Interpretable Probabilistic Latent Variable Models for Automatic Annotation of Clinical Text

Alexander Kotov, Ph.D.[1], Mehedi Hasan, B.S.[1], April Carcone, Ph.D[2], Ming Dong, Ph.D.[1], Sylvie Naar-King, Ph.D[2], Kathryn BroganHartlieb, Ph.D R.D.[3]
[1]Department of Computer Science, [2]Pediatric Prevention Research Center, Wayne State University; [3]Department of Dietetics and Nutrition, Florida International University

Abstract

We propose Latent Class Allocation (LCA) and Discriminative Labeled Latent Dirichlet Allocation (DL-LDA), two novel interpretable probabilistic latent variable models for automatic annotation of clinical text. Both models separate the terms that are highly characteristic of textual fragments annotated with a given set of labels from other non-discriminative terms, but rely on generative processes with different structure of latent variables. LCA directly learns class-specific multinomials, while DL-LDA breaks them down into topics (clusters of semantically related words). Extensive experimental evaluation indicates that the proposed models outperform Naïve Bayes, a standard probabilistic classifier, and Labeled LDA, a state-of-the-art topic model for labeled corpora, on the task of automatic annotation of transcripts of motivational interviews, while the output of the proposed models can be easily interpreted by clinical practitioners.

1. Introduction

Annotation or assignment of codes (labels) from a predefined codebook to fragments (entire documents or their parts) of clinical text is an integral part of medical practice and qualitative research. Such codes can be viewed as semantic labels, or high-level summaries (abstractions) of the raw textual data. Besides cataloging, such abstractions can facilitate the analysis of clinical text in general and clinical interview transcripts in particular. In this work, we focus on the transcripts of motivational interviews with obese adolescents conducted at a Pediatric Prevention Research Center (PPRC).

Childhood and adolescent obesity is a serious public health problem. Recent national data[1] indicate that one in four children aged 2-5 years are overweight or obese. The trend of childhood obesity continues into adolescence – as of 2012, 18% of all adolescents and 23.7% of African American adolescents are obese[1]. Adolescents who are obese are likely to be obese as adults and have a greater risk of heart disease, type 2 diabetes, stroke, cancer, and osteoarthritis[2]. Therefore, childhood and adolescence are critical periods for healthy eating and physical activity interventions to establish healthy weight gain trajectories. To design such interventions, PPRC clinicians conduct interviews with children and their caregivers grounded in the principles of Motivational Interviewing (MI)[3], an evidence-based communication technique to increase intrinsic motivation and self-efficacy for behavior change. Detailed analysis of those interviews aims at identifying clinicians' communication strategies that are effective in triggering patient's motivational statements for the behavioral changes that will ultimately lead to weight loss. Recent literature reviews of mechanisms of effect in MI[4] have concluded that clients' motivational statements about their own desire, ability, reasons and need for or commitment to change (or "change talk") consistently predict actual behavior change[5], as long as 34 months later[6]. Strategies to elicit motivational statements are typically identified via retrospective analysis of past interview transcripts. Part of this analysis involves assignment of codes to patient replies during the interviews. Analyzing sequences of assigned codes allows clinicians to better understand the patient's thought process during the course of the interviews, without having to wade through entire transcripts over and over again. Such understanding, in turn, leads to further specification of the mechanisms of effect for intervention models, which can then be used to refine theory and guide clinical practice[7].

Annotation of interview transcripts has traditionally been performed manually by trained coders, which is a tedious and resource intensive process. Therefore, methods that can efficiently and accurately distinguish the nuances of patient-provider communication can have a tremendous positive impact on many areas of clinical practice and research. Inferring psychological state of the patients during clinical interviews using only lexical content of their transcriptions is a challenging task for several reasons. First, many important indicators of emotions such as gestures, facial expressions and intonations are lost during the transcription process. Second, some utterances from patients during the interview may be too short and lack sufficient context for accurate classification. Furthermore, patients come from a variety of social, cultural, and educational backgrounds and their language is therefore quite

different. This problem is exacerbated when the interviews are conducted with children and adolescents, since children, in general, tend to often use incomplete sentences and frequently change subjects.

