Detecting Health Related Discussions in Everyday Telephone Conversations for Studying Medical Events in the Lives of Older Adults

Golnar Sheikhshabbafghi
B.Sc. in Software Engineering, Sharif University of Technology, Tehran, Iran, 2006
M.Sc. in Artificial Intelligence, Sharif University of Technology, Tehran, Iran, 2009

Presented to the Center for Spoken Language Understanding
within the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of
Master of Science
in
Computer Science & Engineering

August 2014

Center for Spoken Language Understanding
School of Medicine
Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the M.Sc. dissertation of
Golnar Sheikhshabbafghi
has been approved.

Izhak Shafran, Thesis Advisor
Research Scientist, Google, Inc.

Steven Bedrick, On-campus Advisor
Assistant Professor, OHSU

Alexander Kain
Associate Professor, OHSU

Anoop Sarkar
Associate Professor, SFU

Acknowledgment

I would like to thank Dr. Izhak Shafran, Dr. Steven Bedrick, and Dr. Brian Roark for their valuable input, all the faculty members and my fellow classmates in CSLU for what I learnt during the past two years, Ms. Patricia Dickerson for helping me through unfamiliar regulations, and last but not least, my husband for his patience and support.

This research was supported in part by NIH Grants 1K25AG033723, and P30 AG008017, as well as by NSF Grants 1027834, and 0964102. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NIH.

Contents

List of Tables

List of Figures

Abstract

Detecting Health Related Discussions in Everyday Telephone Conversations for Studying Medical Events in the Lives of Older Adults

Golnar Sheikhshabbafghi

Master of Science
Center for Spoken Language Understanding within
Oregon Health & Science University
School of Medicine

August 2014
Thesis advisor: Izhak Shafran, On-campus Advisor: Steven Bedrick

We apply semi-supervised topic modeling techniques to detect health-related discussions in everyday telephone conversations, which has applications in large-scale epidemiological studies and for clinical interventions for older adults. The privacy requirements associated with utilizing everyday telephone conversations preclude manual annotations; hence, we explore semi-supervised methods in this task. We adopt a semi-supervised version of Latent Dirichlet Allocation (LDA) to guide the learning process. Within this framework, we investigate a strategy to discard irrelevant words in the topic distribution and demonstrate that this strategy improves the average F-score on the in-domain task and an out-of-domain task (Fisher corpus). Our results show that the increase in the number of health-related conversations is statistically associated with actual medical events obtained through weekly self-reports.

Chapter 1

Introduction

Daily conversations may contain information about an individual's health status: a person who is suffering from chronic pain is likely to talk about it; those suffering from depression might show more negative sentiments (or little sentiment at all) in their conversations; those with dementia might use more frequent words in comparison to the control group and the grammatical complexity of their language might be significantly lower.

We also may use an individual's conversations to analyze their social network. Stark and colleagues[30] have used language clues in conversations of older adults to predict the nature of those conversations (business vs. social) without violating the subjects' privacy. This work is a sequel of Stark and colleagues' work in that the data is of the same nature (every-day private telephone conversations of older adults) and that the ultimate goal (affordable and home-based monitoring of older adults' health status) is the same. However, our task is different from theirs; we have more subjects and more conversations as well as information about the subjects' medical events such as injuries or medicine change, self reported via weekly questionnaires.

The problem we address is to detect conversations that contain a health related discussion. We are interested in conversations between patients and doctors where the whole conversation is health-related, as well as in general conversation in which, among all different topics being discussed, one is also complaining about pain to a close family member. Our larger goal is to study the association between conversations on health, medical events in their lives, and their social support network. Such associations would be useful in studying depression or suggesting medical visits for patients with chronic illness when their level of pain or discomfort is unusually high.

Among previous work, the one that is similar in spirit to our work is the problem of detecting tweets about tobacco, which took advantage of hashtag and related metadata [25]. In contrast, we utilize only the spoken words and hence our approach is portable across different media and forums. We do not rely on manual transcripts of the conversations and expect the input word sequences to contain ASR errors.

The method we introduce can be used for other topics of interest as well. In our setting, a document corresponds to a whole conversation and we do not care which side of the conversation is mentioning health-related issue. are focused on detecting conversations only one predetermined topic, it is plausible to consider the possibility of specifying a few keywords or seed words and then expanding this set. That is the approach investigated in this thesis. We start with three clearly health-related words (doctor, medicine, and pain) and expand them using topic modeling over the conversations themselves. We use the tf-idf of these vectors as the features for a support vector machine and show that the results of such classifier is more reliable than the ones using the tf-idfs over the whole vocabulary.

For training the classifiers we need labeled data. However, the privacy concerns of recording and analyzing everyday conversations prevents us from manually transcribing or annotating them. We resort to indirect ways of obtaining annotations. We use reverse lookup of caller id to label a small subset of conversations for developing semi-supervised algorithms and for evaluating them.

Before delving into our approach, we discuss related work in Chapter 2, we describe our data in Chapter 3 so as to specify the constraints of our task, constraints that apply more generally to private communications. We describe our semi-supervised topic modeling based approach in Chapter 4. We evaluate the efficacy of our approach empirically and report results in Chapter 5. Finally, we conclude the paper and discuss ideas for future work.

Chapter 2

Related Work

Since our approach is based on a variant of Latent Dirichlet Allocation [6], a state-of-the-art topic modeling method, in this chapter we provide the relevant background on topic modeling and its applications in health-related problems.

