Journal of Biomedical Informatics 62 (2016) 21–31

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories

Mehedi Hasan [a,1], Alexander Kotov [a,1,*], April Idalski Carcone [b], Ming Dong [a], Sylvie Naar [b], Kathryn Brogan Hartlieb [c]

[a] Department of Computer Science, Wayne State University, 5057 Woodward Ave, Detroit, MI 48202, USA
[b] Pediatric Prevention Research Center, School of Medicine, Wayne State University, 540 E Canfield St, Detroit, MI 48201, USA
[c] Department of Dietetics and Nutrition, Florida International University, 11200 SW 8th St, Miami, FL 33199, USA

ARTICLE INFO

Article history:
Received 16 January 2016
Revised 12 May 2016
Accepted 12 May 2016
Available online 13 May 2016

Keywords:
Machine learning
Deep learning
Text classification
Annotation of clinical text
Motivational interviewing

ABSTRACT

This study examines the effectiveness of state-of-the-art supervised machine learning methods in conjunction with different feature types for the task of automatic annotation of fragments of clinical text based on codebooks with a large number of categories. We used a collection of motivational interview transcripts consisting of 11,353 utterances, which were manually annotated by two human coders as the gold standard, and experimented with state-of-art classifiers, including Naïve Bayes, J48 Decision Tree, Support Vector Machine (SVM), Random Forest (RF), AdaBoost, DiscLDA, Conditional Random Fields (CRF) and Convolutional Neural Network (CNN) in conjunction with lexical, contextual (label of the previous utterance) and semantic (distribution of words in the utterance across the Linguistic Inquiry and Word Count dictionaries) features. We found out that, when the number of classes is large, the performance of CNN and CRF is inferior to SVM. When only lexical features were used, interview transcripts were automatically annotated by SVM with the highest classification accuracy among all classifiers of 70.8%, 61% and 53.7% based on the codebooks consisting of 17, 20 and 41 codes, respectively. Using contextual and semantic features, as well as their combination, in addition to lexical ones, improved the accuracy of SVM for annotation of utterances in motivational interview transcripts with a codebook consisting of 17 classes to 71.5%, 74.2%, and 75.1%, respectively. Our results demonstrate the potential of using machine learning methods in conjunction with lexical, semantic and contextual features for automatic annotation of clinical interview transcripts with near-human accuracy.

1. Introduction

Annotation (or labeling) of fragments of clinical text with the codes from a predefined codebook is an integral part of qualitative research. It can also be viewed as a classification of textual fragments into a predefined number of classes (or categories). It has been traditionally performed manually by trained coders, which is a tedious, costly and time-consuming process. Furthermore, manual annotation increases the likelihood of errors due to coder fatigue and bias associated with human subjectivity. To automate tedious cognitive tasks, such as classification, supervised machine learning methods (including interpretable ones [1]) have been recently proposed. Although these methods have been shown to be successful at binary (two-class) classification [2,3] (e.g. classifying textual fragments as neutral or opinionated), we are not aware of any prior studies that evaluate their performance for textual classification tasks involving a large number of classes. Such tasks, however, are fairly common in a clinical setting (e.g. annotation of clinical interviews, assignment of ICD-9/10 codes to patient records). To address this limitation, in this paper, we propose contextual and semantic features and present the results of an extensive experimental evaluation of state-of-the-art supervised machine learning methods in conjunction with lexical and the proposed features for the task of automatic annotation of fragments in the clinical interview transcripts with the codebooks consisting of a large number of classes. This study provides a guideline for clinical informatics researchers and practitioners, who consider

* Corresponding author at: Department of Computer Science, Wayne State University, 5057 Woodward Ave Suite 14001.6, Detroit, MI 48202, USA.
E-mail addresses: mehedi@wayne.edu (M. Hasan), kotov@wayne.edu (A. Kotov).
URL: http://www.cs.wayne.edu/kotov/ (A. Kotov).
[1] The first two authors provided equal contribution.

http://dx.doi.org/10.1016/j.jbi.2016.05.004
1532-0464/

an option of using machine learning methods for automatic annotation of clinical text in their projects.

Annotation of clinical interview transcripts to distinguish different patient behavior types is an important part of clinical research aimed at designing effective interventions for many conditions and disorders. In this paper, we focus on the transcripts of Motivational Interviews with obese adolescents (teens) and their caregivers. Childhood obesity is a serious public health concern in the United States and worldwide. Recent estimates indicate that approximately one-third (31.8%) of US children and adolescents between ages 2 and 19 are overweight and 16.9% are obese [4]. Adolescents who are obese are likely to continue to be obese in adulthood and have a greater risk of heart disease, type 2 diabetes, stroke, cancer, and osteoarthritis [5]. Therefore, there is a need for informatics-based methods to facilitate the development of effective interventions for childhood obesity. One such intervention is Motivational Interviewing (MI), an evidence-based counseling technique to increase intrinsic motivation and self-efficacy for health-related behavior change [6,7]. The goal of a MI counseling session is to encourage patients to explore their own desires, ability, reasons, a need for and commitment to the targeted behavior change. These statements, referred to as "change talk" (or CT), consistently predict actual behavior change [8] that can be sustained for as long as 34 months after an interview [9]. The process of establishing causal linkages to identify effective communication strategies for eliciting change talk and commitment language in MI involves a resource-intensive qualitative coding process. First, clinical interviews are transcribed and then each utterance is manually annotated with a set of codes from a pre-defined codebook designating specific behavior types. Training human coders to reliably and accurately assign codes to textual fragments requires a large investment of manpower, time and money. For example, in a recent MI study [10], training coders to reliability took about four months and, once trained, coders required five hours to code every recorded hour. A similar study reported requiring 60 h of training over six weeks to attain coder reliability, and the actual coding involved two coding passes and six coders [11].

Automatic annotation of patient utterances in clinical communication is a challenging task since patients usually come from a variety of cultural and educational backgrounds and their language use can be quite different [12]. This problem is exacerbated when the interviews are conducted with children and adolescents due to their tendency to use incomplete sentences and frequently change subjects.

Previous quantitative studies of clinical conversation have resulted in creation of Generalized Medical Interaction Analysis System (GMIAS) [13], which uses a codebook with generic hierarchical categories. The small-size codebook in Comprehensive Analysis of the Structure of Encounters System (CASES) [14] was designed to annotate several meta-discursive aspects of medical interviews, such as assigning "ownership" of topics and partitioning them into distinct segments (speech acts). It was also shown that the fragments of transcripts of routine outpatient visits consisting of several speech acts coded using GMIAS and CASES can be annotated as "information giving" and "requesting information" [15]. Other related previous studies focused on categorizing assertions of medical problems in clinical narrative into 5 classes (present, absent, possible, hypothetical, conditional and associated with someone else) using SVM [16] and annotating the utterances in hemodialysis phone dialog with 3 categories using AdaBoost classifier [17]. The present study reports the results of a comprehensive evaluation of 8 state-of-the-art classifiers (Naïve Bayes [18–20], Support Vector Machine [21,22], Conditional Random Fields [23,24], J48 [25], AdaBoost [26], Random Forest [27], DiscLDA [28] and Convolutional Neural Network [29]) for the task of annotating clinical interviews with a codebook, consisting of a large number of classes. We also propose and experimentally evaluate two novel features for this task: contextual features based on the label of the preceding textual fragment and semantic features based on the distribution of words in the annotated fragment over a psycho-linguistic lexicon.

