Developing Machine Learning Models for Behavioral Coding

April Idalski Carcone,¹ PhD, Mehedi Hasan,¹ BS, Gwen L. Alexander,² PhD, Ming Dong,¹ PhD, Susan Eggly,³ PhD, Kathryn Brogan Hartlieb,⁴ PhD, Sylvie Naar,⁵ PhD, Karen MacDonell,¹ PhD, and Alexander Kotov,¹ PhD

¹Wayne State University, ²Henry Ford Health System, ³Wayne State University and Barbara Ann Karmanos Cancer Institute, ⁴Florida International University, and ⁵Florida State University

All correspondence concerning this article should be addressed to April Idalski Carcone, PhD, Department of Family Medicine and Public Health Sciences, Wayne State University, Integrative Biosciences Building #2111, 6135 Woodward, Detroit, MI 48202, USA. E-mail: acarcone@med.wayne.edu

Received May 2, 2018; revisions received December 19, 2018; accepted December 20, 2018

Abstract

Objective The goal of this research is to develop a machine learning supervised classification model to automatically code clinical encounter transcripts using a behavioral code scheme.

Methods We first evaluated the efficacy of eight state-of-the-art machine learning classification models to recognize patient–provider communication behaviors operationalized by the motivational interviewing framework. Data were collected during the course of a single weight loss intervention session with 37 African American adolescents and their caregivers. We then tested the transferability of the model to a novel treatment context, 80 patient–provider interactions during routine human immunodeficiency virus (HIV) clinic visits. **Results** Of the eight models tested, the support vector machine model demonstrated the best performance, achieving a .680 F1-score (a function of model precision and recall) in adolescent and .639 in caregiver sessions. Adding semantic and contextual features improved accuracy with 75.1% of utterances in adolescent and 73.8% in caregiver sessions correctly coded. With no modification, the model correctly classified 72.0% of patient–provider utterances in HIV clinical encounters with reliability comparable to human coders ($k = .639$). **Conclusions** The development of a validated approach for automatic behavioral coding offers an efficient alternative to traditional, resource-intensive methods with the potential to dramatically accelerate the pace of outcomes-oriented behavioral research. The knowledge gained from computer-driven behavioral research can inform clinical practice by providing clinicians with empirically supported communication strategies to tailor their conversations with patients. Lastly, automatic behavioral coding is a critical first step toward fully automated eHealth/mHealth (electronic/mobile Health) behavioral interventions.

Key words: machine learning; motivational interviewing; qualitative research.

Introduction

Motivational interviewing (MI) is an evidence-based strategy for talking with patients about behavior change (Miller & Rollnick, 2012). Developed in adult substance abuse, MI has been widely adapted. Pediatric applications include increasing adolescents' adherence to chronic illness treatments (Schaefer &

Kavookjian, 2017), diet and physical activity (Bean et al., 2015), and caregivers' monitoring of type 1 diabetes self-management (Ellis et al., 2017). Meta-analyses of MI for pediatric health behaviors demonstrated a small, significant effect of MI over active and no treatment (Borrelli, Tooley, & Scott-Sheldon, 2015; Cushing, Jensen, Miller, & Leffingwell, 2014)

with one suggesting MI has greater efficacy for pediatric health behaviors than adult substance abuse (Gayes & Steele, 2014).

Motivational interviewing hypothesizes behavior change occurs through strengthening intrinsic motivation (engaging in an activity for reasons of inherent satisfaction rather than external stimuli, rewards, or consequences [Ryan & Deci, 2000]) as expressed through “change talk.” Change talk is patient statements expressing their internal desire, ability, reasons, need for, or commitment to behavior change. Counselors’ evoke change talk by using “MI-consistent communication skills” (MICO; e.g., open-ended questions and reflections). In contrast, MI-inconsistent communication (MIIN; e.g., warning about behavioral consequences and confronting) leads to arguments against behavior change, referred to as counter change talk or sustain talk (Miller & Rose, 2009). To empirically test this hypothesis, researchers rely upon sequential analysis, an analytic approach to examine the temporal sequencing of behavioral events (Bakeman & Quera, 1997, 2011).

Moyers and Martin (2006) developed the Sequential Code for Observing Process Exchanges (SCOPE), which operationalizes clinical communication into 30 discrete provider and 16 patient behaviors. Utterances (speech segments representing complete thoughts) were characterized using the SCOPE, generating a temporally sequenced stream of communication data for hypothesis testing. Only two other research groups have utilized this approach (i.e., Carcone et al., 2013; Gaume, Gmel, Faouzi, & Daepfen, 2008). The paucity of patient–provider communication process research can partly be attributed to the iterative, resource-intensive, cognitively-demanding coding procedures required.

Recent successes in the development of artificial intelligence techniques, like machine learning, have enabled computers to achieve near-human accuracy for simple cognitive tasks, including classification and pattern recognition (Bishop, 2007), opening up a new frontier in behavioral research. Broadly speaking, machine learning refers to a class of statistical techniques in which computer systems “learn” (e.g., progressively improve their performance on a task) to recognize patterns in data without being explicitly programmed to do so. Machine learning models have been used in a variety of contexts relevant to pediatric psychologists, such as modeling parent–infant interactional patterns to understand how they shape behavior during infant development (Messinger, Ruvolo, Ekas, & Fogel, 2010). Machine learning models have been used to mine biomedical data for diagnostic prediction. To illustrate, computer models have been used to boost the analytic power of functional magnetic resonance

imaging (fMRI) in the classification of brain states in predicting disease and prenatal exposures (Deshpande, Li, Santhanam, Coles, & Lynch, 2010). These models have been used to predict mental health outcomes (e.g., suicide) from medical chart data (Adamou et al., 2018) and developmental outcomes (e.g., developmental language disorders) from screening instrument data (Armstrong et al., 2018). We propose a new application of machine learning models—behavioral coding.