2.1 Topic Modeling

After Blei and colleagues introduced Latent Dirichlet Allocation (LDA) in 2003 as a generative model for discrete data generation, it has been extensively employed in many areas of research including even image processing [9] and genomics [22]. Subsequently, many variants of the method have been proposed to address the limitations of the method and researchers have applied LDA to even problems like word sense disambiguation [8]. In this section we will briefly explain LDA and a few of its variants as well as a few applications in health-related problems.

2.1.1 Latent Dirichlet Allocation

LDA models a document as a distribution over a set of K topics shared among all documents in the corpus. The topics are multinomial distributions over words and each documents distribution over topics is sampled from a prior Dririchlet distribution. Here we review the mechanics of LDA and in doing so, we follow the notations of Andrzejewski and Zhu [2]. Andrzejewski and Zhu describe LDA for a corpus D where w_n is the n^{th} word in D, z_n is the index of the topic that generates w_n and d_n is the document corresponding to w_n. The probability of a word w in a topic j is shown as $\phi_j^{(w)}$) whereas the probability

of a topic j in document d is $\theta_j^{(d)}$. With this terminology the generative process in LDA generates a corpus D as follows:

1. Choose $\theta \sim \text{Dir}(\alpha)$.
2. Choose $\phi \sim \text{Dir}(\beta)$.
3. For n in 1...N:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta^{d_n})$.
   (b) Choose the word $w_n \sim \text{Multinomial}(\phi^{z_n})$.

Where N is the number of the words in the corpus D and α, β, and K (the number of topics) are the hyper parameters. Figure 2.1 illustrates the document generation process in LDA.

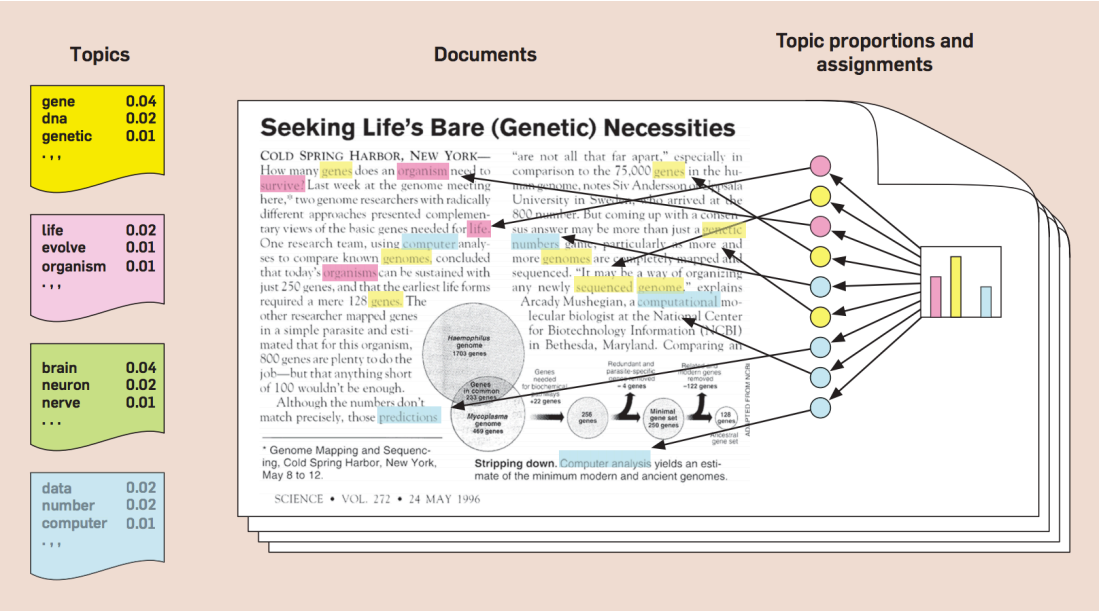

Figure 2.1: A graphical illustration of LDA, adapted from [4], where each document has a fixed topic distribution (right) and each word is then sampled from the topic distribution (left).

Hyperparameters in LDA

Although we don't see K in the generative process, it's implicitly there since θ is a matrix of probabilities of each topic in each document and ϕ is a matrix of probabilities of words in topics. So, one of the dimensions in any of these two matrices is the number of topics, K; therefore K should be known beforehand. In fact ϕ is a $K \times V$ matrix and θ is a $C \times K$ matrix where V is the size of vocabulary and C is the number of documents in D.

Also, α and β, that are vectors in general, usually are chosen to be symmetric vectors of a scalar (α or β) everywhere. In fact, before anything is observed, the probability of a word in a topic is β and the probability of a topic in a document is α.

Parameter inference in LDA

LDA is a generative model. It models how the documents are generated using a set of known document-topic distributions (θ) and topic-word distributions (ϕ). In reality however, we don't know these distributions. In fact, the whole point is to infer these parameters.