2. Materials and methods

2.1. Data collection and preprocessing

The golden standard for the evaluation of machine learning methods was created based on the transcripts of motivational interviews conducted by the clinicians at the Pediatric Prevention Research Center (PPRC) of Wayne State University. Each interview is comprised of two parts: conversation of a clinician with an adolescent followed by a conversation of a clinician with the adolescent's caregiver. All adolescents in this study were between the ages of 12 and 17 ($M = 14.7$, SD = 1.63) and most were female ($n = 27$). Most caregivers were biological mothers ($n = 36$), who were married or living with a partner ($n = 25$). The median family income was $16,000–$21,999 ranging from less than $1000 to $5 0,000–$74,999. Audio recordings of the interviews were first transcribed and segmented into utterances belonging to adolescents, caregivers, and counselors, preserving the sequence of utterances. Transcripts were then manually annotated by trained human coders according to MYSCOPE [10], a specialized codebook including a large number of behavior codes, which was developed by an interdisciplinary team including a clinical psychologist, a nutrition scientist, a communication scientist, a linguist and a community health worker specifically for annotating motivational interviews with obese adolescents. A primary coder independently coded interview sessions and a secondary coder co-coded a randomly selected 20% of the transcripts to monitor reliability ($\kappa = 0.696$) [10]. The MYSCOPE codebook contains a total of 115 different codes that are grouped into the youth, caregiver, and counselor code groups. The experimental datasets for this work were constructed based on the transcripts of 37 motivational interview sessions, which include a total of 11,353 segmented and annotated utterances. These utterances have been further partitioned into two subsets based on the structure of motivational interview sessions: one dataset that includes all utterances from the adolescent sessions (6579 samples) and the other dataset that includes all utterances from the caregiver sessions (4774 samples). A fragment of an adolescent session transcript is presented in Table 1.

To conduct a detailed performance analysis of classification methods, we used the following two-stage process to create codebooks with different number of codes for adolescent and caregiver sessions. In the first stage, we merged conceptually similar behavior codes as well as the codes with similar data distributions, while in the second stage, we eliminated the codes with insufficient data samples. In the case of the adolescent sessions, we started with 55 adolescent session-specific codes and, after merging the codes with subtle differences (e.g. converting variances of change talk CHT+1, CHT+2 and CHT+3 into CHT+), obtained a codebook with 41 classes. We further reduced this codebook to 20 classes after merging 21 classes with similar sample distributions. After eliminating the codes that had less than 10 data samples (to ensure that there can be at least one sample of each class in each fold when using 10-fold cross-validation experimental design), we obtained a third codebook with 17 codes. Using the same approach, we created codebooks containing 58, 19, and 16 caregiver session-specific codes. Table 2 shows the distribution of utterances over 16 classes in the caregiver session transcripts. As follows from Table 2, the distribution of utterances over classes is highly imbalanced even for the codebooks of the smallest size, which is fairly common for clinical text.

Table 1
Fragment of the annotated transcript of a dialog between a counselor and an adolescent.

Annotation	Description	Speaker	Text
331	Open-ended question, elicit change talk positive	Counselor	"do you feel like making healthier choices for your snacks and your meals is something you would be able to do? mm-hmm meaning is that food available for you?"
117	Low uptake, positive	Adolescent	"Yes"
301	Structure session	Counselor	"okay and that's an important thing for us to think about cause I would not want to help you come up with a plan that you would not be able to do without somebody else help so the last part of your plan is how somebody could be supportive to you meaning how they can help you be successful and so we should choose somebody who you feel like is around often enough"
112	Change talk positive	Adolescent	"my um aunt"
301	Structure session	Counselor	"okay so let's stick something my aunt can do"
112	Change talk positive	Adolescent	"she could when I am doing when I am eating something that I should I could not be eating but so I can choose something healthy she could tell me not to eat it"
309	Affirm, low	Counselor	"okay that sounds like a really great suggestion"

Table 2
Distribution of utterances over 16 classes in the caregiver dataset.

Code	Description	Utterance	%
209	Caregiver change talk, negative	297	6.82
212	Caregiver change talk, positive	1107	25.40
232	Low uptake, positive	518	11.89
235	High uptake	231	5.30
301	Structure session	206	4.73
302	General information, positive	309	7.09
305	Emphasize autonomy	148	3.40
306	Closed question, elicit feedback	50	1.15
307	Support	108	2.48
308	Affirm	289	6.63
315	Reflect, change talk positive, about caregiver	659	15.12
329	Self-disclose	44	1.01
330	Statement, other	121	2.78
331	Open-ended question, elicit change talk positive	200	4.59
343	Open-ended question, target behavior neutral	33	0.76
344	Open-ended question, elicit barriers	38	0.87

After creating the codebooks, we pre-processed the dataset using the Snowball stemmer available as part of the Weka [30] machine learning toolkit.[2] We observed that stopword removal decreased the performance of classification models for our task (e.g., in the case of the codebook consisting of 17 classes, the accuracy of Naïve Bayes decreased from 67% to 47.10%, while the accuracy of SVM decreased from 70.76% to 55.26%). A likely reason is that, although negations are typically considered as stopwords, they are fairly important clues for inferring certain behavior types (e.g., removing the stopword "not" completely transforms the meaning of a phrase "not great").

[2] http://www.cs.waikato.ac.nz/ml/weka/.

2.2. Feature set

The different feature types used in our experiments are summarized in Table 3, while Fig. 1 illustrates the process of extracting these features from a sample interview fragment. First, we compared the performance of all classification models using only lexical features, which were derived from a unigram bag-of-words representation of utterances. According to this approach, a set of unique terms (vocabulary) of size N is first determined for a given collection of textual fragments (in our case, interview transcripts) and then each textual fragment f (in our case, adolescent or caregiver utterance) is represented as a feature vector $[nw_{1f}, \ldots, nw_{Nf}]$, where nw_{nf} is a feature representing the number of times an nth word from the collection vocabulary occurred in f. For example, the vocabulary of a collection consisting of only one textual fragment "what you think about your weight right now and your health" would be ("about", "and", "health", "now", "right", "think", "weight", "what", "you", "your") and the unigram bag-of-words feature vector for this fragment based on the representation would be [1,1,1,1,1,1,1,1,1,2]. Since the question mark (?) is an important indicator of some communication types, it was also used as a feature.

Second, we expanded lexical features with the features derived from Linguistic Inquiry and Word Count (LIWC) lexicon [31]. LIWC lexicon consists of dictionaries, which had been manually compiled and validated for over a decade by psychologists, sociologists and linguists. Dictionaries are organized around sixty-eight psychological and social dimensions, which are structured as an ontology-like hierarchy and may overlap. Each dictionary corresponds to a well-defined concept or psychological construct (e.g. social, positive emotions, negative emotions, money). Social dictionary consists of the nouns and pronouns that refer to other people (e.g., "they", "she", "us", "friends") as well as the verbs that indicate interaction (e.g., "talking", "sharing"). Dictionaries of positive (e.g. "happy", "love", "good") and negative (e.g. "sad", "kill", "afraid") words cover the entire spectrum of corresponding emotions from happiness to anxiety. We use the vector of counts of terms in the utterance across LIWC dictionaries as additional semantic features. For example, the sentence "what you think about your weight right now and your health" is represented as a vector [2,. . .,1,. . .,3,. . .,1,. . .,1,. . .,1], in which each element is the number of counts of words that fall under each of the sixty-eight categories [cognitive process,. . .,pronoun,. . .,time,. . .,inclusive,. . .,physical states,. . .,preposition]. LIWC has been applied to successfully predict the onset of depression [32] and characterize postpartum emotional variability based on social media posts by an individual [33]. In the case of annotation of clinical interview transcripts, LIWC features provide important lexical clues related to thought processes, emotional states, intentions, and motivations of patients.

Finally, in addition to lexical features, we also considered the context of interview utterances in the form of the label of the preceding utterance. We hypothesize that contextual features of an utterance play an important role during the annotation process since motivational interviews proceed in a sequential manner with participants asking or responding to questions of the previous speaker. Therefore, we use the automatically assigned category of the preceding counselor utterance as an additional contextual feature when annotating adolescent or caregiver session transcripts, and vice versa. For example, if the set of codes specific to the counselor utterances is [109,. . .,120,. . .,305,. . .,311,. . .,331,. . .,343,. . .,344], then the additional contextual feature vector for the

Table 3
Feature representation of each utterance in machine learning pipeline.