Supervised machine learning refers to methods in which a computer uses a labeled training dataset (i.e., coded data) to learn a mathematical model to map inputs (raw data) to the result of a particular cognitive task (behavior represented by a code, or label), mimicking the process by which human coders learn to perform the same task. We wished to map coded utterances from clinical transcripts to a codebook of MI behaviors. The success of a machine learning application depends on two factors. The first is whether a model has sufficient capacity (i.e., complexity) for the analytical task. Selecting a model with insufficient capacity will result in underfitting, i.e., the model is unable to accurately learn the required mapping. Choosing a model with too much capacity may result in overfitting, or the model learning unnecessary peculiarities of the training data limiting its ability to generalize its learning to novel data. The second factor is whether sufficient training data is available for the model to reliably learn the required mapping, with more complex models generally requiring more training data.

Supervised machine learning models are a natural fit for behavioral coding in pediatric psychological research, such as the analysis of pediatric patient–provider communication during MI sessions. Several researchers have examined the utility of various supervised machine learning methods for the assessment of counselor fidelity to the MI framework in adult treatment contexts. Can et al. (2016) explored the Maximum Entropy Markov Model (MEMM) to classify counselor reflections as operationalized by the MI Skill Code (MISC, a common MI fidelity measure; Miller, Moyers, Ernst, & Amrhein, 2008) achieving 94% accuracy. Atkins, Steyvers, Imel, and Smyth (2014) developed a labeled topic model to predict 12 MISC codes that demonstrated greater performance in the classification of counselor behaviors, but poor performance with patient behavior. Tanana et al. (2015) found recursive neural networks (RNNs) resulted in improved accuracy for the classification of patient behavior but were less accurate than human coders. Discrete Sentence Feature models were also unable to reliably predict patient speech illustrating the difficulty in modeling this behavior (Tanana, Hallgren, Imel, Atkins, & Srikumar, 2016). Pérez-Rosas et al. (2017)

were more successful with a support vector machine (SVM) classifier to distinguish counselor questions (~87% accuracy) and reflections (~82%), but were less successful with a multiclass model to distinguish other counselor behaviors (17-80%) as articulated by the MI Treatment Integrity Scale (MITI; Moyers, Martin, Manuel, Miller, & Ernst, 2010). In summary, research to develop models to automate fidelity coding have been successful with some provider behaviors (i.e., questions and reflections), but less so with other provider behaviors and adult patient behaviors.

Concurrently, we began experimenting with state-of-the-art supervised classification methods for the task of automated coding of patient behaviors in pediatric MI transcripts. In addition to accuracy, we were concerned with interpretability of machine learning models. Interpretability increases the transparency of the classifier and gives MI researchers and clinicians insight into the linguistic characteristics (e.g., individual words and phrases) of different communication behaviors. Specifically, we proposed two interpretable probabilistic generative models, Latent Class Allocation (LCA) and Discriminative Labeled Latent Dirichlet Allocation (DL-LDA), and compared their performance with other popular classifiers, including probabilistic models, such as Naïve Bayes (NB), Labeled Latent Dirichlet Allocation (L-LDA) and SVM (Kotov, Idalski Carcone, Dong, Naar-King, & Brogan, 2014; Kotov et al., 2015). Results indicated that LCA had the best precision and F1 score and DL-LDA had the best recall, however, SVM outperformed all other methods (Kotov et al., 2015). SVM achieved 65% and 74% F1-score for classification of patient behaviors into 5 and 3 distinct behavior categories, respectively. Here we report our efforts to further develop a machine learning model to code clinical transcripts in pediatric settings. These experiments are part of a line of research to develop a fully automated behavioral coding procedure—including models to segment (parse; Hasan, Kotov, Naar, Alexander, & Idalski Carcone, 2019), annotate (code), sequentially analyze (Hasan et al., 2018), and predict the outcome of interview sessions (Hasan, Kotov, Carcone, Dong, & Naar, 2018)—to study patient-provider behaviors expressed during clinical encounters according to the MI framework.

Study 1

The goal of Study 1 was to identify a supervised machine learning model effective at the task of automatic annotation (coding) of patient-provider communication behaviors as operationalized by the MI framework (Miller & Rollnick, 2012). Because we were interested in training the model to recognize

verbalization patterns characteristic of patient-provider communication expressed during MI sessions, this study was a secondary analysis of data previously analyzed using traditional behavioral coding (Carcone et al., 2013).

Materials and Methods

The training dataset was composed of weight loss intervention sessions with 37 African American adolescents (12.0 to 17.0 years; $M = 14.7$, $SD = 1.63$) with obesity (body mass index (kg/m^2) $\geq 95^{\text{th}}$ percentile) and their caregivers (89% biological mothers, 67% 2-parent homes). Caregivers provided informed consent and adolescents provided assent. The institutional review board of the affiliated academic institution approved the research.

Each family took part in one intervention session led by a highly trained MI counselor. The goal of these sessions was to support adolescent autonomy while exploring motivation for weight-related behavior change and setting weight loss goals consistent with the adolescent's motivation for change. The counselor first met with the adolescent alone, then with the caregiver to explore her own weight loss goals and ways to support her child's weight loss goals. Sessions ended with adolescents and caregivers coming together to share their weight loss plans. Because the coding tool used was designed for dyadic interactions, only the adolescent-counselor and caregiver-counselor portions of the session were analyzed.

Sessions were video recorded and professionally transcribed. The transcribed sessions were manually coded using the Minority Youth—Sequential Code for Observing Process Exchanges (MY-SCOPE; Carcone et al., 2013), a qualitative code scheme based upon the SCOPE coding system (Moyers & Martin, 2006). We modified the SCOPE to reflect the language of the target population (minority youth) and expanded the behavior codes to 115 patient and provider communication behaviors operationalized according to the MI framework. When using the MY-SCOPE to code sessions, coders treat each patient speaking turn (an uninterrupted segment of speech) as the unit of analysis. Coders may parse provider speaking turns into utterances (complete thoughts representing distinct behaviors) to capture the fact that providers often use more than one MI strategy during a single speaking turn, for example, a counselor may make a reflection, support the patient's autonomy, and ask a question within a single speaking turn. A primary coder coded all 37 sessions and a second coder coded a randomly selected 20% ($n = 7$) to assess inter-rater reliability, which was assessed with Cohen's kappa ($k = .696$). The coded dataset was composed of 11,353 coded utterances, 6,579 from adolescent-counselor

conversations and 4, 774 from caregiver–counselor conversations.