Before Grifths and Steyvers suggest Gibbs sampling for estimating the parameters of LDA in 2004, sophisticated approximations like variational Bayes [6] or expectation propagation [23] were employed to estimate ϕ [16]. Instead of estimating ϕ and θ directly, Griffith and Steyvers estimated the probability distribution of topic assignments for each word based on all other variables and extracted ϕ and θ from that distribution. In fact they show that

$$P(z_i = v \mid z_{-i}, w, \alpha, \beta) \propto \left(\frac{n_{-i,v}^{(d_i)} + \alpha}{\sum_u^K (n_{-i,u}^{(d_i)} + \alpha)} \right) \left(\frac{n_{-i,v}^{(w_i)} + \beta}{\sum_{w'}^V (n_{-i,v}^{(w')} + \beta)} \right), \tag{2.1}$$

where $n_{-i,v}^{(d_i)}$ is the number of times topic v appeared in document d_i, and $n_{-i,v}^{(w_i)}$ is the number of times it generated the word w_i. The $-i$ notation means that we do not take the i^th word into consideration when calculating the counts. The first part in the right hand side corresponds to the current estimation of θ and the second part corresponds to the current ϕ. So, the estimation of the parameters will be as follows.

$$\hat{\phi}_v^{w_i} = \left(\frac{n_{-i,v}^{(w_i)} + \beta}{\sum_{w'}^{V} (n_{-i,v}^{(w')} + \beta)} \right), \qquad \hat{\theta}_v^{d_i} = \left(\frac{n_{-i,v}^{(d_i)} + \alpha}{\sum_{u}^{K} (n_{-i,u}^{(d_i)} + \alpha)} \right). \tag{2.2}$$

2.1.2 Variants of LDA

Although LDA has proved useful in many applications, it also has limitations. For example the topics, which usually are presented by their most frequent words, are not easy to interpret and the model is not capable of guiding the topics towards the ones that are more interesting. For example in our case, we can't fix any of the topics as the health topic. Besides, LDA has unrealistic assumptions in many practical problems. For example, it assumes that K, the number of topics, is known before training or it assumes that the order of the words in documents or the order of documents themselves are not important (bag of words assumption and document exchangeability assumption). Here we mention a few works to address these limitations.

Adding Supervison

LDA models the most dominant topics in the observed corpus; hence, a particular topic of interest might not show up in the modeled topics. Also, there is no mechanism to use a labeled set of documents to guide the topics. Here we introduce three variants of LDA to address this problem. We will use one of these variants, LDA with topic-in-set knowledge (TLDA henceforth), in our seed word expansion method. So, we will explain it in more detail in section 4.2.1.

TLDA Andrzejewski and Zhu introduced a variation of LDA that incorporates domain knowledge about words into the model [2]. This model assigns a set of valid topics (subset of the set of all topics) to each word and limits the words to have probability of zero in all other topics. So, if we already know a few words related to a topic of interest, we can tell the model that these words should only belong to one of the topics (say, the first one). This way, we make sure that we know which topic in the model, corresponds to our topic of interest. Also, we make it more probable that this topic shows itself; even if it is

dominated by other topics in the corpus. Since we will use TLDA in this work, we will explain it in more detail in section 4.2.1.

Supervised LDA Blei and McAuliffe proposed another extension to LDA that models the generation of labeled documents with topics as hidden variables [3]. In the generative process, they sample not only the topics and the words, but also the document labels from a general form of exponential family. The general form of exponential family is a clever choice for modeling the labels since different members of this family can model different types of labels that usually accompany the documents. For example, we can model binary labels with Binomial distribution and categorial labels with Multinomial distribution; both of which are of the exponential family. Although, their experiments show the models power in label prediction, we can only use this model when a large labeled dataset is available. That is the reason we couldn't use this model for our purpose. Our set of labeled conversations is very small.

Labeled LDA Ramae and colleagues introduced another supervised version of LDA that, unlike supervised topic models [3], allows multiple labels for documents [26]. The model assumes that there are as many topics as labels and in the generative process, the possible topics of a document is restricted to its labels. Since the document labels are generated by the model, we would need a large labeled set of documents for training the model.

Relaxing the assumptions

We will mention five variants of LDA here that address the assumptions of LDA. HLDA will address the pre-known number of topics; Syntax integrated LDA and Syntactic LDA address the bag of words assumption; while Dynamic LDA and Syntactic LDA attempt to relax the document exchangeability assumption of LDA by taking time variable into account.

HLDA Griffiths and colleagues introduced a hierarchical version of LDA (HLDA henceforth) that doesn't need to know the number of topics beforehand [15]. HLDA models

the documents in a generative fashion similar to LDA. however, it models the set of topics as a hierarchy and allows branching to a not-previously-seen topic. Although HLDA outperforms LDA in document modeling and captures the hierarchical structure of topics and subtopics, it is limited in that it allows only one path in the topic hierarchy for each document. This means it can not generate documents with more than one major topics.

Syntax integrated LDA Griffith and colleagues introduce a composite model that relaxes the bag-of-words assumption [17]. They distinguish between so-called content words (words generated by topics) and function words (words with a syntactic role). They model the content words as sampled from k topics and the function words as sampled from c functionality classes. The transition between classes is modeled with a probabilistic Finite State Automaton where one state is set aside for generating function words in which LDA is generating the words from k topics. In other states, function words are sampled from multinomial distributions that correspond to the c classes. A limitation of this model is that it assumes the number of classes and the number of topics are known a priori.

Syntactic LDA Boyd-Graber and Blei introduced a sophisticated model that is a combination of probabilistic topic modeling, probabilistic syntax modeling and Bayesian nonparametric modeling to capture both thematic and syntactic characteristics of the text while allowing new themes as new data arrives [7]. Like LDA, topics are distributions over words and documents have distributions over topics. Here however, each sentence has a parse tree; and each element in a parse tree, has also a distribution over topics of its children. To assign a topic to a word, the model convolves (i.e. multiplies and renormalize) the document distribution over topics and the element distribution over topics and samples the topic from this new distribution. The authors also provide solution to the parameter estimation problem. While this model is powerful in that it ensures both thematic and syntactic consistency, the downside is that unlike LDA, the observed data (models input) is not the collection of documents. Instead, this model needs not only parse trees of the sentences of the documents, but also their dependency-parsed trees.