Feature type	Description	Purpose
Lexical features	One feature per each distinct word in the set of training interview transcripts. The value of each lexical feature is the number of times that the corresponding word appears in the utterance	To capture the vocabulary that is indicative of each label
Contextual features	One feature per each codebook label. The value of the feature is set to 1 if the previous utterance in the dialog was annotated with the corresponding label, and to 0, otherwise	Context changes the likelihood of observing speech acts. For example, if the previous speaker was requesting information, then the next speech act is more likely to be providing the requested information
Semantic features	One feature per each of the sixty-eight LIWC lexicons. The value of each semantic feature is the number of times a word from the corresponding dictionary appears in the utterance	To capture psycho-linguistic clues related to the thought processes, emotional states, intentions and motivations of the speaker

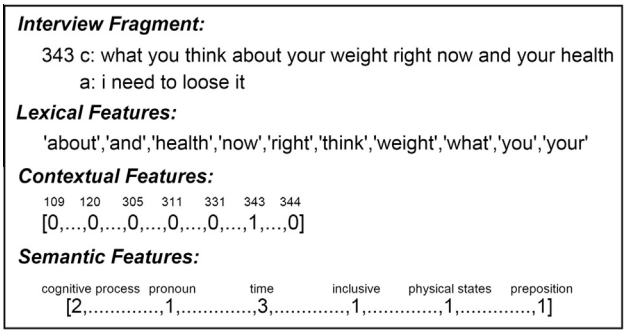

Fig. 1. Features extracted from a sample interview fragment.

adolescent utterance "I need to lose it", which is preceded by the counselor utterance annotated with the code 343, is [0,. . .,0,. . .,0,. . .,0,. . .,1,. . .,0].

2.3. Classification models

We first describe a general architecture of the classification system used in our experiments and then provide a brief overview of each evaluated machine learning method. Fig. 2 shows the architecture of the pipeline used for the classification of medical interview transcripts.

The pipeline consists of two stages: training and testing. Prior to the training stage, we preprocess the collected clinical interview transcripts by performing stemming, punctuation removal, word segmentation and tokenization. Features are then extracted from the preprocessed data. During this stage, previous label and LIWC features are used in conjunction with the lexical features to create feature vectors. After that, classifiers are trained on the feature vectors extracted from the training samples and their associated annotations. In the testing stage, after creation of feature vectors, the previously trained classifiers predict the label of each utterance in the testing sample. Finally, performance of different classifiers is evaluated by calculating standard metrics such as precision, recall, F-score (F1), kappa measure and accuracy. Specifically, we evaluated the performance of the following state-of-the-art supervised machine learning methods.

2.3.1. Naïve Bayes (NB)

Naïve Bayes (NB) [18–20] is a popular probabilistic method [34,35] for text classification due to its robustness and relative simplicity. Experimental results reported in this paper were obtained using standard implementations of binomial Naïve Bayes (NB) and multinomial Naïve Bayes (NB-M) algorithms [19] provided by the Weka toolkit.

2.3.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) [21,22] belongs to a family of generalized linear binary classifiers, which map an input feature vector into a higher dimensional space and find a hyperplane that separates samples into two classes in such a way that the margin between the closest samples in each class is maximized. An open-source implementation of SVM with different kernels in a publicly available LibSVM[3] package [36] was used for the experiments reported in this work. The parameters of each kernel have been empirically optimized using cross-validation. Figs. 3–5 illustrate the variance in accuracy of SVM with different settings of parameters for RBF, polynomial and sigmoid kernels, respectively. As follows from Fig. 4, if the number of classes is large, SVM has optimal performance when a quadratic polynomial kernel is used or when γ is set to 0.1 for a sigmoid kernel. The best performance of SVM among all kernels, however, is achieved when it is used with a Radial Basis Function (RBF) kernel with the parameters C and γ set to 4.0 and 0.1, respectively. We also observed that L1 loss function performs better than L2 loss function for Linear SVM.

2.3.3. Conditional Random Fields (CRF)

Conditional Random Fields (CRF) [23,24] is a probabilistic model, which is different from the rest of the classifiers in that, in addition to lexical features, it also considers the dependencies between the labels of consecutive data samples. We used linear chain CRF provided by MALLET [37], a publicly available machine learning toolkit.[4]

2.3.4. Decision tree (J48)

J48 [25] is an open source implementation of the C4.5 decision tree classification algorithm provided by Weka. Decision trees are interpretable classifiers, which model the classification process as a tree traversal.

2.3.5. AdaBoost

AdaBoost [26] (short for "Adaptive Boosting") is one of the most widely used and studied machine learning meta-algorithms. Boosting algorithms belong to a group of voting techniques [38], which produce classification decision as a linear combination of the output of other classifiers (also called "base" or "weak" learners) [39]. In particular, we used J48 decision tree classifier as a weak learner for AdaBoost.

2.3.6. Random Forest (RF)

Random Forest [27] is an ensemble method that uses bagging to improve classification performance by combining the output of several classifiers. The main idea behind ensemble methods is that a large number of "weak learners" can be used to create a "strong learner". In the case of Random Forest, a "weak learner" is a decision tree. Fig. 6 illustrates the performance of Random Forest by varying the number of individual decision trees. From Fig. 6, it follows that increasing the number of trees beyond 150 results in minor improvement of accuracy. We used 300 trees for RF, which

[3] https://www.csie.ntu.edu.tw/cjlin/libsvm/.
[4] http://mallet.cs.umass.edu/.

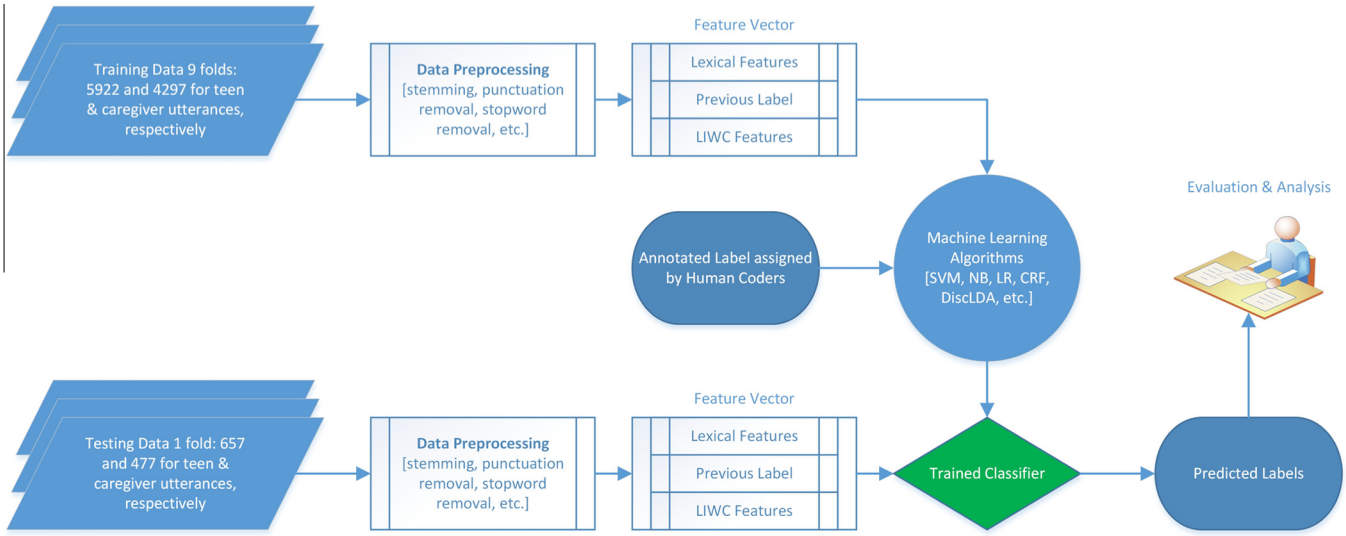

Fig. 2. Architecture of the pipeline for automatic annotation of clinical interview fragments using various supervised machine learning methods.