The goal of the analysis was to identify the classification model demonstrating the greatest accuracy in the prediction of the MY-SCOPE codes. The manually coded data served as the training data for eight candidate machine learning classification models. First, we evaluated the performance of the selected classifiers when only lexical features were used (i.e., when provider or patient utterances were represented as a unigram bag-of words); in other words, single word associations were used to characterize behavior codes. Candidate models (Table I) included Naïve Bayes (NB; Kibriya, Frank, Pfahringer, & Holmes, 2004; McCallum & Nigam, 1998; Rish, 2001), Naïve Bayes-Multinomial (NB-M), J48 (Sharma & Sahni, 2011), AdaBoost (Freund, Schapire, & Abe, 1999), Random Forest (RF; Breiman, 2001), DiscLDA (Lacoste-Julien, Sha, & Jordan, 2009), Convolutional Neural Network (CNN; Kim, 2014), and SVM (Cortes & Vapnik, 1995; Durgesh & Lekha, 2010) using a one-versus-all strategy. These models have been successfully utilized for textual classification, like sentiment analysis (dos Santos & Gatti, 2014; Wang & Manning, 2012).

The top performing model (SVM) was then expanded to include contextual and semantic features to assess the extent to which these features improved model performance. The code assigned to the preceding utterance served as the contextual feature in the SVM-PL model, for example, reflections of change talk are likely to follow a change talk statement. Word distributions derived from the Linguistic Inquiry and Word Count dictionary (LIWC; Tausczik & Pennebaker, 2010) was the semantic feature, denoted as SVM-LIWC. The LIWC is a well-established psycho-linguistic lexicon of words associated with over 80 thought processes, emotional states, intentions, and motivations. In addition, we experimented with Conditional Random Fields (CRF; Lafferty, McCallum, & Pereira, 2001; Sutton & McCallum, 2006), a model that explicitly combines the lexical feature with the previous code when assigning a code to the current utterance (Table II).

Prior to analysis, several data preparation steps were undertaken. First, low frequency codes were merged with other theoretically similar codes (e.g., change talk was originally coded to reflect the strength of client language ranging from the weakest statement “change talk 1” to the strongest “change talk 3”, these categories were merged into a single “change talk” category). We examined the models’ performance with codebooks of varying size by further combining conceptually similar codes (e.g., change talk and commitment language were merged into a single change talk category). This resulted in codebooks with 41, 20,

and 17 codes for the adolescent–counselor interactions and 58, 19, and 16 for caregiver–counselor interactions. The final data preparation step involved word stemming (Snowball stemmer implemented within the Weka [Hall et al., 2009] machine learning toolkit), which truncates inflections from words (e.g., “eating” is stemmed to “eat”). Stop words (the most commonly occurring words in a language) were not removed because their removal significantly reduced the models’ performance; hence, despite their ubiquitous nature, they provided important lexical clues in code determination.

Model performance was evaluated using 10-fold cross-validation (Kohavi, 1995). Evaluation metrics were calculated using weighted macroaveraging over 10-folds. *Accuracy* was the number of true positive or true negative observations identified/total observations. *Precision* was the number of true positive observations identified/total number of predicted positive observations. *Recall* represented the number of true positive observations identified/total real positive observations. The *F1-measure* is a function of precision and recall: $(2 \times \text{precision} \times \text{recall}) / (\text{precision} + \text{recall})$ (Powers, 2011). Cohen’s kappa is a measure of inter-rater agreement corrected for chance agreement; it is computed as $(\text{observed accuracy} - \text{expected accuracy}) / (1 - \text{expected accuracy})$ (Wood, 2007). Kappas <0.40 are considered “fair” to “poor,” 0.41–0.60 are “moderate,” 0.61–0.80 are “substantial,” and >0.81 are “almost perfect” (Landis & Koch, 1977).

Results

First, we compared the performance of all classification models using only lexical features (Table III). Experimental results show that SVM consistently demonstrated the best performance among all machine learning models across all metrics and codebook sizes on both adolescent and caregiver interview session transcripts. The SVM achieved 50.4%, 59.2%, and 68.0% F1-score in adolescent sessions with 41, 20, and 17 codes, respectively, and 41.8%, 53.5%, and 63.9% in caregiver sessions with 58, 19, and 16 codes, respectively. However, the performance of all classification models is consistently lower on caregiver sessions compared to adolescent sessions because adolescents generally responded with simpler language.

We then evaluated model performance improvement when semantic and contextual features were used in conjunction with the lexical features (Table II). The performance of the SVM model improved in both the adolescent and caregiver datasets across codebook sizes. The accuracy of the SVM model improved by 3–10% (F1 score) when the contextual or semantic features were used in addition to the lexical features. When all three feature types were used, SVM outperformed all other methods with the

Table I. Classification Model Performance in Weight Loss Intervention Using Only Lexical Features

