all topic model-based feature, DL-LDA-based topic feature provides the highest performance in "new segment" and "same segment" classification as well as weighted average over "new segment" and "same segment" classification results. In nearly all cases, ML methods perform better, when lexical features are used in combination with punctuation and topic model-based features. This results also indicate that segmentation performance might be improved by adding additional relevant features.

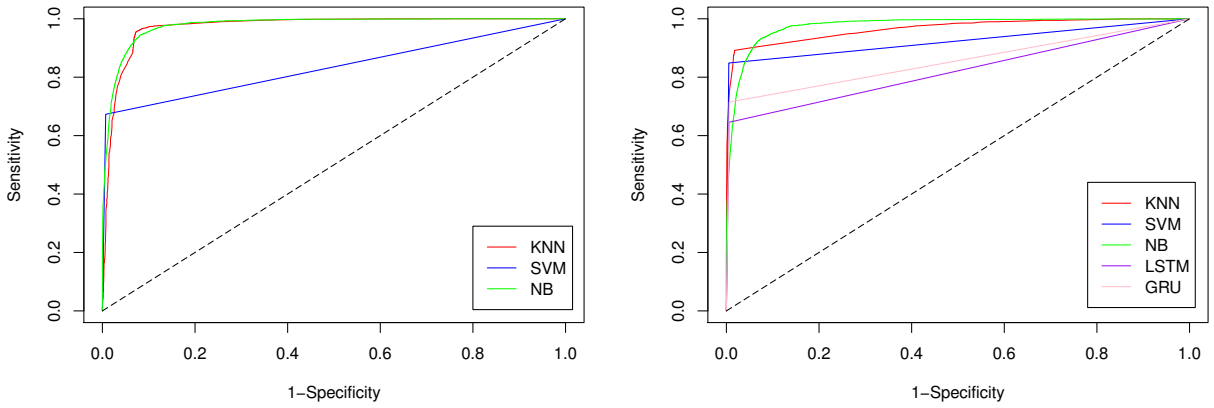

Figure 3: Receiver operating characteristic curves showing the performance of binary classifiers for the segmentation of e-Coaching text when lexical features (left) and combination of lexical, DL-LDA-based topic, and punctuation features (right) are used.

In this paper, results are reported by each class to avoid confusion regarding overall model performance due to severe class imbalance. In addition, standard metrics: precision, recall, and F1-measure were used to eliminate doubt about the model performance, since accuracy is misleading for imbalanced datasets. AUC is also utilized due to its effectiveness in measuring the quality of binary classifiers for imbalanced datasets[33], which are demonstrated by the ROC curves in Figure 3. Although NB provides lowest classification results, it shows the highest AUC values, achieving

0.978 AUC when only lexical features are used and 0.977 AUC when combination of lexical, punctuation and topic features are used. On the other hand, KNN and SVM exhibit 0.972 and 0.833 AUCs when only lexical features are used; and 0.966 and 0.922 AUCs when a combination of lexical, punctuation and topic model-based features are used. Finally, LSTM demonstrates lowest AUC values among all machine learning models, achieving 0.82 AUC, while GRU achieves 4.15% higher AUC, than LSTM. The conclusion drawn from the ROC curves also confirmed the robustness and superiority of KNN model for the segmentation of clinical exchange.

We observed moderate performance of RNN, in particular, LSTM and GRU for the text segmentation. We believe that RNN will perform better if a larger dataset is utilized. In this study, we employed 3,138 examples of boundary case which limit to achieve the best performance. We also observed that GRU performs better than LSTM, which was observed in previous studies[28].

Although punctuation mark plays an important role in segmentation boundary detection, and large numbers of errors were encountered by the false positive of boundary identification. For example, a text block "A1 A2 A3. B1 B2 B3. C1 C2 C3 C4." can be incorrectly segmented at position A3, B3, and C4 where a punctuation mark was encountered. Similarly, additional information is the common reason for the classified original segment into multiple segments. For instance, the above text block can also be incorrectly segmented at position B3 and C4 if third sentence (C) only provides a supportive information of MI code associated with first two sentences (A and B).

Our proposed approach is novel for the segmentation of e-Coaching emails, since previous studies mainly focused on the segmentation of text into sections, headers, and sentences. However, this study segmented clinical exchange into groups of MI behaviors which will significantly reduce the amount of resource and time required to segment clinical exchange manually. Furthermore, these segmentation models could be integrated with auto coding classifiers to build a software pipeline for automated annotation of exchanges.

The limitation of this study is that e-Coaching data is collected from a single medical institute; formatting, style, and email segment can be different in other settings. Therefore, there is a need to replicate the experiments with different data sets. As our future work, we plan to evaluate our approach on other datasets for discourse analysis.

Conclusion

Segmentation of e-Coaching emails is an integral part of developing e-Coaching interventions. Although several studies have focused on clinical interventions, they are limited by the qualitative coding of clinical interactions. In addition, previous studies in the medical domain mainly segmented clinical text into sections and sentences, none of them considered segmentation of text into groups of MI behaviors in the setting of discourse analysis with emails. In this paper, we compared the performance of machine learning models for the task of segmentation of e-Coaching text. We found out that SVM provides the best performance for the segmentation of text in terms of all performance metrics. Manual segmentation of e-Coaching data is very resource-intensive and time-consuming task, which can significantly decrease the time and effort required to develop an effective behavioral intervention. Our proposed methods can help to identify individual text segments, which can be annotated directly with a classification model. This approach will also help for developing fully automated e-Coaching and accelerate the pace of identifying effective communication strategies.