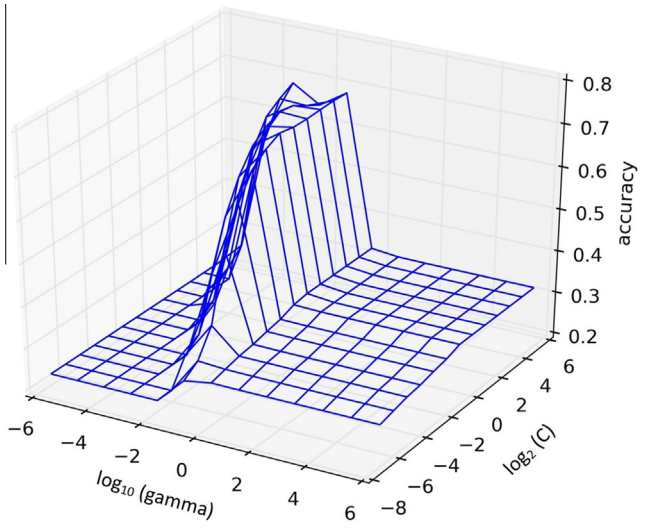

Fig. 3. Accuracy of SVM with RBF kernel by varying kernel parameters C and γ.

we empirically determined to result in the best performance of this classifier for the codebooks of different size.

2.3.7. DiscLDA

DiscLDA [28] is a dimensionality reduction method that incorporates supervision in the form of class labels into Latent Dirichlet Allocation (LDA) [40] to uncover the latent structure in document collections and leverage this structure to improve the accuracy of classification. Experimental results reported in this paper were obtained by setting α to $50/T$ [41], where T is a number of topics, and β to 0.1 and running the model for 150 iterations. Fig. 7 shows the performance of DiscLDA depending on the specified number of topics. From Fig. 7, it follows that DiscLDA achieves the highest accuracy when the number of topics is 250.

2.3.8. Convolutional neural network

Deep Learning exploits the idea of a hierarchy of explanatory factors, in which higher level, more abstract concepts are learned from those at the lower levels. Deep learning helps to disentangle these abstractions and select features that are useful for classification. These architectures are often constructed using a greedy layer-by-layer method. For supervised learning tasks, rather than

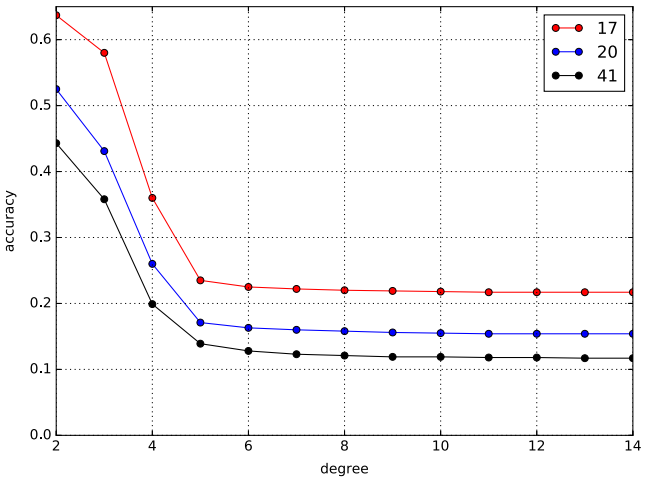

Fig. 4. Accuracy of SVM with polynomial kernel by varying the degree.

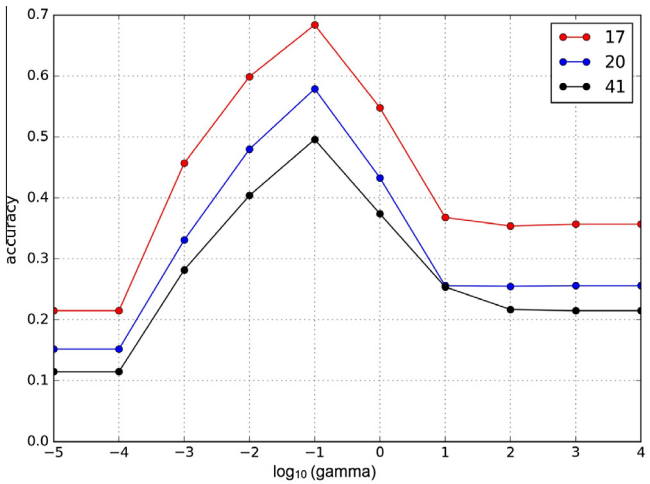

Fig. 5. Accuracy of SVM with sigmoid kernel by varying γ.

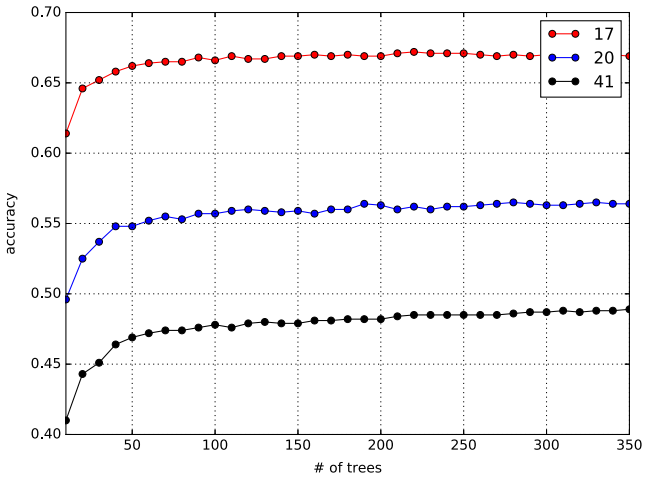

Fig. 6. Accuracy of Random Forest by varying the number of decision trees.

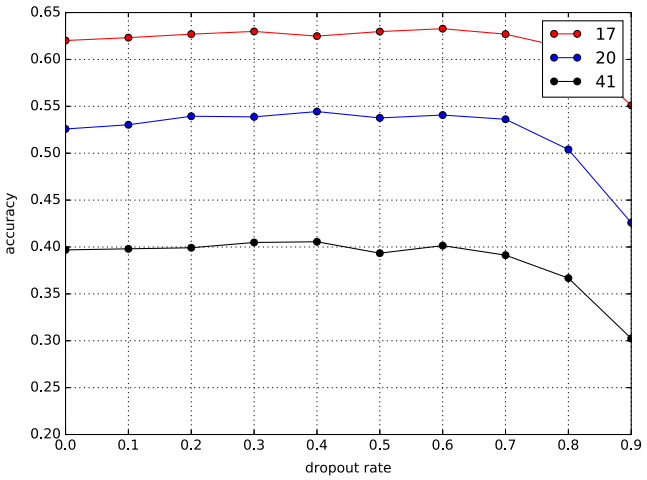

Fig. 8. Accuracy of CNN by varying the dropout rate.

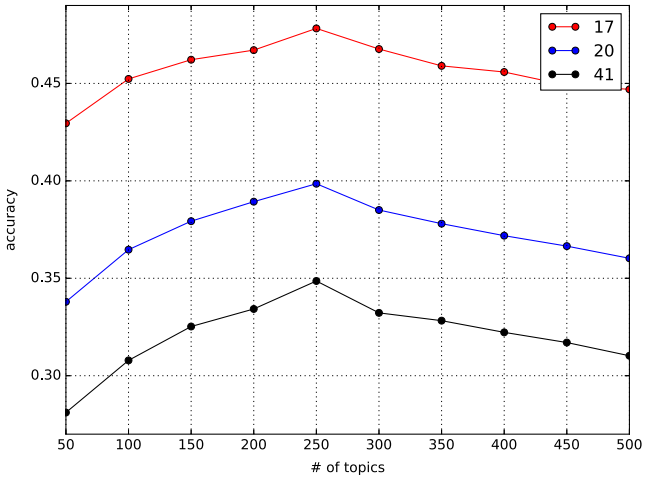

Fig. 7. Accuracy of DiscLDA by varying the number of topics.