Codebook size	Model	Accuracy	Precision	Recall	F1-score	Kappa
Adolescent–counselor interactions						
17	NB	0.544	0.603	0.544	0.552	0.497
	NB-M	0.670	0.662	0.670	0.643	0.622
	J48	0.595	0.573	0.595	0.580	0.539
	AdaBoost	0.627	0.600	0.627	0.609	0.574
	RF	0.670	0.662	0.670	0.625	0.616
	DiscLDA	0.477	0.454	0.477	0.431	0.388
	CNN	0.678	0.633	0.678	0.670	0.509
20	SVM	0.708	0.705	0.708	0.680	0.663
	NB	0.487	0.509	0.487	0.482	0.448
	NB-M	0.579	0.582	0.579	0.559	0.537
	J48	0.479	0.467	0.479	0.470	0.431
	AdaBoost	0.504	0.488	0.504	0.493	0.458
	RF	0.563	0.564	0.563	0.519	0.514
	DiscLDA	0.400	0.410	0.400	0.356	0.330
41	CNN	0.586	0.588	0.586	0.587	0.476
	SVM	0.610	0.611	0.610	0.592	0.571
	NB	0.406	0.434	0.406	0.405	0.375
	NB-M	0.513	0.479	0.513	0.484	0.478
	J48	0.396	0.375	0.396	0.382	0.356
	AdaBoost	0.436	0.412	0.436	0.421	0.398
	RF	0.495	0.487	0.495	0.453	0.455
	DiscLDA	0.362	0.387	0.362	0.301	0.304
	CNN	0.396	0.369	0.396	0.382	0.170
	SVM	0.537	0.513	0.537	0.504	0.502
Caregiver–counselor interactions						
16	NB	0.571	0.608	0.571	0.575	0.518
	NB-M	0.633	0.629	0.633	0.604	0.573
	J48	0.578	0.563	0.578	0.567	0.514
	AdaBoost	0.602	0.582	0.602	0.588	0.539
	RF	0.640	0.631	0.640	0.596	0.574
	DiscLDA	0.482	0.442	0.482	0.421	0.362
	CNN	0.657	0.641	0.657	0.648	0.512
19	SVM	0.664	0.653	0.664	0.639	0.606
	NB	0.477	0.504	0.477	0.467	0.434
	NB-M	0.536	0.539	0.536	0.512	0.487
	J48	0.436	0.431	0.436	0.432	0.382
	AdaBoost	0.467	0.457	0.467	0.460	0.415
	RF	0.507	0.508	0.507	0.467	0.450
	DiscLDA	0.374	0.370	0.374	0.333	0.287
58	CNN	0.510	0.498	0.510	0.504	0.401
	SVM	0.545	0.547	0.545	0.535	0.497
	NB	0.379	0.392	0.379	0.370	0.350
	NB-M	0.442	0.404	0.442	0.386	0.401
	J48	0.340	0.321	0.340	0.328	0.302
	AdaBoost	0.381	0.359	0.381	0.366	0.344
	RF	0.402	0.358	0.402	0.352	0.358
	DiscLDA	0.288	0.258	0.288	0.234	0.229
	CNN	0.118	0.102	0.118	0.109	0.032
	SVM	0.451	0.420	0.451	0.418	0.414

Note. The F1-score is a function of model precision and recall. CNN = Convolutional Neural Network; NB = Naive Bayes; NB-M = Naive Bayes-Multinomial; RF = Random Forest; SVM = support vector machine. Bold text indicates the best performing model for each codebook.

greatest accuracy observed in the adolescent codebook with 17 classes (75.1%) and the caregiver codebook with 16 classes (73.8%).

Study 2

In Study 2, we tested the accuracy and reliability of the machine learning classification model developed in Study 1 in a new treatment setting, human

immunodeficiency virus (HIV) medical care. The training dataset for Study 2 was composed of 80 patient–provider clinical interactions during routine HIV clinic visits previously coded with the MYSCOPE coding instrument (Idalski Carcone & Naar, 2017). Our working hypothesis was that the classification model developed in Study 1 would demonstrate transferability of knowledge by achieving a high level of coding accuracy.

Table II. Classification Model Performance in Weight Loss Intervention Using Multiple (Lexical, Contextual, and Semantic) Features

Codebook size	Model	Accuracy	Precision	Recall	F1-score	Kappa
Adolescent–counselor interactions						
17	CRF	0.682	0.673	0.682	0.677	0.636
	SVM	0.708	0.705	0.708	0.680	0.663
	SVM-PL	0.715	0.711	0.715	0.696	0.673
	SVM-LIWC	0.742	0.740	0.742	0.727	0.704
	SVM-AF	0.751	0.750	0.751	0.739	0.715
20	CRF	0.581	0.579	0.581	0.580	0.540
	SVM	0.610	0.611	0.610	0.592	0.571
	SVM-PL	0.639	0.642	0.639	0.630	0.604
	SVM-LIWC	0.653	0.653	0.653	0.657	0.619
	SVM-AF	0.682	0.685	0.682	0.674	0.651
41	CRF	0.493	0.485	0.493	0.457	0.502
	SVM	0.537	0.513	0.537	0.504	0.502
	SVM-PL	0.565	0.543	0.565	0.542	0.535
	SVM-LIWC	0.538	0.518	0.538	0.507	0.503
	SVM-AF	0.568	0.549	0.568	0.546	0.538
Caregiver–counselor interactions						
16	CRF	0.654	0.652	0.654	0.653	0.603
	SVM	0.664	0.653	0.664	0.639	0.606
	SVM-PL	0.670	0.658	0.670	0.651	0.614
	SVM-LIWC	0.730	0.730	0.730	0.717	0.686
	SVM-AF	0.738	0.733	0.738	0.727	0.696
19	CRF	0.539	0.541	0.539	0.540	0.492
	SVM	0.545	0.547	0.545	0.535	0.497
	SVM-PL	0.566	0.570	0.566	0.559	0.522
	SVM-LIWC	0.620	0.625	0.620	0.613	0.581
	SVM-AF	0.638	0.639	0.638	0.631	0.601
58	CRF	0.438	0.409	0.438	0.423	0.385
	SVM	0.451	0.420	0.451	0.418	0.414
	SVM-PL	0.480	0.462	0.480	0.456	0.446
	SVM-LIWC	0.459	0.445	0.459	0.429	0.422
	SVM-AF	0.488	0.466	0.488	0.462	0.454

Note. The F1-score is a function of model precision and recall. Bold font denotes the model with the best performance in each dataset. CRF = Conditional Random Field; SVM = support vector machine; SVM-AF = Support Vector Machine All Features; SVM-LIWC = Support Vector Machine Linguistic Inquiry and Word Count Dictionaries, SVM-PL = Support Vector Machine.

Materials and Methods

Participants were recruited from a multidisciplinary HIV outpatient clinic located within a large urban teaching hospital providing primary medical care annually to over 200 adolescents and young adults living with HIV. All identified members of the multidisciplinary care team were eligible and invited to participate; this included physicians, nurses, psychologists, social workers, outreach workers/advocates, and case managers. Patients up to age 25 presenting in the HIV clinic for routine HIV follow-up care with a consenting provider were eligible. A total of 192 patient–provider encounters were observed; of these, 64 were excluded because they were either less than 5 min in duration ($n = 61$) or involved participants other than the patient and provider ($n = 3$). Of the remaining 128, 80 were randomly selected for coding. A little more than half (56%, $N = 45$) were psychosocial providers ($n = 23$ psychologists/psychiatrists, $n = 20$ social workers, and $n = 2$ health/outreach workers); the rest (44%, $N = 35$) were medical providers ($n = 15$

physicians, $n = 15$ nurses, and $n = 5$ residents/fellows). No demographic information was collected from either providers or patients in this descriptive study; however, the clinic serves a diverse patient population composed primarily of African American patients.