Automating the annotation of clinical documents is one of the fundamental problems in medical informatics, which can have tremendous implications for clinical practice. It falls under a general class of classification problems, which are typically addressed using supervised machine learning methods (or classifiers). Given a set of pre-classified data samples (called the training set) represented as a feature vector, in which each feature is the value of a feature function calculated based on a data sample, these methods learn the mapping from feature vectors to their classifications. Once learned, such mapping can be applied to classify new, unlabeled data samples (called the testing set). Classification problems arise in many different domains, from analysis of scientific literature[8] and on-line reviews[9] to digital forensics[10] and medical informatics[11,12]. Performance of different classifiers (including Naïve Bayes[13]) on most common text classification tasks has been examined in detail in previous work[14,15]. However, biomedical context places additional restriction of interpretability on machine learning models, as they are not only required to make correct classification decisions, but to also allow humans to easily understand how they arrived at these decisions. Interpretability of classification models is particularly important for psychological studies, such as Motivational Interviewing, since each class needs to have distinct interpretation (i.e. clearly correspond to a certain communication or behavior type). Furthermore, annotation models for these studies often need to be manually corrected. While the effectiveness of non-interpretable classifiers leveraging external resources, such as concepts from the Unified Medical Language System (UMLS) or clusters derived from a large external corpus, has been previously studied[16], there is still a need for designing interpretable models for annotating clinical interviews for behavioral studies, which typically contain very limited, domain-specific terminology and thus render general purpose medical lexicons ineffective for this task.

In this work, we focus on the problem of designing an interpretable model for automatic annotation of utterances in clinical interview transcripts with fine-grained semantic class (such as behavior type) and propose two new latent variable probabilistic models, Latent Class Allocation and Discriminative Labeled Latent Dirichlet Allocation as effective solutions to this problem. Both methods model how human annotators approach classification using probabilistic generative process. In particular, during training, they learn to distinguish the vocabularies that are highly indicative of the given classes from the general and non-discriminative terms via probabilistic assignment of latent variables to each term in the training corpus. The learned vocabularies in the form of multinomial distributions (or language models) are used to probabilistically classify new textual fragments and are easily interpretable by clinical practitioners. Although all experiments in this paper were conducted using clinical conversation data, the proposed methods can be applied to annotate any other type of clinical text.

2. Methods

2.1 Classes

As a golden standard for all experiments in this work, we used a sample of obesity Motivational Interview transcripts, in which patient utterances were manually annotated by human coders according to the "Minority Youth Sequential Code for Observing Process Exchanges" (MY-SCOPE)[17] coding manual. Each utterance in the golden standard is labeled with one class. Among others, MY-SCOPE defines the following five classes of patient utterances, which correspond to major target patient behaviors clinicians were focused on when conducting obesity Motivational Interviews:

- **CL-:** negative commitment language;
- **CL+:** positive commitment language;
- **CT-:** negative change talk;
- **CT+:** positive change talk;
- **AMB:** ambivalence.

Commitment language (CL) is statements about patients' intentions or plans for enacting weight related changes, which are positive, when supportive of behavior change, and negative, when against behavior change. Change talk (CT) corresponds to utterances that describe patients' own desires, abilities, reasons, and need for adhering to weight loss recommendations and are also positive, when supportive of behavior change, and negative, when against behavior change. Ambivalent utterances (AMB) are change talk or commitment language statements that contain a

combination of positive and negative sentiments about changing one's behavior. Examples of utterances for each class are presented in Table 1.

Table 1. Examples of utterances representing the language and behavior types considered in this work.

Category	Example
CL-	I eat a lot of junk food. Like, cake and cookies, stuff like that.
CL+	Well, I've been trying to lose weight, but it really never goes anywhere.
CT-	It can be anytime; I just don't feel like I want to eat (before) I'm just not hungry at all.
CT+	Hmm. I guess I need to lose some weight, but you know, it's not easy.
AMB	Fried foods are good. But it's not good for your health.

In the following sections, we present the classifiers used for the task of differentiating the above classes and report their performance in terms of standard evaluation metrics.