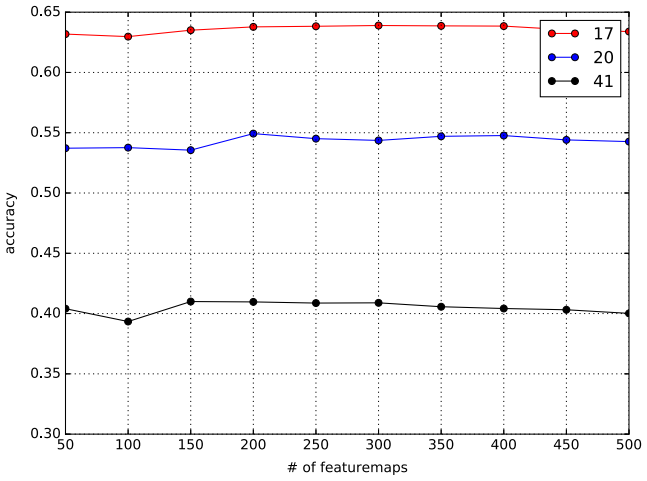

Fig. 9. Accuracy of CNN by varying the number of featuremaps.

extracting manually designed features from the data, deep learning methods translate the data into a compact intermediate representation, similar to other dimensionality reduction techniques, and derive layered structures, which eliminate redundancy in feature representation. We used a convolutional neural network (CNN) with one layer of convolution [29] on top of the latent dimensional representation of each word in an interview fragment using the publicly available WORD2VEC vectors, which were obtained from an unsupervised neural language model [42] estimated on 100 billion word corpus from Google News. If a WORD2VEC vector was not available for a particular word, we applied random initialization for its latent dimensional representation. In the architecture of this CNN, shown in Fig. 10, an interview fragment consisting of n words is represented by n 300-dimensional WORD2VEC vectors, which were fine-tuned for our dataset through backpropagation. A convolution operation with multiple filters corresponding to the windows of size 3, 4 and 5 words was then applied to produce new features. After that, a max-over-time pooling [43] is used to capture the most important feature for each particular filter. These features form the penultimate layer and are then passed to a fully connected softmax layer whose output is a probability distribution over category assignments for a given interview fragment. Based on empirical analysis in [44], we tuned two important parameters to improve the performance of CNN: dropout rate and a number of featuremaps. The effect of the dropout rate and the number of featuremaps on the performance of CNN is shown in Figs. 8 and 9, respectively. As follows from Fig. 9, the number of faturemaps does not have a significant effect on the performance of CNN when the number of classes is large.

2.4. Evaluation

To ensure the robustness of performance estimates, we used 10-fold cross-validation [45] as an experimental design. The performance of different classifiers and feature sets was evaluated in terms of precision, recall, F1 score (F1), kappa measure and accuracy using weighted macro-averaging over 10 folds.

3. Results

Experimental evaluation of automatic annotation using machine learning methods in this work spanned several dimensions:

- determining the performance of classifiers on codebooks of different size;
- determining the effectiveness of the proposed contextual and semantic features.

Since clinical researchers typically annotate caregiver and adolescent sessions separately, we first created two experimental

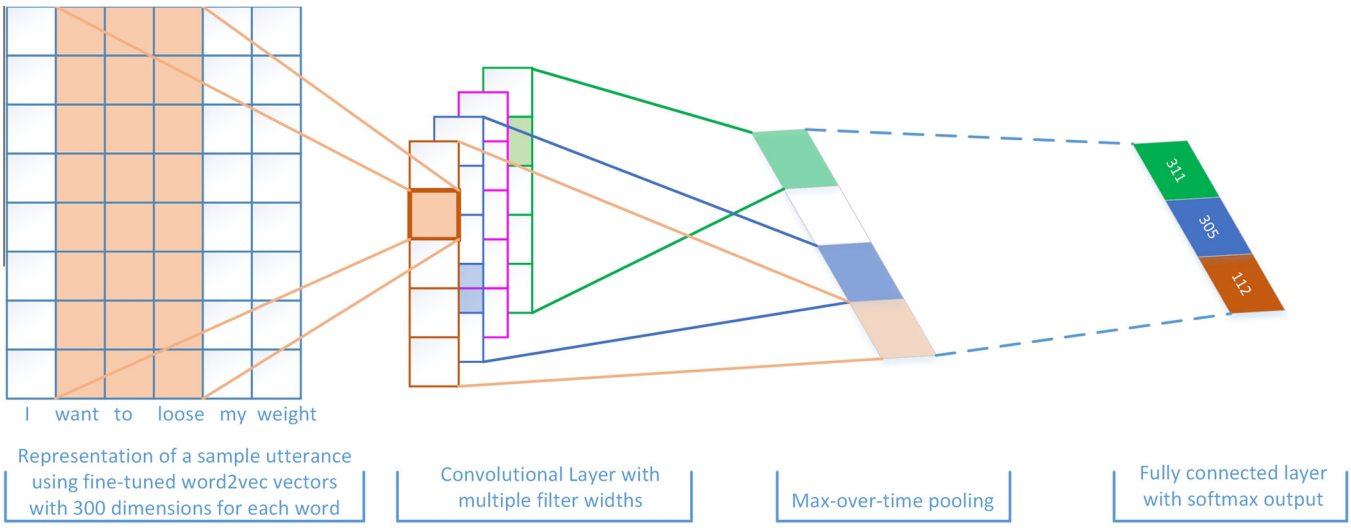

Fig. 10. Architecture of convolutional neural network for automatic annotation of clinical interview transcripts.

datasets consisting of only adolescent and only caregiver session transcripts. Second, besides evaluating the accuracy of annotating adolescent and caregiver transcripts with codebooks containing an entire set of codes, we also conducted a series of experiments with the codebooks of smaller size created as outlined above. Third, besides training and testing NB, SVM, CRF, Decision Tree, AdaBoost, DiscLDA, Random Forest and CNN classifiers using only lexical features, we also evaluated the effectiveness of the proposed contextual and semantic features.

3.1. Quality of automatic annotation using only lexical features

Standard performance metrics[5] of different classification models using only lexical features for the task of annotating adolescent and caregiver session transcripts are summarized in Tables 4 and 5, respectively.

Several observations can be made based on Tables 4 and 5. First, SVM consistently demonstrates the best performance while DiscLDA and J48 consistently have the worst performance in terms of all metrics and for codebooks of all sizes on both adolescent and caregiver interview session transcripts. In the case of DiscLDA, this indicates that dimensionality reduction is less effective when the number of classes is large. In the case of J48, this indicates that decision trees are not effective when feature vectors are sparse and high-dimensional as well as when the number of classes is large. Furthermore, the difference in performance between SVM and other classifiers keeps increasing with the number of classes in the codebook. For example, in the case of adolescent interview transcripts, the difference in accuracy between SVM and CNN (the best and the second best methods) is 3% when the codebook with 17 labels is used, 2.4% when the codebook with 20 labels is used and 14.1% when the codebook with 41 labels is used. This indicates superior robustness of SVM compared to other machine learning methods. Second, although boosting with AdaBoost consistently improves the performance of J48 in terms of all metrics and for codebooks of all sizes and on both adolescent and caregiver interview session transcripts, SVM and, in many cases, multinomial NB, outperformed AdaBoost, particularly in the case of codebooks with a large number of codes (41 labels in the case of adolescent and 58 labels in the case of caregiver session-specific codebooks), which indicates that boosting is also less effective for classification tasks involving a large number of classes. Third, CNN outperforms all other classifiers except CRF and SVM, when the codebooks of all sizes except 41 and 58 are used. The differences in accuracy between SVM and CNN are 0.7%, 3%, 3.5%, 2.4%, 14.1% and 33.3% when the codebooks of size 16, 17, 19, 20, 41 and 58 are used, respectively. These results indicate that CNN is less effective for classification problems when the number of classes is large. Fourth, the performance of all classification models is consistently lower on caregiver utterances compared to adolescent utterances, which can be explained by relative simplicity of the language used by adolescents.

