Machine Learning Models for the Segmentation of eCoaching Text

Mehedi Hasan, BS[1*], Alexander Kotov, PhD[1*], April Idalski Carcone, PhD[2], Ming Dong, PhD[1], Sylvie Naar, PhD[2]
[1]Department of Computer Science, Wayne State University, Detroit, Michigan
[2]Department of Family Medicine and Public Health Sciences, School of Medicine, Wayne State University, Detroit, Michigan

Abstract *Machine learning-based models has the potential to to efficiently and accurately identify patient-provider communication behaviors during eCoaching intervention sessions targeting fruit and vegetable intake among GenY young adults.*

1) Motivation: Why do we care about the problem and the results?
2) Problem statement: What problem is the paper trying to solve and what is the scope of the work?
3) Approach: What was done to solve the problem?
4) Results: What is the answer to the problem?
5) Conclusions: What implications does the answer imply?

Introduction

Unhealthy eating habits, particularly low fruit and vegetable intake, is a growing, serious public health concern, particularly among young adults age 21-30, referred to as Generation Y (GenY)[1,2]. This generation has adopted a lifestyle that involves eating accessible, "no mess", quick, "grab and go" foods[3,4]. They mainly eat "out" and infrequently shop and prepare food, limiting access to fruit and vegetables (FV)[5,6]. Unfortunately, less than one-third of US adults[1,7] and only 20% of GenY[1,8,9] eat the recommended 5 servings of fruit and vegetables daily. Those in inner city urban and rural settings have among the poorest eating habits[1,2,7–9]. GenY's poor dietary practices placing them at high risk for obesity and many chronic diseases, such as type 2 diabetes, as well as declines in predicted health status and life expectancy. Thus, there is a need to develop effective interventions to improve GenY's eating habits.

GenY is a tech-savvy generation requiring an intervention matched to their mobile lifestyle. Growing numbers use the internet to access health information with the largest increases in internet access among low-income Americans, making the internet well-suited for health promotion intervention[10]. MENU GenY[11] (Making Effective Nutrition Choices for Generation Y) is a technology-based public health intervention to encourage increased fruit and vegetable intake among GenY. A critical component of MENU GenY is personalized eCoaching. eCoaches use email to deliver motivation-enhancing coaching to encourage healthy eating, grounded in the principles of Motivational Interviewing (MI), an evidence-based communication technique to increase intrinsic motivation and self-efficacy for behavior change[12–14]. Patient "change talk", statements of intrinsic motivation about their desire, ability, reasons, need for and commitment to behavior change, is an established mediator of health behavior change[15]. Identifying specific communication strategies linked to behavior change and integrating these strategies into communication-based interventions (e.g., brief, motivation-enhancing interventions delivered in a variety of settings or public health initiatives) can increase these interventions' potency.

A major drawback of this research is the qualitative methods traditionally used to analyze the communication process which are resource-intensive, requiring an iterative process of human (subjective) interpretation of text. Rapidly developing computational technologies, specifically machine learning combined with classification models, offer a unique opportunity to accelerate this process. Our research group has recently applied machine learning-based models to similar communication data[16,17]. A simple communication code scheme was automated to characterize patient communication and achieved accuracy comparable to human coders[16]. The ultimate goal of the research study is to leverage innovative machine learning models to fully automate the communication coding process in eCoach-patient communication to increases in fruit and vegetable intake.

However, a significant barrier of fully automate eCoaching is the unsegmented text data. Developing an automatic

[*] Authors provided equal contribution.

classification of clinical interactions required segmented text. eCoaching data comprised of email responses which need to be segmented into “block of text” refers to a particular MI behavior. Automatic segementation of eCoaching intervention sessions is a challenging task due to the 2 important reasons. First, the email is an unstructured text that contains informal email exchange in non-traditional formats. Second, a text segment not necessarily belongs to the entire sentence or collection of sentences. One sentence can be segmented into several MI behaviors, and vice versa. Figure 1 illustrates that the marked sentence taken from an eCoaching email exchanage, segmented into 2 different MI behaviors, CHT and HUP.

In this paper, we addressing these issues by developing several state-of-the-art machine learning based models for the segmentation of eCoaching text to promote the automatic identification of best communication strategies without human interference. More specifically, we develop Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbor (KNN), Long Short Term Memory (LSTM), and Gated Recurrent Unit (GRU) by utlizing lexical, topic and punctuation mark features for the segmentation of eCoaching text.

Previous studies mainly focus on segemntation of text into sections, headers or sentences.

The authors are not aware of any other work where this approach has been considered. This approach will accelerate the pace of identifying effective communication strategies linked to healthy eating. This study can also be used to tailor interventions and develop fully automated eCoaching.

What is the problem?

Why is the problem important?