This study used naturalistic observation of medical encounters. Prior to initiating data collection, providers signed informed consent. Patients were approached upon arrival to the HIV clinic for routine appointments. After obtaining patients' informed consent, audio recorders were placed in patient exam rooms and recorded their clinical encounter. No additional data were collected from patients or providers. A professional transcription company completed a verbatim transcription of the audio recordings for qualitative coding. Interruptions (e.g., other providers entering the room to speak to the provider, patient telephone conversations) were excluded.

These encounters were coded with the MY-SCOPE which had been adapted to study patient–provider communication during HIV medical care visits

Table III. Classification Model Performance in Human Immunodeficiency Virus (HIV) Medical Encounters Using Multiple (Lexical, Contextual, and Semantic) Features

Behavior codes	Accuracy	Precision	Recall	F1-score	Kappa	N or %
Overall	0.720	0.701	0.720	0.696	0.663	11,889
Patient codes						
Change talk	0.426	0.609	0.426	0.501	0.483	4%
Sustain talk	0.451	0.665	0.451	0.537	0.526	3%
High uptake HIV-related	0.500	0.522	0.500	0.511	0.495	3%
High uptake other	0.913	0.848	0.913	0.879	0.828	29%
Low uptake	0.747	0.743	0.747	0.745	0.721	9%
Counselor codes						
Question to elicit change talk	0.410	0.442	0.410	0.425	0.415	2%
Question to elicit sustain talk	0.495	0.495	0.495	0.495	0.487	2%
Neutral question	0.270	0.445	0.270	0.336	0.325	2%
Other question	0.885	0.777	0.885	0.828	0.783	19%
Reflection of change talk	0.182	0.377	0.182	0.245	0.239	1%
Reflection of sustain talk	0.364	0.392	0.364	0.377	0.375	<1%
Other reflection	0.132	0.531	0.132	0.212	0.201	3%
Affirmation	0.194	0.737	0.194	0.308	0.306	1%
Emphasize autonomy	0.182	0.769	0.182	0.294	0.293	<1%
Provide information, not patient-centered	0.210	0.46	0.210	0.289	0.283	1%
Provide information, patient-centered	0.453	0.459	0.453	0.456	0.437	3%
MI-inconsistent behavior (Confront, Warn, Advise)	0.112	0.474	0.112	0.181	0.177	1%
Structure session	0.195	0.473	0.195	0.276	0.268	2%
Other statement	0.887	0.647	0.887	0.748	0.700	14%

Note. The F1-score is a function of model precision and recall. MI = motivational interviewing.

(Idalski Carcone et al., 2018). Specifically, we edited the examples of patient–provider communication behavior to illustrate the target behaviors in HIV clinical care: HIV medication adherence, regular HIV clinic attendance, and behaviors that place the patient at risk for transmitting the HIV virus (e.g., unprotected sex, drug use, and alcohol abuse). Two coders were trained to reliably use MY-SCOPE. Coders were randomly assigned equal numbers of encounters to code with 20% co-coded for inter-rater reliability, which was assessed with Cohen’s kappa ($k = .688$).

The SVM model with all feature types (lexical, contextual, and semantic) developed in Study 1 was applied to the HIV dataset. Model performance was evaluated using the same 10-fold cross-validation procedure and metrics (precision, recall, F-1, kappa). As in Study 1, data preparation involved merging low frequency and conceptually similar codes, which resulted in a codebook with 20 behavioral codes and word stemming.

Results

The dataset included 11,889 utterances, of which 6,592 (55%) were providers and 5,297 (45%) were patients. The SVM model, with no modifications from Study 1, achieved 69.6% F1-score with 70.1% precision and 72.0% recall for the task of automatic annotation of utterances in patient–provider encounters in HIV clinic (Table III). The model demonstrated good reliability compared to human coders as assessed by Cohen’s kappa, $k = .663$ indicating that the model performed better than random chance at

differentiating MY-SCOPE codes overall. The kappa values for individual communication behaviors were variable, ranging from 0.177 to 0.828.

To better understand the variability observed at the individual code level, we conducted a post hoc review of the confusion matrix (a cross-tabulation table of the manually assigned and computer assigned behavior codes). The confusion matrix revealed the model misclassified patient language as provider language <1% of the time ($n = 47$) and provider language as patient language 5% of the time ($n = 624$). Patient language was classified as the wrong patient language category 5% of the time ($n = 585$). Provider language was classified as the wrong provider language category 17% of the time ($n = 2075$). The most confused codes were providers’ MI-inconsistent communication behaviors, other reflections, statements emphasizing patient autonomy, and affirmations with 11%, 13%, 18%, and 19% accuracy, respectively. The least confused codes were patients’ high uptake statements (91% accuracy), providers’ other questions (89%), providers’ other statements (89%), and patients’ low uptake (75%).

Discussion

We developed a machine learning supervised classification model to code pediatric/young adult patient–provider communication behaviors during treatment encounters within the MI framework. Our SVM model, using lexical (discrete words observed), contextual (code assigned to the previous utterance), and

semantic features (latent cognitive states informed by the Linguistic Inquiry and Word Count dictionaries), accurately coded 75% of pediatric patient–provider communication behaviors during weight loss treatment sessions and, with no modification, 72% of patient–provider communication behaviors during routine HIV clinical care. The inter-rater reliability between human coders and the classification model was very good in both datasets (Study 1: $k = .715$, Study 2: $k = .663$) and comparable to the inter-rater reliability observed between the human coders (Study 1 $k = .696$, Study 2: $k = .688$). These results illustrate the accuracy of our newly developed SVM-based machine learning classification model with lexical, contextual, and semantic features is comparable to that of human coders. And, the effectiveness of transfer learning strategies or applying machine learning models trained on one clinical context (e.g., weight loss) to another clinical context (e.g., HIV patient visits). Effective transfer of machine learned models can significantly reduce the time and resources needed to develop the training datasets for different types of clinical discourse.