3.2. Quality of automatic annotation using lexical and non-lexical features

Summary of performance[6] of CRF and SVM using the combinations of lexical and contextual (SVM-PL), lexical and semantic (SVM-LIWC) and all features (SVM-AF) on adolescent and caregiver session transcripts is provided in Tables 6 and 7, respectively.

Several important conclusions can be made by comparing the experimental results in Tables 6 and 7 with Tables 4 and 5. First, CRF outperformed multinomial NB, achieving 1.2% and 0.2% higher accuracy as well as 3.4% and 2.1% higher F1 score when the codebooks consisting of 17 and 20 classes, respectively, were used to annotate the adolescent transcripts and 2.1% and 0.3% higher accuracy as well as 4.9% and 2.8% higher F1 score when the codebooks consisting of 16 and 19 classes, respectively, were used to annotate the caregiver transcripts. However, CRF has 2% and 2.7% lower accuracy as well as 2% and 2.7% lower F1 score when the codebooks with 41 and 58 classes are used, respectively. On the other hand, the accuracy of CRF is lower than the accuracy of SVM using lexical features by 2.6%, 2.9% and 4.4% when the codebooks of size 17, 20 and 41, respectively, are used to annotate adolescent transcripts and by 7.4%, 0.6% and 1.3% when the codebooks of size 16, 19 and 58, respectively, are used to annotate caregiver transcripts. Nevertheless, since CRF considers both lexical features as well as the labels of previous utterances, these results highlight the importance of accounting for context when annotating the utterances in clinical interview transcripts.

[5] Cls.: # of classes, Acc.: Accuracy, Prec.: Precision, Rec.: Recall.

[6] Cls.: # of classes, Acc.: Accuracy, Prec.: Precision, Rec.: Recall.

Table 4
Performance of classification models using only lexical features according to different evaluation metrics for the task of annotating adolescent interview session transcripts. Highest value for each metric and codebook size across all models is highlighted in boldface.

Cls.	Model	Acc.	Prec.	Rec.	F1	Kappa
	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
17	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
	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
20	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
	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
41	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**

Table 5
Performance of classification models using only lexical features according to different evaluation metrics for the task of annotating caregiver interview session transcripts. Highest value for each metric and codebook size across all models is highlighted in boldface.

Cls.	Model	Acc.	Prec.	Rec.	F1	Kappa
	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
16	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
	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
19	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
	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
58	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**

Second, the performance of SVM improves in terms of all metrics on both adolescent and caregiver datasets and for the codebooks of all sizes when either contextual (SVM-PL) or semantic (SVM-LIWC) features are used in addition to the lexical ones. When both of these features are combined (SVM-AF), the annotation performance of SVM improves even further achieving the best performance in terms of all metrics using the codebooks of all sizes on both adolescent and caregiver transcripts. In particular, by using contextual and semantic features in addition to the lexical ones, the accuracy of SVM improves by 4.3%, 7.2%, and 3.1%, while its F1 score improves by 5.9%, 8.2%, and 4.2%, when the codebooks with 17, 20, and 41 labels, respectively, are used to annotate the adolescent transcripts. When contextual and semantic features are used, the accuracy of SVM improves by 7.4%, 9.3%, and 3.7% and its F1 score improves by 8.8%, 9.6%, and 4.4% when the codebooks with 16, 19, and 58 labels, respectively, are used to annotate the caregiver transcripts.

Table 6
Performance of classification models using a combination of lexical and different types of non-lexical features according to standard metrics for the task of annotating adolescent interview session transcripts. Highest value for each metric and codebook size across all models is highlighted in boldface.

Cls.	Model	Acc.	Prec.	Rec.	F1	Kappa
	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
17	SVM-LIWC	0.742	0.740	0.742	0.727	0.704
	SVM-AF	**0.751**	**0.750**	**0.751**	**0.739**	**0.715**
	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
20	SVM-LIWC	0.653	0.653	0.653	0.657	0.619
	SVM-AF	**0.682**	**0.685**	**0.682**	**0.674**	**0.651**
	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
41	SVM-LIWC	0.538	0.518	0.538	0.507	0.503
	SVM-AF	**0.568**	**0.549**	**0.568**	**0.546**	**0.538**

Table 7
Performance of classification models using a combination of lexical and different types of non-lexical features according to standard metrics for the task of annotating caregiver interview session transcripts. Highest value for each metric and codebook size across all models is highlighted in boldface.

Cls.	Model	Acc.	Prec.	Rec.	F1	Kappa
	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
16	SVM-LIWC	0.730	0.730	0.730	0.717	0.686
	SVM-AF	**0.738**	**0.733**	**0.738**	**0.727**	**0.696**
	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
19	SVM-LIWC	0.620	0.625	0.620	0.613	0.581
	SVM-AF	**0.638**	**0.639**	**0.638**	**0.631**	**0.601**
	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
58	SVM-LIWC	0.459	0.445	0.459	0.429	0.422
	SVM-AF	**0.488**	**0.466**	**0.488**	**0.462**	**0.454**

3.2.1. Comparison of performance of different classification models

The accuracy of NB-M, SVM, CNN, CRF, SVM-AF, J48 decision tree, Random Forest, AdaBoost, and DiscLDA classification models for the task of annotating adolescent and caregiver datasets is compared across codebooks of different sizes in Figs. 11 and 12.

From Figs. 11 and 12, it follows that SVM and CRF achieve around 52%, 60%, and 70% accuracy when using the codebooks consisting of 41, 20, and 17 labels, respectively, to annotate adolescent session transcripts and 45%, 55%, and 66% accuracy when using the codebooks consisting of 58, 19, and 16 labels, respectively, to annotate caregiver session transcripts. CNN also has approximately the

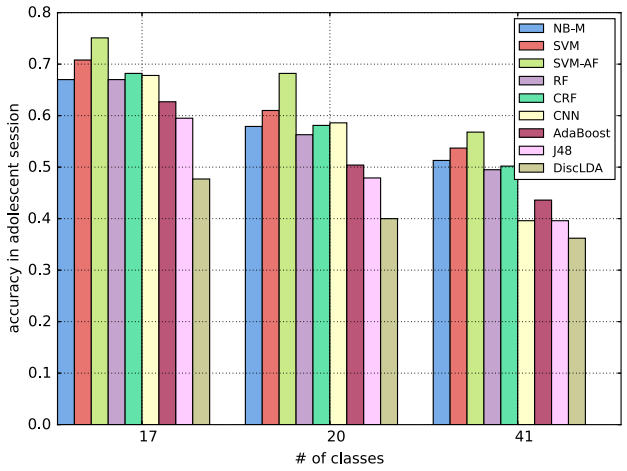

Fig. 11. Comparison of annotation accuracy of adolescent interview fragments with different machine learning methods and feature sets.

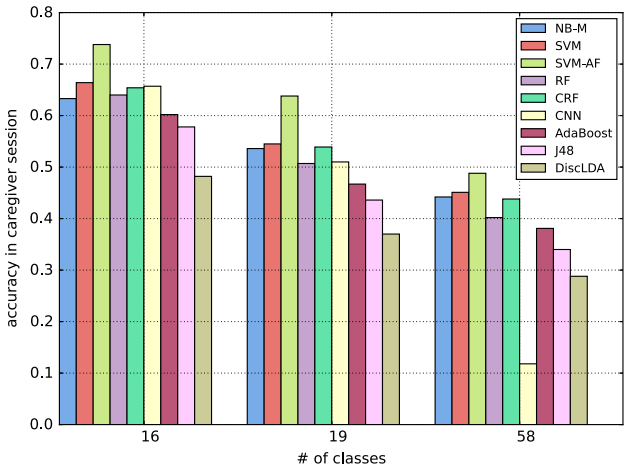

Fig. 12. Comparison of annotation accuracy of caregiver interview fragments with different machine learning methods and feature sets.

same performance as SVM and CRF, when the codebooks consisting of 16, 17 and 20 labels are used. However, CNN has significantly lower performance compared to SVM and CRF in terms of all metrics when the codebooks of size 41 and 58 labels are used. SVM-AF consistently outperforms all other methods across the codebooks of all sizes on both datasets, achieving the highest accuracy of 75.1% (which is close to human accuracy), when the codebook consisting of 17 classes is used for annotating adolescent interview session transcripts, and of 73.8%, when the codebook consisting of 16 classes is used for annotating caregiver interview session transcripts.