Dynamic LDA Blei and Lafferty introduced dynamic topic models to track topic changes over time [5]. They partition the documents into time slices and model documents of each time slice by a k-topic LDA-like topic modeling component and evolve the topics of each time slice from the corresponding topics in the previous time slice. I called each component of their model LDA-Like because they look exactly like LDA except for the fact that they do not use Dirichlet distribution as priors for distributions over topics. The authors explain that since the Dirichlet is not amenable to sequential modeling it can not be used as the prior distribution in their model. instead, they evolve the mean of each topic, which is a Multinomial distribution, over time with a Gaussian noise (See equation 1 in the paper). Although the authors provide two algorithms for estimating the parameters of the method, since the Gaussian distribution is not a conjugate to Multinomial distribution, the parameter estimation for the model is not as simple as it is in LDA.

On-line LDA AlSumait and colleagues introduce On-line LDA to address the same problem as dynamic topic models do [1]. Their approach is similar to dynamic models in that they partition the documents into time slices and evolve topics over time. The difference however, is in choosing the priors. Instead of using a Gaussian distribution to evolve the topics they use the current topic distributions, themselves; and thus, can provide an easy solution for parameter estimation. They also describe an algorithm for detecting emerging topics, which is also pluggable to dynamic models. This algorithm computes the KL Divergence (a measure of distance between distributions) of two corresponding topics in consecutive time slices (e.g. the first topic in time slice t and the first topic in time slice $t + 1$).Then, it nominates the topic in the later time slice as emerging if this distance is more than, say, ninety percent of the pair distances.

2.2 Applications of Topic Modeling on Health-Related Data

Here, we present the health-related applications of topic modeling on textual data in four categories: structured and unstructured clinical data, as well as public and private non-clinical data. Although there has been health-related applications of topic modeling on

non-text data (*ex.* on radiology images [9]), here we only cover applications on text or transcribed speech.

2.2.1 Structured Clinical Data

In this section, we will briefly explain how topic modeling can help provide security for Electronic Medical Records (EMR), detect fraud in healthcare claims, analyze patient pathway traces, and retrieve more related educational documents for patients based on their EMR notes. All these data are considered structured although in the last application, the method uses the EMR as free text.

Healthcare claims

Chandola and colleagues apply LDA on the healthcare (insurance) claim data to create profiles for providers to detect fraud [10]. Figure 2.2 shows the entities and their relationships. Chandola and colleagues transform the data to a format for which LDA is applicable. They make 6 matrices of the form XY where X belongs to provider, beneficiary and Y belongs to Diagnosis, Procedure, Drug. Then they treat providers and beneficiaries as documents while different diagnoses, procedure, and drugs are treated as words in documents. Their results show that the topic distributions are different between fraudulent and non-fraudulent providers.

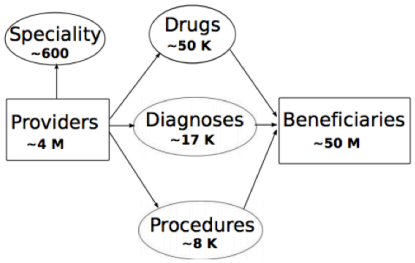

Figure 2.2: Entities and relationships in healthcare claims data along with approximate number of entries in each entity set. [10]

Security of Electronic Medical Records

Gupta and Siddharth attempt to detect illegitimate access to patients' data in a hospital [18]. The challenge is, electronic medical records have role-based security and since

no employee should be denied access when the data is important for patients' health, the security enforcement is weak. However, all access transactions are logged for future reference. But this can't be enough by itself because the log file is too big for human screening. Gupta and Siddharth apply LDA to medical records to find a set of topics and suggest monitoring the access to the set of topics instead of access to the records. For example, a nurse in a stroke unit may regularly access neurological diagnoses. However, when he/she accesses unrelated topics, say obstetrics, then the system should flag it as an anomaly.

Information retrieval by looking at EMR

A very interesting application of topic modeling is the work of Kandula and colleagues [21]. They combine two different sets of data. The electronic medical record of patients suffering from diabetes and educational documents on diabetics that are available through different resources . Since all patients are diabetics and all educational documents are on diabetes, one could say that all documents are related to all patients. But the goal of this study is to find best documents for patients with a specific problem or using a specific treatment. Kandula and colleagues apply LDA on all the educational documents, then they calculate a topic coverage vector for the EMR notes based on the td-idf of words in their content and the frequency of those words in each topic. Finally, they match documents and EMR notes by their topic coverage similarity.

Patient pathway trace analysis

Huang and colleagues apply LDA on patient pathway traces (a sequence of events happened for a patient including admission, tests, etc.) and use the cosine similarity between LDA topics' probability vectors as a similarity measure between two patient traces [20]. Note that, it is meaningful to use LDA in this context because there actually exist latent variables, treatment behaviors, which decide the clinical event types in each patient trace. For applying LDA, each patient trace is considered a document, clinical event types are considered words and the final topics are interpreted as treatment behaviors. Figure 2.3, shows patient trace samples and topic samples. They use this similarity measure along with conventional similarity measures, Euclidean distance (ED) and term vector (TV)

i.e. the TF-IDF vector cosine similarity, in three Clinical pathway analysis (CPA) tasks: patient trace retrieval, patient trace clustering, and anomaly detection. The results show that LDA-based similarity measure outperforms the conventional measures. It specially makes sense in the anomaly detection task where ED and TV similarities are very small and thus, many patient traces are falsely detected as anomalous.