These results add to the nascent literature on applications of machine learning to behavioral coding in MI. Previous research has focused primarily on the assessment of provider fidelity in adult treatment contexts. Researchers have successfully developed models for the accurate classification of provider questions (Pérez-Rosas et al., 2017) and reflections (Can et al., 2016; Pérez-Rosas et al., 2017). Efforts to develop a model to accurately classify multiple patient and/or provider behaviors, however, have been less successful (Pérez-Rosas et al., 2017; Tanana et al., 2015, 2016). Not only does our research extend the focus of machine learning methods used in MI to include a code scheme designed to test an important tenet of MI theory, the technical hypothesis, we have developed a model that, overall, accurately classifies multiple pediatric patient and provider behaviors. Automated classification of behavioral data provides the opportunity to accelerate the pace of outcomes-oriented behavioral research by offering an efficient alternative to traditional, resource-intensive methods of behavioral coding.

Although we are optimistic about these results, we do acknowledge that there are opportunities for model improvement. Overall, we found the number of classification errors decreased when conceptually similar behaviors were combined; however, there was significant variability in the model's accuracy at the level of the individual behavior code. Atkins et al. (2014) pointed out that machine learning models more accurately predict behaviors with a syntactically predictable structure (i.e., questions, which often start with key words such as “What”, “Why”, and “How”, and

reflections, which frequently include key words like “It sounds like” or “You said”). Other provider behaviors (e.g., affirmations) and patient motivational statements (e.g., change talk and sustain talk) often indicate an underlying cognitive state that is not necessarily having a predictable syntactic structure. This assertion is supported by the fact that our model's performance improved when contextual and semantic features were used in addition to the lexical features, that is, provided information related to the speaker's cognitive state. Despite this overall improvement in model accuracy, our model did misclassify 28% of communication behaviors. Behaviors with the highest rates of misclassification were low frequency provider behaviors with a syntactically indistinct structure due to either reflecting an underlying cognitive state (statements emphasizing patient autonomy and affirmations) or being a combination of behaviors (MI-inconsistent communication behaviors) or subject matter (other reflections). Thus, a challenge of applying machine learning approaches to behavioral coding is the identification and integration of features that illuminate the underlying cognitive state expressed through spoken language. One strategy might be to combine the Linguistic Inquiry and Word Count dictionaries with the dependency tree features employed by Tanana et al. (2015, 2016).

Another strategy to improve model performance is to increase the amount of training data. In particular, affirmations, statements emphasizing autonomy, and reflections of sustain talk represented a very small proportion of the HIV dataset (<0.06% of utterances), which led to their greater misclassification. The primary barrier for increasing the volume of training data is the time-consuming and resource-intensive behavioral coding process. Active learning is a strategy used to train classification models that involves interaction between the model and manual coders (Tong & Koller, 2001). During this interaction, the model selectively requests coders to provide training data only for the behavioral codes that are most challenging to classify. Careful selection of training examples during active learning can dramatically reduce the size of the training dataset required for the classifier to achieve a given level of accuracy.

The knowledge gained from computer-driven behavioral coding could also inform clinical practice. The ability of machine learning models to classify large quantities of clinical data into more and less effective communication sequences can increase the ability of clinicians, like pediatric psychologists, to integrate evidence-based communication strategies into their practice. These models' capacity to adapt and improve their performance by continuously incorporating new data allows the development of more effective interventions via tailoring. Clinicians could

individualize treatment by selecting evidence-based communication strategies using any number of relevant factors, such as patient age, target behavior, and treatment context. Leveraging machine learning models in this manner could be an effective strategy to align behavioral medicine with the precision medicine initiative (Collins & Varmus, 2015).

Automatic behavior coding is a critical first step in developing fully automated pediatric behavioral eHealth/mHealth interventions. Technology-delivered interventions are easily streamlined into point-of-care clinical service delivery (e.g., medical care visits) or delivered to anyone with an internet connection. With MIs demonstrated efficacy for the treatment of pediatric health behaviors (Borrelli et al., 2015; Cushing et al., 2014; Gayes & Steele, 2014), e/mHealth-delivered MI could vastly expand the reach of pediatric behavioral intervention. To illustrate, our research group has developed MI-based eHealth interventions to support pediatric (Rajkumar et al., 2015) and young adult (MacDonell, Naar, Gibson-Scipio, Lam, & Secord, 2016; Naar-King et al., 2013) patients' chronic illness management. These interventions have a highly structured, preprogrammed architecture that relies upon patients selecting the best fitting option from a predefined list of commonly reported experiences culled from the literature and clinical experience. Integrating an automated coding algorithm into these interventions would allow patients to explain their experiences in their own words, rather than choosing an option that might not fully describe their experience. The coding algorithm would then analyze their comments, provide appropriate feedback and guide them to intervention content consistent with their motivational state, more closely approximating a traditional clinical encounter.

We view this research as promising while acknowledging that the research is not without limitations. First, these results are based on two relatively small, nonrandomly selected samples of patient-provider interactions. While the number of coded utterances (>11,000) is sufficient for machine learning modeling, further testing the model's performance would be improved with additional examples of communication behaviors in different samples. Second, the classification model reported here is restricted to the application of MY-SCOPE codes to previously parsed and coded data. Efforts are underway to develop a machine learning model to parse a clinical transcript into utterances that could then be coded with MY-SCOPE machine learning model. Together, the parsing and MY-SCOPE classification models would be a step closer toward the goal of developing a fully automated behavioral coding procedure. A final limitation of this approach is its dependence on the support of a computer scientist. Translating the model to a different

clinical context or behavioral code scheme would require developing a new training dataset and model specification. Despite these limitations, machine learning models hold promise for scaling up resource-intensive cognitive tasks, such as qualitative coding.