Depending on the type of interview transcript and codebook size, SVM-AF achieves 3–9% higher accuracy and 4–10% higher F1 score than SVM and 4–10% higher accuracy and 4–11% higher F1 score than CRF, which highlights the importance of contextual and semantic features.

ROC curves [46] in Figs. 13–15 illustrate the relative performance of classifiers for annotation of adolescent interview transcripts using the codebooks of different size, while Table 8 provides the corresponding AUC (Area Under Curve) values. As follows from Table 8, SVM-AF has the highest AUC compared to all other classifiers. The AUC for automatic annotation using 17-class codebook by SVM using only lexical features is 0.758. Adding contextual and semantic features helps increase the AUC by 7.7%.

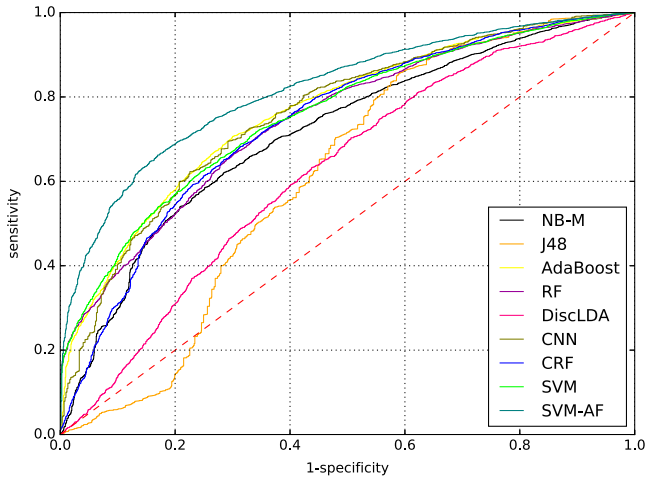

Fig. 13. ROC curves for all classifiers when the codebook with 17 classes is used.

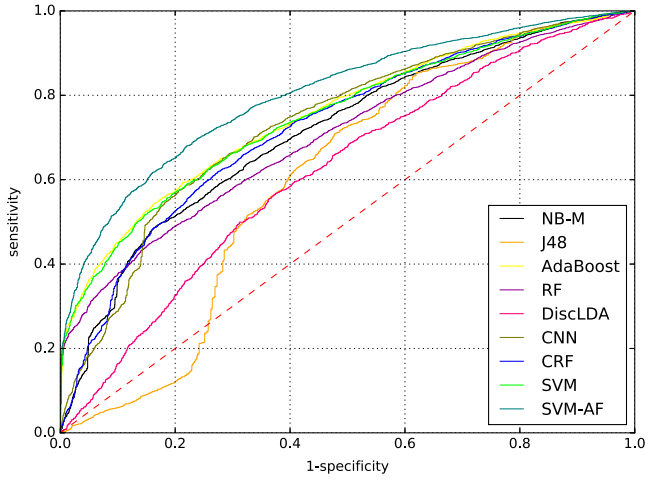

Fig. 14. ROC curves for all classifiers when the codebook with 20 classes is used.

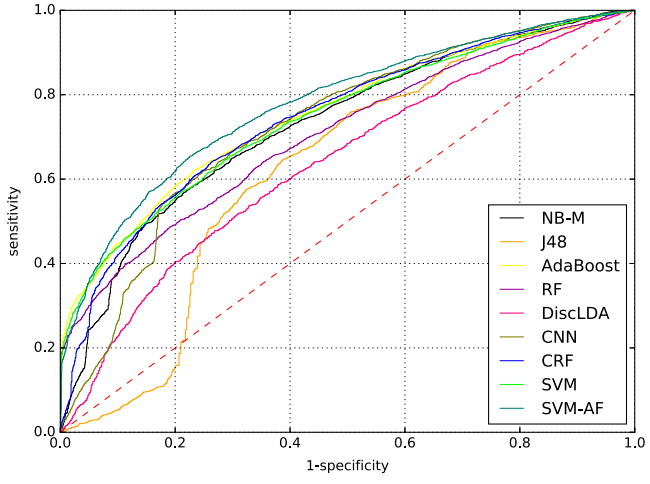

Fig. 15. ROC curves for all classifiers when the codebook with 41 classes is used.

We also observed that NB-M provides the average AUC comparable to all other classifiers and that DiscLDA and J48 have the lowest AUC values around 0.610. AdaBoost, RF, CRF and CNN demonstrate similar performance in terms of AUC that is close to 0.75.

Table 8
AUC values of all classifiers for the codebooks of different size. Highest value for each codebook size across all models is highlighted in boldface.

Model	17 Classes	20 Classes	41 Classes
NB-M	0.714	0.712	0.730
J48	0.613	0.606	0.629
AdaBoost	0.764	0.754	0.755
RF	0.747	0.702	0.706
DiscLDA	0.622	0.618	0.637
CNN	0.760	0.733	0.727
CRF	0.739	0.724	0.745
SVM	0.758	0.750	0.746
SVM-AF	**0.816**	**0.802**	**0.778**

4. Discussion

Experimental evaluation of supervised machine learning methods for the task of automatic annotation of clinical interview transcripts resulted in several important observations and conclusions. First, although CNN has comparable performance to SVM when the number of classes is relatively small, its performance drastically decreases when the number of classes gets large. Remarkably, for very large number of classes (41 and 56, in our case) CNN is less accurate than a random guess. Second, multinomial and binomial Naïve Bayes, AdaBoost, Random Forest, and DiscLDA have been consistently outperformed on both datasets and all codebook sizes by CNN, CRF and SVM, when all models use only lexical features. Superior generalization ability of SVM even in the case of a large number of classes and features (which is the case when lexical features are used) can be attributed to its ability to learn the classification model independent of the dimensionality of a feature space.

We also observed a consistent trend of performance improvement for SVM when adding non-lexical features, such as the label of the preceding utterance and the features derived from LIWC dictionaries, to the lexical ones. The first result indicates that the context of an utterance in clinical interview transcripts in the form of the label of the preceding utterance plays an important role in the classification process, besides the content of the utterance itself. The second result indicates that, for the purpose of classification, the semantics of an utterance in clinical interviews can be approximated with a distribution of its words across LIWC dictionaries.

5. Conclusions

In this work, we propose novel features and report the results of an extensive experimental evaluation of state-of-the-art supervised machine learning methods for text classification using those features, to help clinical researchers and practitioners assess the feasibility of using these methods for the task of automatic annotation of clinical text using codebooks of realistic size. We found out that Support Vector Machine using only lexical features consistently outperforms all other classifiers on caregiver and adolescent datasets according to most metrics. Adding contextual and semantic features further improves the performance of SVM on both datasets, achieving close to human accuracy when the codebooks consisting of 16 and 17 classes are used to annotate caregiver and adolescent transcripts, respectively.

This work has important practical implications. First, it can facilitate researchers to establish a causal relationship between different communication strategies and desired behavioral outcomes without having to repeatedly wade through pages of interview transcripts. Second, since automatic annotation is significantly faster than manual, it can dramatically accelerate the pace of research in the behavioral sciences. Although all experiments were conducted on interview transcripts, the proposed methods and features are not specific to a particular domain of Motivational Interviewing, and thus there is no prima facie reason to believe that they will not be effective for annotation of any other type of clinical conversation.

Ethics approval

Wayne State University IRB provided ethics approval.