c_1	⟨(Adm, 1), (EPT, 1), (EKS, 1), (EBT, 1), (OxS, 1), (Iso, 2), (Oxl, 2), (OxS, 2), (BT, 2), (Mic, 2), (Coa, 2), (Uri, 2), (Vei, 2), (EBT, 2), (EKS, 2), (Hig, 2), (Hep, 2), (Ele, 2), (Cat, 2), (Sto, 2), (Rep, 3), (Rep, 3), (IhH, 4), (OxS, 4), (Vei, 5), (Pos, 6), (Ind, 6), (Rep, 6), (Vei, 7), (Oxl, 7), (BT, 8), (Hep, 8), (Ele, 8), (Vei, 9), (Vei, 9), (Vei, 11), (Vei, 11), (KAR, 14), (Osm, 16), (Oxl, 16), (Vei, 16), (Hep, 16), (Ele, 16), (Vei, 18), (Dis, 18)⟩
c_2	⟨(Adm, 1), (Con, 1), (Sto, 1), (Cat, 1), (Ele, 1), (LKF, 1), (Sil, 1), (EPT, 1), (EKS, 1), (EBT, 1), (Thy, 1), (Uri, 1), (Coa, 1), (Vei, 1), (Sex, 1), (BT, 1), (OxS, 1), (Iso, 1), (InH, 2), (Oxl, 3), (Vei, 3), (Pos, 5), (Rep, 5), (Lum, 6), (aCB, 6), (rCF, 6), (Lum, 7), (CFA, 7), (rCF, 7), (Lum, 8), (BT, 8), (CFA, 8), (rCF, 8), (CSF, 8), (aCC, 8), (Hep, 8), (Ele, 8), (Oxl, 11), (OxS, 11), (Ind, 11), (Rep, 11), (Dis, 11)⟩
c_3	⟨(Adm, 1), (Det, 1), (Tum, 1), (Oxs, 1), (Hem, 1), (Hem, 1), (Ane, 1), (Ane, 1), (Coa, 1), (Coa, 1), (Ana, 1), (Ana, 1), (Uri, 1), (Uri, 1), (Vei, 1), (Thy, 1), (Thy, 1), (EBT, 1), (EKS, 1), (EKS, 1), (EUS, 1), (EPT, 1), (EPT, 1), (LKG, 1), (LKG, 1), (Sto, 1), (Sto, 1), (Oxs, 2), (Vei, 2), (Con, 2), (Oxs, 3), (ERS, 5), (FAC, 7), (Oxs, 7), (Oxs, 8), (BTH, 10), (Ele, 10), (LKF, 10), (Hem, 10), (Det, 11), (CDR, 11), (Dis, 18)⟩

Topic 1	Admission; Oxygen saturation monitoring; Vein catheterization; Conventional ECG Exam; Blood test; Coagulation + D-dimer; Electrolyte; Emergency blood test; Emergency kidney and sugar; Emergency PT; Liver, kidney and the glycolipid heart enzyme (hospitalization); Urine + sediment test; The hepatorenal sugar (hospitalization); Thyroid function (7 items); Emergency ultra-sensitivity CRP; Stool examination + OB; Tumors (10 items); Blood test + Hypersensitive CRP; Hemorheology; Discharge
Topic 2	Oxygen saturation monitoring; High-frequency oxygen/hour; Vein catheterization; Intracranial hematoma surgery (including simple epidural); Postoperative drainage; Indwelling catheter; Replacement of drainage bag; Lumbar puncture; Blood test; Electrolyte; Stool examination; Oxygen inhalation; Cerebrospinal fluid biochemical; CSF routine; The hepatorenal sugar (hospitalization); Osmotic pressure; Bacteria and fungi were cultured and identified; Blood test + Hypersensitive CRP; Hemorheolopgy; B-D Heparin cap

Figure 2.3: Patient trace samples (above) and topic samples (below). [20]

2.2.2 Free-Text Clinical Data

The following examples are applications of topic modeling on free textual data that is collected in clinical settings for diagnosis.

Predicting depression and neuroticism

Resnik and colleagues extract language features from students' essays to predict Neuroticism and Depression [27]. They asked college students to write their thoughts and feelings as they come to them. Then they counted the words in each of the 64 LIWC categories as one set of features for each essay and then expanded the feature set by adding topic probabilities extracted from applying LDA on the data set. They conclude that the expanded feature set increases the prediction power significantly. They also investigated the topics and found out that 7 of them were meaningfully related to depression.

Detecting language impairment

One test for detecting language impairment in children is to provide the child with a wordless picture story-book and ask him/her to narrate the story. Different aspects of

language, such as coherence, is annotated traditionally be human annotators, and these annotations are used for diagnosis. Hassanali and colleagues use topic modeling of the child narratives to find meaningful topics (and subtopics) and develop features that can indicate coherence and language impairment [19]. Even if the performance was only as well as the human annotators, the fact that this automates the process is valuable. But the results show improvement over human annotations due to high coverage of topics induced by LDA.

2.2.3 Online Health-Related Textual Data

There are all sorts of health-related texts online. In forums and online communities people post information about their own (or other peoples) health problems and sometimes they share information in these somewhat anonymous settings that they are reluctant to share elsewhere. Others can comment on the posts with information on how to treat/handle the problem or they can comment to be supportive. The online information could also be a good source for analyzing public health. we talk about these applications in this section.