2.2 Features and baselines

We use standard bag-of-words feature generation framework, in which a predefined set of lexical features (or vocabulary), $V = \{w_1,\ldots,w_N\}$, can appear in a given textual fragment. For example, one such feature could be the number of times a word (unigram) "exercise" appears in a given textual fragment. This way each textual fragment *f* is represented as a feature vector ($n_{w_1,f}$, ...,$n_{w_N,f}$), where $n_{w,f}$ is the number of times feature (word) *w* occurs in *f*.

To determine the best classification model for this task, in the following sections, we experimentally compare standard supervised machine learning methods, such as Naïve Bayes[13] and Labeled Latent Dirichlet Allocation[18], with our proposed probabilistic classification models.

2.2.1 Naïve Bayes

Naïve Bayes (NB) is a standard probabilistic classifier, which annotates a given textual fragment $f = \{w_1,\ldots,w_{N_f}\}$, consisting of N_f words, with a class *c**, such that *c** = arg maxc p(c|f), where *p(c|f)* is estimated by applying the Bayes' rule as follows:

$$p(c|f) = \frac{p(f|c)p(c)}{p(f)} \propto p(f|c)p(c)$$

In order to estimate *p(f|c)*, Naïve Bayes classifier makes an assumption about conditional independence of features given c, the class of fragment *f*:

$$p(f|c) = \prod_{i=1}^{N_f} p(w_i|c)^{n_{w_i,f}}$$

Despite its relative simplicity, NB has been experimentally demonstrated to be one of the most effective text classification algorithms ever created. In this work, we used a standard implementation of Multinomial NB algorithm from the Weka text mining toolkit.

2.3 Probabilistic models

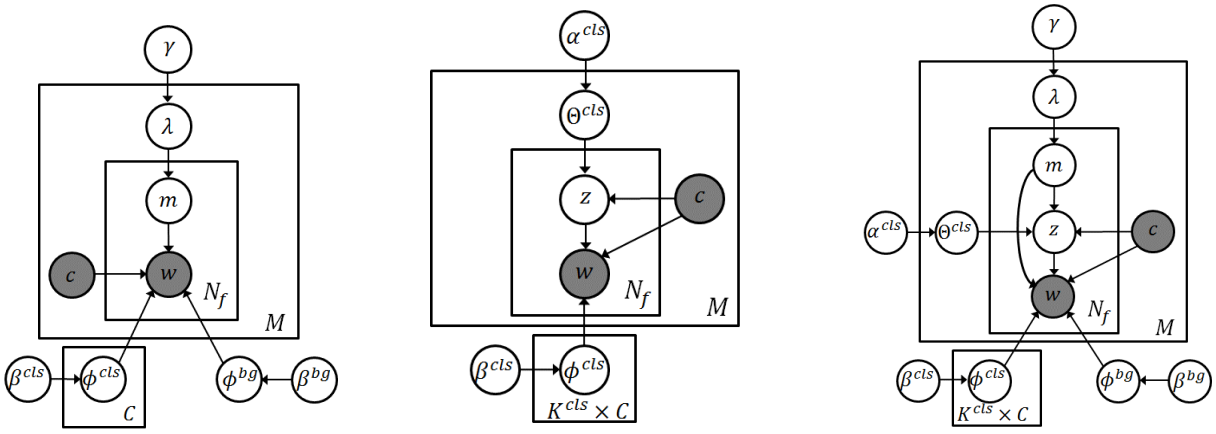
We propose Latent Class Allocation (LCA) and Discriminative Labeled Latent Dirichlet Allocation (DL-LDA), two novel probabilistic generative latent variable models for the task of automatic coding of clinical interview transcripts, and compare their performance with Naïve Bayes and Labeled Latent Dirichlet Allocation (L-LDA), a state-of-the-art probabilistic model for labeled data, on the task of annotating utterances in clinical text. LCA associates only one latent variable *m* with each word, which determines its type (whether a word is general or characteristic of a certain class). DL-LDA is an extension of L-LDA that makes a different set of assumptions about the structure of latent variables. Rather than directly associating each word with a latent variable determining its topic for a certain class, DL-LDA, similar to LCA, first associates with each word a latent variable, which determines whether a word is general or characteristic of a certain class and, *only in the second case*, associates it with another latent variable *z*, determining its class-specific topic. Thus, DL-LDA can be viewed as a more structured version of Labeled LDA.