In conclusion, the goal of this research is to scale up MI research by testing MIs technical hypothesis via machine learning models. The methods discussed here offer an opportunity to accelerate the pace of outcomes-oriented behavioral research faster than any coding laboratory staffed by human coders could feasibly achieve. This efficiency represents not only vast research opportunities, but great clinical relevance as well.

Acknowledgments

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 preparing the training datasets.

Funding

This work was supported by the National Institute of Diabetes, Digestive, and Kidney Diseases (grant number R21DK108071) and the National Institute of Mental Health (grant number R34MH103049).

Conflicts of interest: None declared.

References

- Adamou, M., Antoniou, G., Greasidou, E., Lagani, V., Charonyktakis, P., & Tsamardinos, I. (2018). *Mining free-text medical notes for suicide risk assessment*. Paper presented at the Proceedings of the 10th Hellenic Conference on Artificial Intelligence, Patras, Greece.
- Armstrong, R., Symons, M., Scott, J., Arnott, W., Copland, D., McMahon, K., & Whitehouse, A. (2018). Predicting language difficulties in middle childhood from early developmental milestones: A comparison of traditional regression and machine learning techniques. *Journal of Speech, Language, and Hearing Research: JSLHR*, 61, 1926–1944.
- Atkins, D. C., Steyvers, M., Imel, Z. E., & Smyth, P. (2014). Scaling up the evaluation of psychotherapy: Evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, 9, 1–11. doi:10.1186/1748-5908-9-49
- Bakeman, R., & Quera, V. (1997). *Observing interaction: An introduction to sequential analysis* (2nd edn). New York, NY: Cambridge University Press.
- Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences*. New York, NY: Cambridge University Press.
- Bean, M. K., Powell, P., Quinoy, A., Ingersoll, K., Wickham, E. P., 3rd., & Mazzeo, S. E. (2015). Motivational interviewing targeting diet and physical activity improves

- adherence to paediatric obesity treatment: Results from the MI values randomized controlled trial. *Pediatric Obesity*, 10, 118–125. doi:10.1111/j.2047-6310.2014.226.x
- Bishop, C. M. (2007). *Pattern recognition and machine learning*. New York: Springer.
- Borrelli, B., Tooley, E. M., & Scott-Sheldon, L. A. J. (2015). Motivational interviewing for parent-child health interventions: A systematic review and meta-analysis. *Pediatric Dentistry*, 37, 254–265.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- Can, D., Marín, R. A., Georgiou, P. G., Imel, Z. E., Atkins, D. C., & Narayanan, S. S. (2016). 'It sounds like...': A natural language processing approach to detecting counselor reflections in motivational interviewing. *Journal of Counseling Psychology*, 63, 343–350. doi:10.1037/cou0000111
- Carcone, A. I., Naar-King, S., Brogan, K. E., Albrecht, T., Barton, E., Foster, T., ... Marshall, S. (2013). Provider communication behaviors that predict motivation to change in black adolescents with obesity. *Journal of Developmental and Behavioral Pediatrics*, 34, 599–608. doi:10.1097/DBP.0b013e3182a67daf
- Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *The New England Journal of Medicine*, 372, 793–795.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.
- Cushing, C. C., Jensen, C. D., Miller, M. B., & Leffingwell, T. R. (2014). Meta-analysis of motivational interviewing for adolescent health behavior: Efficacy beyond substance use. *Journal of Consulting and Clinical Psychology*, 82, 1212–1218. doi:10.1037/a0036912
- Deshpande, G., Li, Z., Santhanam, P., Coles, C., Lynch, M., Hamann, S., & Hu, X. (2010). Recursive cluster elimination based support vector machine for disease state prediction using resting state functional and effective brain connectivity. *PloS one*, 5(12), e14277. doi: 10.1371/journal.pone.0014277
- dos Santos, C., & Gatti, M. (2014). *Deep convolutional neural networks for sentiment analysis of short texts*. Paper presented at the 25th International Conference on Computational Linguistics (COLING 2014), Dublin, Ireland.
- Durgesh, K. S., & Lekha, B. (2010). Data classification using support vector machine. *Journal of Theoretical and Applied Information Technology*, 12, 1–7.
- Ellis, D. A., Carcone, A. I., Ondersma, S. J., Naar-King, S., Dekelbab, B., & Moltz, K. (2017). Brief computer-delivered intervention to increase parental monitoring in families of African American adolescents with type 1 diabetes: A randomized controlled trial. *Telemedicine and e-Health*, 23, 493–502. doi:10.1089/tmj.2016.0182
- Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society for Artificial Intelligence*, 14, 1612.
- Gaume, J., Gmel, G., Faouzi, M., & Daeppen, J. B. (2008). Counselor behaviors and patient language during brief motivational interventions: A sequential analysis of speech. *Addiction*, 103, 1793–1800.
- Gayes, L. A., & Steele, R. G. (2014). A meta-analysis of motivational interviewing interventions for pediatric health behavior change. *Journal of Consulting and Clinical Psychology*, 82, 521–535.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11, 10–18.
- Hasan, M., Carcone, A. I., Naar, S., Eggly, S., Alexander, G. L., Hartlieb, K. E. B., & Kotov, A. (2018). Identifying effective motivational interviewing communication sequences using automated pattern analysis. *Journal of Health Informatics*, p.1–21, https://doi.org/10.1007/s41666-018-0037-6.
- Hasan, M., Kotov, A., Carcone, A. I., Dong, M., & Naar, S. (2018). Predicting the outcome of patient-provider communication sequences using recurrent neural networks and probabilistic models. *AMIA Summits on Translational Science Proceedings*, 2017, 64.
- Hasan, M., Kotov, A., Naar, S., Alexander, G. L., & Idalski Carcone, A. (2019). *Deep neural architectures for discourse segmentation in e-mail based behavioral interventions*. Paper presented at the American Medical Informatics Association (AMIA) 2019 Informatics Summit, San Francisco, CA.
- Idalski Carcone, A., Kotov, A., Hasan, M., Dong, M., Eggly, S., Brogan Hartlieb, K. E., ... Lu, M. (2018). *Using natural language processing to understand the antecedents of behavior change*. Paper presented at the 39th Annual Meeting & Scientific Sessions of the Society of Behavioral Medicine, New Orleans, LA.
- Idalski Carcone, A., & Naar, S. (2017). *Provider behaviors that predict change talk in HIV: A study of communication using the motivational interviewing framework*. Paper presented at the Fifth International Conference on Motivational Interviewing (ICMI-5), Philadelphia, PA.
- Kibriya, A. M., Frank, E., Pfahringer, B., & Holmes, G. (2004). *Multinomial naive bayes for text categorization revisited*. Paper presented at the Australasian Joint Conference on Artificial Intelligence, Berlin, Heidelberg.
- Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv Preprint arXiv*, 1408.5882.
- Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection*. Paper presented at the International Joint Conference on Artificial Intelligence, Montreal, Canada.
- Kotov, A., Hasan, M., Carcone, A., Dong, M., Naar-King, S., & BroganHartlieb, K. (2015). *Interpretable probabilistic latent variable models for automatic annotation of clinical text*. Paper presented at the AMIA Annual Symposium Proceedings, San Francisco, CA.
- Kotov, A., Idalski Carcone, A., Dong, M., Naar-King, S., & Brogan, K. E. (2014). *Towards automatic coding of interview transcripts for public health research*. Paper presented at the Proceedings of the Big Data Analytic Technology For Bioinformatics and Health Informatics Workshop (KDD-BHI) in conjunction with ACM SIGKDD