Social support for people with mental health issues

De Choudhury and colleagues investigate the social support people can get about mental health issues [13]. The authors looked at posts in reddit (a social media where accounts are temporary; so, users tend to share more uninhibited feelings or socially unacceptable thoughts) on mental health related subdomains (alcoholism, anxiety, bipolarreddit, depression, mentalhealth, MMFB (Make Me Feel Better), socialanxiety, SuicideWatch) and investigate the characteristics of the posts themselves and the number of them as well as type of the comments the get. For categorizing the comments, they used LDA with 20 topics and then, human annotators identified the topics that bore prescriptive, informational, instrumental, and emotional information. They observed that there were more emotional (36%) and prescriptive comments (32%), than informational (13%) or instrumental (19%) ones; and that the difference was significant.

Tweet analysis

Tuarob and colleagues classify the tweets into health-related vs. not health-related [32]. They also incorporate health-related information (using 3 vocabularies: disease, symptom, and anatomy) and sentiment bearing information (using SentiStrength algorithm proposed by Thelwall and colleagues [31]) along with N-Gram features and topic model probability vectors. They take advantage of all these features by 1) having a classifier trained on the combination of features and 2) via ensemble learning, i.e. combining the final decision of different classifiers (each using one group of features). They find out that the best results are obtained when the weight of the classifiers are as follows. $\langle NGrams : 0.1; DomainAware : 0.2; TopicModeling : 0.1; SentimentRelated : 0.1; Combined : 0.5\rangle$ Interestingly, the combination of the features seems to be heavily more informative than any of the other feature sets alone. However, it could also have been interesting if they had reported results in absence of any feature set.

Modeling tweets for public health analysis

Paul and Dredze propose an extension of LDA, Ailment Topic Aspect Model (ATAM), that can model all twitter messages into health-related topics with general, symptom, and treatment words (aspects) and also allows for a background topic to capture words that are not health-related [24]. Figure 2.4 shows examples of the health-related topics they found in twitter messages. They performed Mechanical Turk experiments to show that the topics extracted from their method where more coherent and interpretable for human in comparison with topics induced by LDA. They also suggest that their model can be used in syndromic surveillance since it can discover ailments in twitter.

In a subsequent work in 2011, they extract knowledge about disease specific vocabulary from published articles and incorporate this knowledge into ATAM by changing the prior distribution of health-related topics. They use this method (which they call ATAM+) for syndromic surveillance (tracking temporal impact on medical well being), analyzing geographic behavioral risk factors by categorizing the tweets based on users' location (in US states), and finally combining the two to come up with geographic syndromic

	Allergies	Insomnia	Obesity	Injuries	Respiratory	Dental	Aches/Pains
General	allergies	sleep	blood	knee	throat	ow	body
	nose	asleep	weight	leg	stop	teeth	need
	eyes	fell	eat	right	better	tooth	neck
	allergy	awake	healthy	ankle	voice	wisdom	hurts
	allergic	hours	fat	shoulder	hurts	dentist	head
Symptoms	sneezing	insomnia	pressure	pain	cough	pain	aches
	coughing	fall	weight	sore	coughing	toothache	pain
	cold	burning	loss	arthritis	cold	sore	sore
	nose	pain	blood	limping	sneezing	infection	muscle
	runny	falling	high	neck	sneeze	tooth	aching
Treatments	medicine	sleeping	diet	surgery	medicine	braces	massage
	benadryl	pills	exercise	brace	antibiotics	pain	exercise
	claritin	caffeine	dieting	crutches	codeine	relief	massages
	zyrtec	tylenol	insulin	physical	vitamin	muscle	bath
	drops	pill	exercising	therapy	tylenol	surgery	hot

Figure 2.4: Example output of the most likely words for ailments from the Ailment Topic Aspect Model. [24]

surveillance. They also use their model for analyzing the relationships between ailments and their symptoms or popular treatments based on the tweets. The motivation for that is the fact that for many health problems (like flu or obesity) people don't go to health care providers. As a result information about occurrence of these problems, their symptoms, and the treatments people will not appear in clinical reports.

2.2.4 Every-day Private Telephone Conversations

While online data usually provide information about the whole population, the health status of individuals may be tracked via their private data such as their every-day telephone conversations. Topic modeling can help get the gist of the conversations without violating the privacy of the conversant. Stark and colleagues recorded the private conversations of 56 older adults over periods of 6 months to one year and used topic modeling to extract topic probability features for classifying the conversations into social or business. The motivation is to measure the social engagement of senior citizens and relate this measurement to presence/absence of depression and dementia. [30] The work presented in this thesis is the sequel of this work.

Chapter 3

Data

Everyday Telephone Conversations Our corpus consists of everyday telephone conversations recorded in two phases from older native English speakers. Initially, conversations were recorded from 10 volunteers, 79 years or older, over a period of approximately 12 months and consists of about 12,000 digitized land-line telephone conversations. Subsequently, we recruited more subjects and recorded about 29,000 digitized land-line telephone conversations from 46 volunteers, 65 years or older, over a period of approximately 6 months. In addition to the speech, our corpus includes a rich set of meta-data, such as call direction (in- coming vs outgoing), time of call, duration and caller ID when available.

Extra information on subjects' health In the everyday telephone conversations corpus, we also have access to the subjects' weekly self-reports on their medical status during the week indicating medical events such as injuries or going to emergency rooms, which we use to relate the health-related conversations to actual medical events in the subjects' lives.

Labeling Data As mentioned previously, the corpus was collected with the assurance that the recordings will not be manually transcribed or annotated. For developing and evaluating our algorithms, we identified conversations using reverse look-up of caller ids. We fed the phone numbers into the reverse phone service of "whitepages.com" and retrieved tags it returned under the "search related" part. Calls associated with tags such as "hospital" and "pharmacy" were labeled as medical and those with tags like "car repair"

and "real estate" were marked as non-medical. While this may seem a straight forward method to classify the conversations, the reverse look-up has severe limitations.