Acknowledgments

This study was supported by a grant from the National Institutes of Health, NIDDK R21DK108071, Carcone and Kotov, MPIs. We would like to thank research staff and student assistants in the Department of Family Medicine and Public Health Sciences at Wayne State University School of Medicine for their help in preparing the training dataset.

References

[1] Miller WR, Rollnick S. Motivational interviewing: Helping people change. Guilford press; 2012.

[2] Miller WR, Rollnick S. Ten things that motivational interviewing is not. Behavioural and cognitive psychotherapy. 2009;37(2):129–140.

[3] Miller WR, Rose GS. Toward a theory of motivational interviewing. American psychologist. 2009;64(6):527.

[4] Apodaca TR, Longabaugh R. Mechanisms of change in motivational interviewing: a review and preliminary evaluation of the evidence. Addiction. 2009;104(5):705–715.

[5] Nigam K, McCallum AK, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using EM. Machine learning. 2000;39(2-3):103–134.

[6] Wang S, Manning CD. Baselines and bigrams: Simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. Association for Computational Linguistics; 2012. p. 90–94.

[7] Hasan M, Kotov A, Carcone AI, Dong M, Naar S, Hartlieb KB. A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories. Journal of biomedical informatics. 2016;62:21–31.

[8] Kotov A, Hasan M, Carcone A, Dong M, Naar-King S, Hartlieb KB. Interpretable probabilistic latent variable models for automatic annotation of clinical text. In: AMIA Annual Symposium Proceedings. vol. 2015. American Medical Informatics Association; 2015. p. 785.

[9] Hasan M, Kotov A, Carcone AI, Dong M, Naar-King S. Predicting the outcome of patient-provider communication sequences using recurrent neural networks and probabilistic models. In: Proceedings of the 2018 AMIA Informatics Summit. American Medical Informatics Association; 2018. .

[10] Webber B, Egg M, Kordoni V. Discourse structure and language technology. Natural Language Engineering. 2012;18(4):437–490.

[11] Apostolova E, Channin DS, Demner-Fushman D, Furst J, Lytinen S, Raicu D. Automatic segmentation of clinical texts. In: Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE. IEEE; 2009. p. 5905–5908.

[12] Denny JC, Spickard III A, Johnson KB, Peterson NB, Peterson JF, Miller RA. Evaluation of a method to identify and categorize section headers in clinical documents. Journal of the American Medical Informatics Association. 2009;16(6):806–815.

[13] Tepper M, Capurro D, Xia F, Vanderwende L, Yetisgen-Yildiz M. Statistical Section Segmentation in Free-Text Clinical Records. In: LREC; 2012. p. 2001–2008.

[14] Cho PS, Taira RK, Kangarloo H. Text boundary detection of medical reports. In: Proceedings of the AMIA Symposium. American Medical Informatics Association; 2002. p. 998.

[15] Griffis D, Shivade C, Fosler-Lussier E, Lai AM. A quantitative and qualitative evaluation of sentence boundary detection for the clinical domain. AMIA Summits on Translational Science Proceedings. 2016;2016:88.

[16] Kreuzthaler M, Schulz S. Detection of sentence boundaries and abbreviations in clinical narratives. In: BMC medical informatics and decision making. vol. 15. BioMed Central; 2015. p. S4.

[17] Treviso MV, Shulby C, Aluísio SM. Sentence segmentation in narrative transcripts from neuropsychological tests using recurrent convolutional neural networks. arXiv preprint arXiv:161000211. 2016;.

[18] Galley M, McKeown KR, Fosler-Lussier E, Jing H. Discourse segmentation of multi-party conversation. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics; 2003. .

[19] Alexander GL, Lindberg N, Firemark AL, Rukstalis MR, McMullen C. Motivations of Young Adults for Improving Dietary Choices: Focus Group Findings Prior to the MENU GenY Dietary Change Trial. Health Education & Behavior. 2017;p. 1090198117736347.

[20] Hashimoto K, Kontonatsios G, Miwa M, Ananiadou S. Topic detection using paragraph vectors to support active learning in systematic reviews. Journal of biomedical informatics. 2016;62:59–65.

[21] Lu HM, Wei CP, Hsiao FY. Modeling healthcare data using multiple-channel latent Dirichlet allocation. Journal of biomedical informatics. 2016;60:210–223.

[22] Ramage D, Hall D, Nallapati R, Manning CD. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. Association for Computational Linguistics; 2009. p. 248–256.

[23] McCallum A, Nigam K, et al. A comparison of event models for naive bayes text classification. In: AAAI-98 workshop on learning for text categorization. vol. 752. Citeseer; 1998. p. 41–48.

[24] Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995;20(3):273–297.

[25] Keller JM, Gray MR, Givens JA. A fuzzy k-nearest neighbor algorithm. IEEE transactions on systems, man, and cybernetics. 1985;(4):580–585.

[26] Lipton ZC, Berkowitz J, Elkan C. A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:150600019. 2015;.

[27] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;9(8):1735–1780.

[28] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:14123555. 2014;.

[29] Joachims T. Text categorization with support vector machines: Learning with many relevant features. In: European conference on machine learning. Springer; 1998. p. 137–142.

[30] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014;.

[31] Aas K, Eikvil L. Text categorisation: A survey. Technical report, Norwegian computing center; 1999.

[32] Kumar R, Indrayan A. Receiver operating characteristic (ROC) curve for medical researchers. Indian pediatrics. 2011;48(4):277–287.

[33] Hu J, Yang H, King I, Lyu MR, So AMC. Kernelized Online Imbalanced Learning with Fixed Budgets. In: AAAI; 2015. p. 2666–2672.