2.3.1 LCA

LCA models each textual fragment f labeled with class c_f as a set of alternating draws from a background multinomial ϕ^{bg} that is drawn from a symmetric Dirichlet prior β^{bg} and a multinomial ϕ^{cls} specific to c_f that is drawn from a symmetric Dirichlet prior β^{cls}. The proportion of words drawn from ϕ^{bg} and ϕ^{cls} is controlled by a binomial distribution λ_f. LCA generates annotated textual fragments according to the following probabilistic process:

1. draw $\lambda_f \sim$ Beta(γ), a binomial distribution controlling the mixture of words in f drawn from the background and class-specific multinomials
2. for each word position i of N_f in f:
   - *a)* draw Bernoulli switching variable $m_{f,i} \sim \lambda_f$
   - b) if $m_{f,i}$ = bg:
     - draw a word $w_{f,i} \sim \phi^{bg}$
   - c) if $m_{f,i}$ = cls:
     - draw a word $w_{f,i} \sim \phi^{cls,c_f}$

The generative process of LCA in plate notation is presented in Figure 1a.



a) LCA **b) L-LDA** **c) DL-LDA**

Figure 1. Generative processes of the proposed and baseline latent variable models in plate notation.

Annotation of textual fragments in the testing set with LCA is done using class-specific multinomials ϕ^{cls} (or $p(w|c)$) determined as a result of posterior inference on the training set to derive $p(c|w)$, distributions showing how indicative each word w is for each class c:

$$p(c|w) = \frac{p(w|c)p(c)}{p(w)}$$

where $p(c) = \frac{n_{f,c}}{M}$ ($n_{f,c}$ is the number of interview fragments labeled with class c and M is the total number of fragments) and $p(w)$ is a probability of word w in a collection language model estimated using maximum likelihood. $p(c|w)$ are then used to classify f according to the following formula:

$$c^* = \arg\max_c p(c|f) = \prod_{i=1}^{N_f} p(c|w_i)^{n_{w_i,f}}$$

2.3.3 L-LDA

L-LDA directly associates a latent variable z with each word that determines its assignment to a topic specific to c_f. It is state-of-the-art topic model for labeled textual collections that has been shown to outperform standard classifiers, such as SVM, for the task of multi-class classification[18]. The generative process of L-LDA in plate notation is presented in Figure 1b. L-LDA along with NB is used as a baseline in our experimental evaluation.

2.3.2 DL-LDA

DL-LDA models each textual fragment f labeled with class c_f as a mixture of the background topic ϕ^{bg} drawn from a symmetric Dirichlet prior β^{bg} and K^{cls} topics drawn from a uniform Dirichlet prior β^{cls}. DL-LDA generates the textual fragments in clinical interviews according to the following probabilistic process:

1. draw $\lambda_f \sim$ Beta(γ), a binomial distribution controlling the mixture of background and class-specific topics for f
2. draw $\theta_f^{cls} \sim$ Dir(α^{cls}), a distribution of class-specific topics for f
3. for each word position i of N_f in f:

   (a) draw Bernoulli switching variable $m_{f,i} \sim \lambda_f$

   (b) if $m_{f,i}$ = bg:

   - draw a word $w_{f,i} \sim \phi^{bg}$

   (c) if $m_{f,i}$ = cls:

   - draw a topic $z_{f,i} \sim \theta_f^{cls}$
   - draw a word $w_{f,i} \sim \phi_{z_{f,i}}^{cls,c_f}$

The generative process of DL-LDA in plate notation is presented in Figure 1c. Classification using DL-LDA is performed by first deriving a class-specific multinomial $p(w|c)$ per each class c from class-specific topics $\phi^{cls,cf}$ (or $p(w|c, z)$) by marginalizing over z:

$$p(w|c) = \sum_{z=1}^{K^{cls}} p(w|c_i, z)$$

and then using class-specific multinomials to directly classify f, similar to LCA.