- Caller ids are not very reliable. Many retirement communities have legacy telephone systems that do not support caller id. The DTMF tones of outgoing calls are noisy.
- The reverse look-up can be ambiguous, for example, those associated with the tags such as "insurance", which were of course discarded in our labeled subset.
- The reverse look-up is only applicable for calls associate with medical and non-medical businesses. It can be used for detecting which personal conversations are about health.
- Even for businesses, these tags are not always available.

In our corpus, we were only able to label 420 out of 29000 conversations – 137 and 283 were medical and non-medical respectively. We divided the labeled set further into training and evaluation set, consisting of 105 vs. 218 and 32 vs. 65 respectively.

Fisher Test Set In order to understand how well our approach generalizes, we also evaluated our algorithm on a subset of a much larger corpus, the Fisher corpus [12].In Fisher, strangers spoke on 40 different topics, of which only one relates to medical (illness). We identified 338 conversations on illness as medical conversations and sampled 702 conversations from the other 39 topics to build the non-medical subset.

Automatic Speech Recognition System Conversations in our corpus were automatically transcribed using an ASR system. Our ASR system is structured after IBMs conversation telephony system [29]. The acoustic models were trained on about 2000 hours of telephone speech from Switchboard and Fisher corpora [14]. The system has a vocabulary of 47K and uses a trigram language model with about 10M n-grams, estimated from a mix of transcripts and web-harvested data. Decoding is performed in three stages using speaker-independent models, vocal-tract normalized models and speaker-adapted models. The three sets of models are similar in complexity with 4000 clustered pentaphone states

and 150K Gaussians with diagonal covariances. Our system does not include discriminative training and performs at a word error rate of about 24% on NIST RT Dev04 which is comparable to state of the art performance for such systems. The privacy requirements in place for our corpus prohibit human listening precluding the transcriptions needed reporting recognition accuracy. However, while our corpus differs from Switchboard, we expect the performance of the 2000 hour recognizer to be relatively close to results on NIST benchmark.

Chapter 4

Method

4.1 Overview

As we explained in previous chapter, we can label a small set of conversations in the everyday telephone conversations corpus as health-related vs.ñon-health related. Using this labeled set we can train a support vector machine (SVM) to classify the conversations. The conversations are represented by a vector of tf-idf scores of words where tf-idf is a score for measuring the importance of a word in one document of a corpus. We used the following formula for calculating tf-idf:

$$\frac{count(t,d)}{\sum_{t'} count(t',d)} log(\frac{|D|}{|\{d \in D | t \in d\}|}) \tag{4.1}$$

where t and t' are word types, d is a document, $count(t,d)$ is the number of times t appears in d, and $|D|$ is the number of documents in corpus D.

As we will see in the experiments, such a classifier doesn't generalize to the out-of-domain Fisher task. Generalizability is important in our case, especially because the data we use for training is limited in number and the nature of conversations.

One way to improve generalization is to perform feature selection. That is, instead of using tf-idf scores for the whole vocabulary, we would like to rely only on features relevant to detecting the health topic. The idea is to specify a handful of words that exemplify the topic, then expand this set of words using TLDA and use them as features to represent conversations. In our task, the topic of interest is health. So, we choose *doctor, medicine,* and *pain* – often used while discussing health – as our seed words. Topics in LDA based

methods such as TLDA are usually represented using the n most probable words; where n is an arbitrary number. So, the first candidate sets for expanding our seed words are the sets of 50 most probable words in the topic of health in different runs of TLDA. As our experiments reveal, these candidate sets contain many words that are unrelated to health . To solve this problem, we use the small labeled set of conversations to filter out the unrelated words. Thus, our approach is algorithmic and is not focused on optimizing an overall cost function.

4.2 Iterative Seed Words Expansion

We propose a method to iteratively expand a small set of initial seed words. One simple approach for expanding the set is to identify words which have a high co-occurrence with the seed words. This is easy to compute for pairwise co-occurrences but quickly becomes infeasible when co-occurrence between more number of words needs to be computed. Instead, we adopt TLDA, briefly reviewed in 4.2.1, for identifying words topically related to the seed words as illustrated in Figure 4.1.

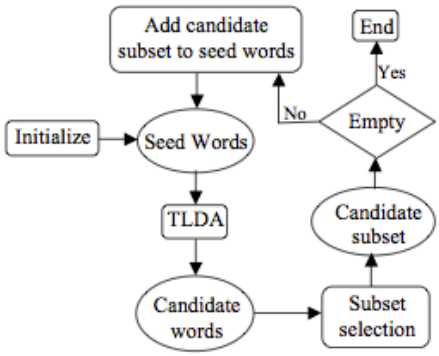

Figure 4.1: Expanding the set of seed words

The iterative algorithm starts with initializing the seed words to a few words that exemplify the topic.. In each iteration, TLDA, explained in 4.2.1, does the topic modeling and returns the 50 most probable words in the supposedly health topic as the candidate words. A subset of these candidate words are selected as described in 4.2.2 to be added

to the seed words. The algorithm terminates when we decide in subset selection that no word from the candidate words should be added to the seed words.

4.2.1 TLDA

As we mentioned in the Background Chapter, TLDA is a semi-supervised variant of LDA, introduced by Andrzejewski and Zhu in 2009. TLDA adds domain knowledge to the model by assigning valid topics to the words.