- Conference on Knowledge Discovery and Data Mining, New York, NY.
- Lacoste-Julien, S., Sha, F., & Jordan, M. I. (2008). *DiscLDA: Discriminative learning for dimensionality reduction and classification*. Paper presented at the Neural Information Processing Systems 2008, Whistler, British Columbia, Canada.
- Lafferty, J., McCallum, A., & Pereira, F. C. (2001). *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. Paper presented at the Eighteenth International Conference on Machine Learning (ICML 2001), Williamstown, MA.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.
- MacDonell, K. K., Naar, S., Gibson-Scipio, W., Lam, P., & Secord, E. (2016). The detroit young adult asthma project: Pilot of a technology-based medication adherence intervention for African-American emerging adults. *Journal of Adolescent Health*, 59, 465–471.
- McCallum, A., & Nigam, K. (1998). *A comparison of event models for naive bayes text classification*. Paper presented at the ICML/AAAI-98 workshop on learning for text categorization, Madison, Wisconsin.
- Messinger, D. M., Ruvolo, P., Ekas, N. V., & Fogel, A. (2010). Applying machine learning to infant interaction: The development is in the details. *Neural Networks*, 23, 1004–1016. doi:10.1016/j.neunet.2010.08.008
- Miller, W. R., Moyers, T. B., Ernst, D. B., & Amrhein, P. C. (2008). *Manual for the Motivational Interviewing Skill Code (MISC), Version 2.1*. New Mexico: Center on Alcoholism, Substance Abuse, and Addictions, The University of New Mexico.
- Miller, W. R., & Rollnick, S. (2012). *Motivational interviewing: Helping people change* (3rd edn). New York, NY: The Guilford Press.
- Miller, W. R., & Rose, G. S. (2009). Toward a theory of motivational interviewing. *American Psychologist; American Psychologist*, 64, 527.
- Moyers, T. B., & Martin, T. (2006). Therapist influence on client language during motivational interviewing sessions. *Journal of Substance Abuse Treatment*, 30, 245–251. doi:10.1016/j.jsat.2005.12.003
- Moyers, T. B., Martin, T., Manuel, J. K., Miller, W. R., & Ernst, D. (2010). 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).
- Naar-King, S., Outlaw, A. Y., Sarr, M., Parsons, J. T., Belzer, M., MacDonell, K., ... Ondersma, S. J. (2013). Motivational enhancement system for adherence (MESA): Pilot randomized trial of a brief computer-delivered prevention intervention for youth initiating antiretroviral treatment. *Journal of Pediatric Psychology*, 38, 638–648. doi:10.1093/jpepsy/jss132
- Pérez-Rosas, V., Mihalcea, R., Resnicow, K., Singh, S., Ann, L., Goggin, K. J., & Catley, D. (2017). *Predicting Counselor Behaviors in Motivational Interviewing Encounters*. Paper presented at the Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Valencia, Spain.
- Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation.
- Rajkumar, D., Ellis, D. A., May, D. K., Carcone, A., Naar-King, S., Ondersma, S., & Moltz, K. C. (2015). Computerized intervention to increase motivation for diabetes self-management in adolescents with type 1 diabetes. *Health Psychology and Behavioral Medicine*, 3, 236–250.
- Rish, I. (2001). *An empirical study of the naive Bayes classifier*. Paper presented at the IJCAI 2001 workshop on empirical methods in artificial intelligence.
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55, 68.
- Schaefer, M. R., & Kavookjian, J. (2017). The impact of motivational interviewing on adherence and symptom severity in adolescents and young adults with chronic illness: A systematic review. *Patient Education and Counseling*, 100, 2190–2199. doi:10.1016/j.pec.2017.05.037
- Sharma, A. K., & Sahni, S. (2011). A comparative study of classification algorithms for spam email data analysis. *International Journal on Computer Science and Engineering*, 3, 1890–1895.
- Sutton, C., & McCallum, A. (2006). *An introduction to conditional random fields for relational learning* (Vol. 2): *Introduction to statistical relational learning*. Cambridge, MA: MIT Press.
- Tanana, M., Hallgren, K., Imel, Z., Atkins, D., Smyth, P., & Srikumar, V. (2015). *Recursive Neural Networks for Coding Therapist and Patient Behavior in Motivational Interviewing*. Paper presented at the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Denver, Colorado.
- Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., & Srikumar, V. (2016). A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of Substance Abuse Treatment*, 65, 43–50. doi:10.1016/j.jsat.2016.01.006
- Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29, 24–54.
- Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2, 45–66.
- Wang, S., & Manning, C. D. (2012). *Baselines and bigrams: Simple, good sentiment and topic classification*. Paper presented at the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, Jeju Island, Korea.
- Wood, J. M. (2007). Understanding and computing Cohen's Kappa: A tutorial. *WebPsychEmpiricist. Web Journal*. Retrieved from <http://wpe.info/>