We would like to note that the inference algorithm for LCA, L-LDA and DL-LDA is adaptable to distributed environment and therefore the proposed methods can be scaled up to very large datasets[19,20].

3. Results

The dataset used for experiments in this work consists of 2966 manually annotated fragments of interview transcripts. The distribution of the number of samples per each class is shown in Table 2.

Table 2. Number of samples per class in experimental dataset.

Class	# Samples	%
CL-	73	2.46 %
CL+	875	29.50 %
CT-	278	9.37 %
CT+	1657	55.87 %
AMB	83	2.80 %
Total	**2966**	**100 %**

The dataset was first pre-processed by removing very frequently occurring terms (that occur in more than 25% of textual fragments). We also used the following pre-processing methods to study their effect on classification performance:

- **RAW:** no preprocessing, original dataset is used;
- **STEM:** Porter stemmer is applied to each term in the dataset to eliminate morphological variation;
- **STOP:** stopwords are removed, but stemming is not applied;

- **STOP-STEM:** Porter stemmer is applied to each term in the dataset and stopwords are removed.

For all experiments we used randomized 5-fold cross-validation. The Gibbs sampler for posterior inference of parameters of LCA, L-LDA and DL-LDA was run for 1000 iterations. Classification performance of NB, L-LDA, LCA and DL-LDA on the task of differentiating 5 language categories when experimental dataset is pre-processed with different methods is summarized in Tables 3, 4, 5 and 6, respectively.

Table 3. Performance of Naïve Bayes using different pre-processing methods. Best result for each performance metric is highlighted in boldface.

Method	Recall	Precision	F1 score
RAW	0.522	0.523	0.506
STEM	**0.534**	**0.534**	**0.518**
STOP	0.511	0.526	0.510
STOP-STEM	0.510	0.519	0.506

Table 4. Performance of Labeled Latent Dirichlet Allocation using different pre-processing methods. Best result for each performance metric is highlighted in boldface.

Method	Recall	Precision	F1 score
RAW	0.537	0.530	0.480
STEM	**0.544**	**0.540**	0.474
STOP	0.530	0.520	**0.478**
STOP-STEM	0.538	0.517	0.475

Table 5. Performance of Latent Class Allocation using different pre-processing methods. Best result for each performance metric is highlighted in boldface.

Method	Recall	Precision	F1 score
RAW	0.543	0.534	0.537
STEM	**0.557**	**0.542**	**0.549**
STOP	0.541	0.508	0.520
STOP-STEM	0.543	0.515	0.525

Table 6. Performance of Discriminative Labeled Latent Dirichlet Allocation using different pre-processing methods. Best result for each performance metric is highlighted in boldface.

Method	Recall	Precision	F1 score
RAW	**0.591**	**0.533**	**0.537**
STEM	0.586	0.515	0.527
STOP	0.560	0.504	0.508
STOP-STEM	0.557	0.492	0.498

Table 7. Summary of the best performance of different methods for the task of annotation of 5 original language types. Best result for each performance metric is highlighted in boldface.

Algorithm	Recall	Precision	F1 score
Naïve Bayes	0.522	0.523	0.506
LCA	0.543	**0.534**	**0.537**
L-LDA	0.537	0.530	0.480
DL-LDA	**0.591**	0.533	0.537

Since classification accuracy of DL-LDA is dependent on the number of per-class topics, which is a parameter that needs to be specified a priori, we first optimized it with respect to F1 score. Figure 2 indicates that the optimal classification results for DL-LDA in combination with different pre-processing methods are achieved when the number of topics is small (2 or 3 in most cases).

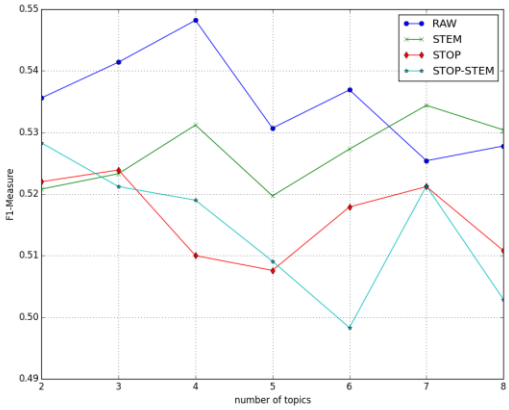


Figure 2. F1 score of DL-LDA by varying the number of topics and in combination with different pre-processing methods on the task of classification of all language categories.