What has so far been done on the problem?
1) describes previous work in more technical detail
2) as far as needed for a proper understanding of the contribution of the paper

What is the contribution of the paper on the problem?
1) What are the rival approaches?
2) What are the drawbacks of each?
3) How has the battle between different approaches progressed?
4) What are the major outstanding problems? (This is where you come in)

Is the contribution original? Explain why

Is the contribution non-trivial? Explain why

Methods

Data collection

The experimental dataset for this work was constructed from the transcripts of 129 motivational interviews, which include a total of 50,239 segmented and annotated utterances. Each transcript consists of an MI interview session involving counselor, adolescent, and caregiver. The utterances were annotated based on MYSCOPE codebook[18], in which the codes are grouped into patient (adolescent and caregiver) codes and counselor codes. Utterances were divided into successful and unsuccessful communication sequences. Successful communication sequences result in positive change talk and commitment language statements by an adolescent or caregiver, while unsuccessful sequences are the ones that result in negative change talk or commitment language and the sequences, in which no change talk or commitment language statements occur. Out of 5143 observed sequences, 4225 were positive and 918 were negative. Successful sequences had an average length of 9.79 utterances, while unsuccessful sequences had on average 9.65

utterances. For each of the probabilistic models (MC and HMM), two models were trained, one model was trained using successful sequences and another one was trained using unsuccessful sequences.

Part 1

1) What algorithms or data structures did you select? Who created them? What is their asymptotic behavior? What other specific characteristics are worth noting for this study?
2) What programming language and platform did you use? What impact did these choices have on your project?

Part 2

1) How specifically did you implement the algorithms?
2) How did you handle instrumentation code? Why?
3) Did you perform any optimizations? Why or why not?
4) How did you choose to test and benchmark your code?
5) What inputs (data) did you select to test your implementations? Why?

Segmentation method

Generally, a sequence can be viewed as a temporally ordered set of events. In this study, an event is a behavior code that also has a symbolic representation.

Naive Bayes: comming soon...

Support Vector Machine: comming soon...

K-Nearest Neighbour: comming soon...

Recurrent Neural Networks: comming soon...

Evaluation metrics

Performance of the proposed method was evaluated in terms of precision, recall, and F-measure using 10 folds cross-validation and weighted macro-averaging of these metrics over the folds. However, LSTM and GRU are trained on 80% of the data and validated on 10%. The remaining 10% of the data is used as a test set for reporting the performance of the model.

Results

Experimental evaluation of the proposed method is conducted on both under and over-sampled sequences. Predictive performance summary of the proposed methods on under and over-sampled sequences is presented in Table.

1) In general, the pure, unbiased results should be presented first without interpretation (van Wagenen 1990).
2) These results should present the raw data or the results after applying the techniques outlined in the methods section. The results are simply results; they do not draw conclusions.

Predictive performance summary of the proposed methods on under and over-sampled sequences is presented in Table.

We took the inspiration for the representation of behavior codes from the idea of word embeddings. Word embedding is a representation of words in low-dimensional space by vectors, which contain the features of the words. In our study, we employed embedding in place of one-hot vectors for representation of behavior codes as input to LSTM and GRU, since one-hot vectors are high-dimensional and sparse.

Table 1: Performance of MC, HMM, and RNN for predicting the success of under and over-sampled patient-provider communication sequences. The highest value for each performance metric is highlighted in bold.

Method	lexical features only			lexical features + topics		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Naive Bayes Multinomial	0.7060	0.7044	0.7038	0.7932	0.7799	0.7775
Support Vector Machine	0.6395	0.6385	0.6379	0.7111	0.7029	0.7000
K-Nearest Neighbour	0.6244	0.6143	0.6067	0.7775	0.7567	0.7520
Long Short Term Memory	**0.8733**	**0.8681**	**0.8677**	**0.8424**	**0.8385**	**0.8381**
Gated Recurrent Unit	0.8705	0.8676	0.8673	0.8412	0.8377	0.8373

Table 2: Performance of MC, HMM, and RNN for predicting the success of under and over-sampled patient-provider communication sequences. The highest value for each performance metric is highlighted in bold.

Method	lexical features only			lexical features + topics		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Naive Bayes Multinomial	0.7060	0.7044	0.7038	0.7932	0.7799	0.7775
Support Vector Machine	0.6395	0.6385	0.6379	0.7111	0.7029	0.7000
K-Nearest Neighbour	0.6244	0.6143	0.6067	0.7775	0.7567	0.7520
Long Short Term Memory	**0.8733**	**0.8681**	**0.8677**	**0.8424**	**0.8385**	**0.8381**
Gated Recurrent Unit	0.8705	0.8676	0.8673	0.8412	0.8377	0.8373