Conflicts of interest

The authors whose names are listed certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Acknowledgements

We would like to thank student assistants in the Pediatric Prevention Research Center and Textual Data Analytics Laboratory at Wayne State University for their help with transcribing, annotating and pre-processing the interview transcripts and the codebooks used for experiments reported in this paper. This research was funded by the grants from NHLBI (1U01HL097889-01 Naar-King & Jen, PIs) and the Karmanos Cancer Institute Behavioral and Field Research Core (P30CAP30CA022453-23 Bepler, PI).

References

[1] A. Kotov, M. Hasan, A. Carcone, M. Dong, S. Naar-King, K. Brogan Hartlieb, Interpretable probabilistic latent variable models for automatic annotation of clinical text, AMIA Annu. Symp. Proc. (2015) 785–794.
[2] B. Pang, L. Lee, Opinion mining and sentiment analysis, Found. Trends Inform. Retr. 2 (1–2) (2008) 1–35.
[3] A. Nguyen, D. Moore, I. McCowan, M.J. Courage, Multi-class classification of cancer stages from free-text histology reports using support vector machines, in: Proc. of the 29th Annual International Conference of the IEEE, 2007, pp. 5140–5143.
[4] C.L. Ogden, M.D. Carroll, B.K. Kit, et al., Prevalence of obesity and trends in body mass index among us children and adolescents, 1999–2010, JAMA 307 (5) (2012) 483–490.
[5] U.S. Department of Health and Human Services, The Surgeon Generals vision for a healthy and fit nation, U.S. Department of Health and Human Services, Office of the Surgeon General, Rockville, MD, January 2010.
[6] W.R. Miller, S. Rollnick, Motivational Interviewing: Helping People Change, third ed., The Guilford Press, New York, 2012.
[7] S. Rollnick, W.R. Miller, C.C. Butler, Motivational Interviewing in Health Care: Helping Patients Change Behavior, Guilford Press, 2007.
[8] T.R. Apodaca, R. Longabaugh, Mechanisms of change in motivational interviewing: a review and preliminary evaluation of the evidence, Addiction 104 (5) (2009) 705–715.
[9] D. Walker, R. Stephens, J. Rowland, et al., The influence of client behavior during motivational interviewing on marijuana treatment outcome, Addict. Behav. 36 (6) (2011) 669–673.
[10] A. Idalski Carcone, S. Naar-King, K. Brogan, et al., Provider communication behaviors that predict motivation to change in African American adolescents with obesity, J. Develop. Behav. Pediatr. 34 (8) (2013) 599–608.
[11] T.B. Moyers, T. Martin, J.M. Houck, P.J. Christopher, J.S. Tonigan, From in-session behaviors to drinking outcomes: a causal chain for motivational interviewing, J. Consult. Clin. Psychol. 77 (6) (2009) 1113.
[12] C.R. Teal, R.L. Street, Critical elements of culturally competent communication in the medical encounter: a review and model, Soc. Sci. Med. 68 (3) (2009) 533–543.
[13] M.B. Laws, M.C. Beach, Y. Lee, et al., Provider-patient adherence dialog in HIV care: results of a multisite study, J. AIDS Behav. 17 (1) (2013) 148–159.
[14] M.B. Laws, T. Taubin, T. Bezreh, et al., Problems and processes in medical encounters: the cases method of dialogue analysis, Patient Educ. Couns. 91 (2) (2013) 192–199.
[15] E. Mayfield, M.B. Laws, I.B. Wilson, et al., Automating annotation of information-giving for analysis of clinical conversation, J. Am. Med. Inform. Assoc. 21 (2014) e122–e128.

[16] K. Roberts, S.M. Harabagiu, A flexible framework for deriving assertions from electronic medical records, J. Am. Med. Inform. Assoc. 18 (5) (2011) 568–573.
[17] R. Lacson, R. Barzilay, Automatic processing of spoken dialogue in the home hemodialysis domain, AMIA Annu. Symp. Proc. (2005) 420–424.
[18] I. Rish, An empirical study of the Naive Bayes classifier, IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, 3 (22), 2001, pp. 41–46.
[19] A.K. McCallum, K. Nigam, A comparison of event models for Naive Bayes text classification, Proc. AAAI-98 Workshop on Learning for Text Categorization, 752, 1998, pp. 41–48.
[20] A.M. Kibriya, E. Frank, B. Pfahringer, et al., Multinomial Naive Bayes for text categorization revisited, in: Advances in Artificial Intelligence, 2005, pp. 488–499.
[21] C. Cortes, V. Vapnik, Support-vector networks, Machine Learn. 20 (3) (1995) 273–297.
[22] K.S. Durgesh, B. Lekha, Data classification using support vector machine, J. Theor. Appl. Inform. Technol. 12 (1) (2010) 1–7.
[23] J.D. Lafferty, A.K. McCallum, F.C.N. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: Proc. of 18th International Conference on Machine Learning, 2001, pp. 282–289.
[24] C. Sutton, A.K. McCallum, An introduction to conditional random fields for relational learning, Introduct. Stat. Relational Learn. (2006) 93–128.
[25] A.K. Sharma, S. Sahni, A comparative study of classification algorithms for spam email data analysis, Int. J. Comput. Sci. Eng. 3 (5) (2011) 1890–1895.
[26] Y. Freund, R.E. Schapire, A short introduction to boosting, J. Japanese Soc. Artif. Intell. 14 (5) (1999) 771–780.
[27] L. Breiman, Random forests, Machine Learn. 45 (1) (2001) 5–32.
[28] S. Lacoste-Julien, F. Sha, M.I. Jordan, DiscLDA: discriminative learning for dimensionality reduction and classification, in: Advances in Neural Information Processing Systems, 2009, pp. 897–904.
[29] Y. Kim, Convolutional neural networks for sentence classification, in: Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1746–1751.
[30] M. Hall, E. Frank, G. Holmes, et al., The WEKA data mining software: an update, ACM SIGKDD Explor. Newslett. 11 (1) (2009) 10–18.
[31] Y.R. Tausczik, J.W. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods, J. Lang. Soc. Psychol. 29 (1) (2010) 24–54.
[32] M. De Choudhury, M. Gamon, S. Counts, et al., Predicting depression via social media, in: Proc. of the 7th International Conference on Weblogs and Social Media, 2013.
[33] M. De Choudhury, S. Counts, E. Horvitz, Predicting postpartum changes in emotion and behavior via social media, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2013, pp. 3267–3276.
[34] G.H. John, P. Langley, Estimating continuous distributions in Bayesian classifiers, in: The Proc. the 11th Conference on Uncertainty in Artificial Intelligence, 1995, pp. 338–345.
[35] K. Nigam, A.K. McCallum, S. Thrun, et al., Learning to classify text from labeled and unlabeled documents, in: Proc. of the 10th Conference on Artificial Intelligence, 1998, pp. 792–799.
[36] C.C. Chang, C.J. Ling, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 27.
[37] A.K. McCallum, MALLET: A Machine Learning for Language Toolkit, 2002.
[38] Y. Freund, Boosting a weak learning algorithm by majority, Inform. Comput. 121 (2) (1995) 256–285.
[39] S. Gnter, H. Bunke, New boosting algorithms for classification problems with large number of classes applied to a handwritten word recognition task, in: Multiple Classifier Systems, 2003, pp. 326–335.
[40] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Machine Learn. Res. 3 (2003) 993–1022.
[41] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proc. of the National Academy of Sciences, 101 (1), 2004, pp. 5228–5235.
[42] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.
[43] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcouglu, P. Kuksa, Natural language processing (almost) from scratch, J. Machine Learn. Res. 12 (2011) 2493–2537.
[44] Y. Zhang, B. Wallace, A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification, 2015. Available from: arXiv:1510.03820.
[45] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, Proc. of the 14th International Joint Conference on Artificial Intelligence, 2, 1995, pp. 1137–1145.
[46] T. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. 27 (8) (2006) 861–874.