The best results for each proposed model and the baselines are summarized and compared in Table 7, while the per-class breakdown of the best results for all models is provided in Table 8.

Table 8. Summary of the best per class performance in terms of F1 score of different classifiers for the task of distinguishing 5 original language types. Best result for each class is highlighted in boldface.

Algorithm	1	2	3	4	5
Naïve Bayes	**0.129**	0.329	0.164	0.691	0.170
LCA	0.094	**0.437**	**0.223**	0.682	**0.162**
L-LDA	0.025	0.252	0.066	0.708	0.128
DL-LDA	0.025	0.396	0.114	**0.729**	0.061

In the second set of experiments, we aggregated the interview fragments labeled as positive and negative commitment language (CL+ and CL-) and change talk (CT+ and CT-) into one combined class for commitment language (CL) and one combined class for change talk (CT), respectively, and evaluated the accuracy of our classifiers in distinguishing the interview fragments labeled with the resulting three broader classes.

Table 9. Number of samples per aggregated positive and negative sub-classes of CL and CT.

Class	Samples	%
CL	948	31.96 %
CT	1935	65.24 %
AMB	83	2.80 %
Total	**2966**	**100 %**

Distribution of samples across these three classes is shown in Table 9. For this task we used raw data for each classifier (no preprocessing). We optimized the number of topics for DL-LDA with respect to the F1 score (Figure 3) and found out that again the optimal number of topics is 3. Performance of different classifiers on the task of differentiating the interview fragments labeled with CL, CT and AMB is summarized in Table 10. In the third set of experiments, we aggregated the interview fragments labeled as positive sub-classes of commitment language (CL+) and change talk (CT+) into one combined positive class (+) and negative sub-classes of commitment language (CL-) and change talk (CT-) into one combined negative class (-) and evaluated the accuracy of our classifiers in distinguishing the interview fragments labeled with the resulting sentiment modality-based broader classes.

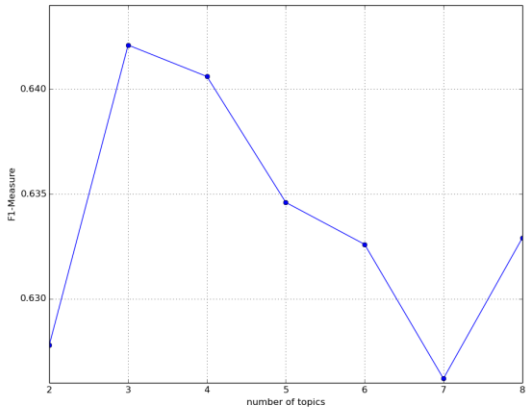


Figure 3. F1 score of DL-LDA by varying the number of topics on the task of classification of aggregated positive and negative sub-classes within CL and CT.

Table 10. Performance of the proposed methods and the baselines on the task of distinguishing aggregated positive and negative sub-classes within CL and CT. Best result for each metric is highlighted in boldface.

Algorithm	Recall	Precision	F1 score
Naïve Bayes	0.617	0.627	0.611
LCA	**0.674**	**0.651**	**0.656**
L-LDA	0.634	0.631	0.587
DL-LDA	0.673	0.637	0.633

Distribution of samples across these three classes in shown in Table 11. Similarly to the task of differentiating CT, CL and AMB, we tuned DL-LDA with respect to F1 score (Figure 5) and found out that this time the optimal number of per-class topics is 5.

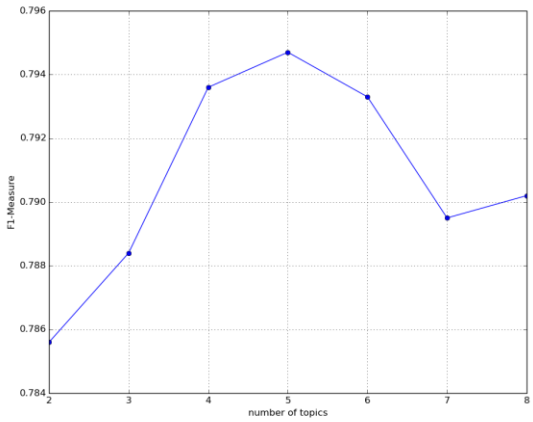


Figure 5. F1 score of DL-LDA by varying the number of topics on the task of classification of aggregated positive and negative sub-classes across CL and CT.