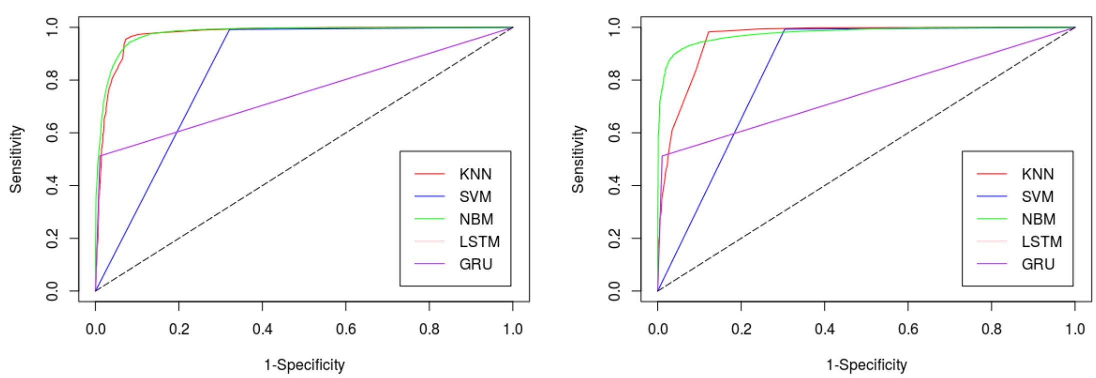

Figure 1: ROC curves compared the performance of different models.

Discussion

By analyzing the experimental results of different communication sequence outcome prediction methods proposed in this paper, we arrived at the following conclusions. First, the overall predictive performance of RNN models is substantially better than probabilistic models. In particular, the RNN-based method achieves near-human accuracy for predicting the

1) What, specifically, did you learn from comparing these algorithms or data structures?
2) What do your results say about the problem or question you were investigating?
3) Was your hypothesis confirmed or disproved?
4) Are the results what you expected?
5) If you obtained anomalies or other unexpected results, can you explain them? If not, how could you set about in the future to identify what caused them?
6) How do your results compare to past findings? Are they consistent? Different?Why?
7) How would you respond to objections or questions that other researchers might have about your methods, results,

or interpretations?
8) What is new and significant?

Conclusion

In this paper, we compared the performance of machine learning models for the task of segmentation of e-coaching text. We found out that k-nearest neighbour provides the best performance for the segmentaion 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 effective behavioral interventions. Our proposed methods can help to identify individual text segments, which can be annotated directly with a classification model and increase the effectiveness of behavioral interventions.

1) The hypothesis and the evidence for and against it are briefly restated
2) The original motivation is recapitulated
3) The state of the field in the light of this new contribution is reassessed
4) describes future research and new directions suggested by the contribution
5) in particular, research that would improve the evidence for or against the hypothesis

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 the student assistants in the Department of Family Medicine and Public Health Sciences at Wayne State University School of Medicine for their help in developing the training dataset by manually annotating the dataset using the MYSCOPE codebook.

References

[1] Blanck HM, Gillespie C, Kimmons JE, Seymour JD, Serdula MK. Trends in fruit and vegetable consumption among US men and women, 1994–2005. Preventing chronic disease. 2008;5(2).

[2] for Disease Control C, (CDC P, et al. Fruit and vegetable consumption among adults–United States, 2005. MMWR Morbidity and mortality weekly report. 2007;56(10):213.

[3] Nebeling L, Yaroch AL, Seymour JD, Kimmons J. Still not enough: can we achieve our goals for Americans to eat more fruits and vegetables in the future? American journal of preventive medicine. 2007;32(4):354–355.

[4] Brug J, Campbell M, van Assema P. The application and impact of computer-generated personalized nutrition education: a review of the literature. Patient education and counseling. 1999;36(2):145–156.

[5] Nelson MC, Lytle LA, Pasch KE. Improving literacy about energy-related issues: the need for a better understanding of the concepts behind energy intake and expenditure among adolescents and their parents. Journal of the American Dietetic Association. 2009;109(2):281–287.

[6] Larson NI, Perry CL, Story M, Neumark-Sztainer D. Food preparation by young adults is associated with better diet quality. Journal of the American dietetic association. 2006;106(12):2001–2007.

[7] Ogden CL, Carroll MD, Curtin LR, McDowell MA, Tabak CJ, Flegal KM. Prevalence of overweight and obesity in the United States, 1999-2004. Jama. 2006;295(13):1549–1555.

[8] Association ACH, et al. American college health association national college health assessment (ACHA-NCHA) spring 2005 reference group data report (abridged). Journal of American College Health. 2006;55(1):5.

[9] Thompson TG, Veneman AM. Dietary guidelines for Americans 2005. United States Department of Health and Human Services and United States Department of Agriculture. 2005;.

[10] Strecher V. Internet methods for delivering behavioral and health-related interventions (eHealth). Annu Rev Clin Psychol. 2007;3:53–76.

[11] 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.

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

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

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

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

[16] 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.

[17] Kotov A, Hasan M, Carcone A, Dong M, Naar-King S, BroganHartlieb K. 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.

[18] Carcone AI, Naar-King S, Brogan K, Albrecht T, Barton E, Foster T, et al. Provider communication behaviors that predict motivation to change in black adolescents with obesity. Journal of developmental and behavioral pediatrics: JDBP. 2013;34(8):599.