Performance of different classifiers on the task of differentiating the utterances labeled with +, - and AMB is summarized in Table 12.

4. Discussion

Several important observations can be made from Tables 3, 4, 5 and 6. First, stemming and stopwords removal

degrade classification performance of DL-LDA, which performs best without any pre-processing, while Naïve Bayes, LCA and L-LDA achieve the best classification performance when stemming is applied.

Table 11. Number of samples per aggregated positive and negative sub-classes across CL and CT.

Class	# Samples	%
-	351	11.83 %
+	2532	85.37 %
AMB	83	2.80 %
Total	**2966**	**100 %**

Table 12. Performance of the proposed methods and the baselines on the task of distinguishing aggregated positive and negative sub-classes across CL and CT. Best result for each metric is highlighted in boldface.

Algorithm	Recall	Precision	F1 score
Naïve Bayes	0.734	**0.778**	0.753
LCA	0.818	0.771	0.790
L-LDA	0.814	0.774	0.781
DL-LDA	**0.838**	0.770	**0.793**

However, for all 4 classifiers used in this work stopwords removal by itself and in combination with stemming decreases the accuracy of classification, which suggests that common stopwords might be important indicators for some of the language categories. Second, as follows from Table 7, LCA and DL-LDA outperform both baselines (NB and L-LDA) in terms of all three performance measures (Recall, Precision and F1 score). Across all models, DL-LDA achieves the best performance in terms of Recall, while LCA achieves the best performance in terms of both Precision and F1 score. As follows from Table 8, LCA and DL-LDA also achieve the best per-class performance for 4 out of 5 classes. These results lead to two important conclusions. First, explicitly accounting for discriminativity of terms (general or class-specific) in an utterance with a latent variable allows to improve annotation performance using probabilistic methods. Second, additional division of class-specific multinomials into class-specific topics by DL-LDA allows to improve recall, but not precision and F1 score.

Results of classifying language types without taking into account modality (Table 10) indicate that LCA is particularly suited for this task and again support our assumption about the utility of differentiating the terms by their discriminativity. LCA and DL-LDA use the strength of statistical associations of terms with the class labels as a measure of their discriminativity. Since non-discriminative words are the ones that occur in many fragments labeled with different classes, statistical associations of these terms with class labels are relatively weak, which is effectively captured by LCA and DL-LDA.

Table 13. Most characteristic words for each utterance label according to LCA.

Class	Words
CL-	drink sugar gatorade lot hungry splenda beef tired watch tv steroids sleep home nervous confused starving appetite asleep craving pop fries computer
CL+	stop run love tackle vegetables efforts juice swim play walk salad fruit
CT-	got laughs sleep wait answer never tired splenda fault phone joke weird hard don't
CT+	time go mom brother want happy clock boy can move library need adopted reduce sorry solve overcoming lose
AMB	what taco mmm know say plus snow pain weather

Results of classifying modality (Table 12) indicate that DL-LDA is the best in detecting the attitude of the speaker. This can be explained by the fact that only a portion of vocabularies indicative of specific classes reflect the sentiment modality of an utterance, therefore splitting class-specific multinomials into class-specific topics by DL-LDA allows to isolate the sentiment-specific vocabularies and leverage them during classification.

Examples of the most characteristic terms for each utterance label determined by LCA are provided in Table 13. As follows from Table 13, negative commitment language is strongly associated with the words reflecting poor diet ("sugar", "pop", "fries") and sedentary lifestyle ("watch", "tv", "computer"), while positive commitment language is strongly associated with the terms related to exercise ("walk", "play", "run") and healthy food options ("salad", "vegetables", "fruit"). The words characteristic of CT- and CT+ generally reflect negative ("don't", "never", "tired") and positive ("can", "need", "lose", "happy") attitudes towards weight loss, respectively.

5. Conclusion

In this paper, we proposed Latent Class Allocation, a novel interpretable probabilistic model for supervised text classification, and Discriminative Labeled LDA, an extension of Labeled LDA, that differentiates between class-specific and general terms. Through extensive experimental evaluation, we demonstrated that the proposed models have consistently better performance for the task of single class annotation of fragments of Motivational Interviewing transcripts than state-of-the-art methods, such as Naïve Bayes and Labeled LDA.

References

1. Ogden CL, Carroll MD, Kit BK et al. Prevalence of obesity and trends in body mass index among US children and adolescents 1999 – 2010. *JAMA* 2012 Feb 1, 307(5):483–90.
2. U.S. Department of Health and Human Services. The Surgeon General's vision for a healthy and fit nation. Rockville (MD): *U.S. Department of Health and Human Services, Office of the Surgeon General*; 2010.
3. Miller WR, Rollnick S. Motivational Interviewing: Helping People Change. *The Guilford Press*; 2012.
4. Pollak KI, Alexander SC, Coffman CJ, et al. Physician communication techniques and weight loss in adults: Project CHAT. *Am J Prev Med.* 2010 Oct;39(4):321-28.
5. Apodaca TR, Longabaugh R. Mechanisms of change in motivational interviewing: A review and preliminary evaluation of the evidence. *Addiction.* 2009 May;104(5):705-15.
6. Walker D, Stephens R, Rowland J, Roffman R. The influence of client behavior during motivational interviewing on marijuana treatment outcome. *Addict Behav.* 2011 Jun;36(6):669-73.
7. Spear BA, Barlow SE, Ervin C, et al. Recommendations for treatment of child and adolescent overweight and obesity. *Pediatrics.* 2007 Dec;120(4):S254-88.
8. Bergsma S, Post M, Yarowsky D. Stylometric analysis of scientific articles. *Proc. Conf. of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies*. 2012:327-37.
9. Wang S, Manning CD. Baselines and bigrams: Simple, good sentiment and topic classification. *Proc. 50th Annual Meeting of the Association for Computational Linguistics*. 2012:90-4.
10. de Vel OY, Corney MW, Anderson AM, Mohay GM. Language and gender author cohort analysis of e-mail for computer forensics. *Digital Forensic Research Workshop*. 2002.
11. Wallace BC, Laws MB, Small C, et al. Automatically annotating topics in transcripts of patient-provider interactions via machine learning. *Med Decis Making*. 2014 May; 34(4):503-12.
12. Mayfield E, Laws MB, Wilson IB et al. Automating annotation of information-giving for analysis of clinical conversation. *J Am Med Inform Assoc.* 2014 Feb;21(e1):e122-8.
13. McCallum A, Nigam K. A comparison of event models for naive bayes text classification. *Proc. AAAI-98 workshop on learning for text categorization*. 1998:41-8.
14. Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C. Learning word vectors for sentiment analysis. *Proc. 49th Annual Meeting of the Association for Computational Linguistics - Human Language Technologies. 2011*:142-50.
15. Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques. *Proc. 2002 Conf. on Empirical Methods in Natural Language Processing*. 2002:79-6.
16. Lacson R, Barzilay R. Automatic processing of spoken dialogue in the home hemodialysis domain. *AMIA Annu Symp Proc*. 2005:420-24.
17. Idalski Carcone A, Naar-King S, Brogan K, et al. Provider communication behaviors that predict motivation to change in African American adolescents with obesity. *J Dev Behav Pediatr*. 2013 Oct;34(8):599-608.
18. Ramage D, Hall D, Nallapati R, Manning CD. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *Proc. 2009 Conf. on Empirical Methods in Natural Language Processing*. 2009:248-56.
19. Newman D, Asuncion A, Smyth P, et al. Distributed algorithms for topic models. *Journal of Machine Learning Research*. 2009;10:1801-28.
20. Smola A, Narayanamurthy S. An architecture for parallel topic models. *Proceedings of the VLDB Endowment*. 2010 Sep;3(1-2):703-10.