TLDA uses $C^{(i)}$ as the set of valid topics for $z^{(i)}$, the latent topic for $w^{(i)}$ and adds a hard constraint to the Gibbs sampling by multiplying the right hand side of equation 2.1 to an indicator function $\delta(v \in C^{(i)})$:

$$\delta(v \in C^{(i)}) = \begin{cases} 1 & v \in C^{(i)} \\ 0 & \text{otherwise.} \end{cases}$$

We use TLDA by forcing the seed words into the first topic only while other words can fall into any topic. This means we defined $C^{(i)}$ as follows:

$$C^{(i)} = \begin{cases} \{1\} & \text{if } w^{(i)} \in SeedWords \\ \{1, 2, ..., T\} & \text{otherwise.} \end{cases}$$

This will ensure that the first topic will represent the topic of interest, health. We train TLDA using all 47000 unlabeled conversations, and chose the number of topics, K, to be 20.

The 50 most probable words of the first topic (health), are the candidate words that serve as the input of subset selection. Table 4.1 shows a set of candidate words returned by TLDA in the first iteration of one run of our algorithm with our three health-related initial seed words *doctor*, *medicine* and *pain*.

As the Table 4.1 shows, the TLDA picks up several words (in bold) that are highly related to our topic of interest. But along with those, three are many words that are clearly unrelated and several that are not unambiguously related to health.

pain, **medicine**, appointment, **medical**, **doctors**, **emergency**, **prescription**, contact, **medication**, dial, **insurance**, **pharmacy**, schedule, moment, reached, questions, services, **surgery**, telephone, record, appointments, options, address, **patient**, advice, quality, tuesday, position, answered, records, wednesday, **therapy**, **healthy**, correct, department, ensure, numbers, act, **doctor**, personal, test, senior, **nurse**, plan, **kaiser**, reach, threatening, transportation, cats, women

Table 4.1: One set of candidate words identified by TLDA.

4.2.2 Subset Selection

Having the candidate words list in hand, we need to extract the best words from it and add them to our current seed words. We do this in a greedy iterative manner based on the performance of a simple classifier that classifies the documents based on presence of the key words, essentially a key word spotter (KWS) . KWS classifies a document as a hit if and only if it contains any of the current seed words. Assuming that S is the set of seed words, we define the performance of KWS as

$$F(S) = \alpha Recall(S) + (1 - \alpha) Precision(S). \tag{4.2}$$

Then in iteration j of the subset selection, pick a word w_j from the candidate words that is not already a seed word and increases the KWS performance the most if added to the seed words.

$$w_j = argmax_{w \in CW - S}\{F(S \bigcup \{w\})\} \tag{4.3}$$

Subset selection terminates when for no candidate word cw, $F(S) < F(S \bigcup \{cw\})$.

Table 4.2 shows the effect of α in equation 4.2. The words are the ones selected by one iteration of the iterative seed word expansion method shown in Figure 4.1. For smaller αs more relevant words are chosen whereas for bigger αs because of the importance of recall, even not so relevant words are selected. So, while bigger α emphasizes recall, smaller α is better for applications that precision is more important. In our experiments, presented in the next chapter, we set α to 0.5.

α	Selected Words
0.1	medication, appointments, emergency, patient
0.3	medication, appointments, doctors, emergency, patient
0.5	doctors
0.7	prescription, blood, doctors, appointments, appointment, problems, surgery, medication, emergency, patient
0.9	prescription, appointment, insurance, blood, doctors, problems, calls, monday, surgery, feeling, schedule, information, medication, emergency, nurse, appointments, physical, patient

Table 4.2: The effect of α in equation 4.2 in one run of TLDA.

Chapter 5

Experiments

In this chapter, we present the classification results of an SVM using tf-idf vectors over the whole vocabulary, or over smaller sets of words. we train the SVM classifier with the relevant features using the popular libSVM [11] implementation. We chose the parameters of the SVM using a 30-fold cross-validated (CV) grid search over the training data. We use a 4-fold cross validation over the labeled set of conversations to maximize the use of the relatively small labeled set. That is, we trained the feature selection algorithm on 3-folds and tested the resulting SVM tested on the fourth. In the in-domain task we always report the average performance across the folds.

Table 5.1 shows the results of our experiments using different input features. We report recall, precision and F-measure on the in-domain and the out-of-domain (Fisher) tasks. In addition, we report average F-measure across both tasks. The justification for considering the average F-measure is that we want our algorithm to work well on both in-domain corpus and Fisher corpus since we need to make sure that our classifier is generalizable (i.e. it works well on Fisher) and it works well on the private and natural telephone conversations (i.e. the ones similar to the in-domain corpus)

When using the full vocabulary, the in-domain performance (the performance on the everyday telephone conversations data) is relatively good with 75.1% recall and 83.5% precision. But the out-of-domain recall (recall on the Fisher data set) is considerably low at 2.8%. Ideally, we want a classifier that performs well in both domains. Rows 2 to 5 can be seen as steps to get to such a classifier.

The second row shows the performance of the other extreme end of feature selection: the features include the manually chosen words *doctor, medicine, and pain* only. While

	Recall		Precision		F-measure		
Feature Words	In-Domain	Fisher	In-Domain	Fisher	In-Domain	Fisher	Avg.
Full vocabulary (no feature selection)	75.2	2.8	83.5	91.1	79.1	5.4	42.3
Initial words (*doctor, medicine, pain*)	45.1	69.2	94.8	94.5	61.1	79.9	70.5
50 most probable words in *health* returned by TLDA (average over 30 runs)	58.4	57.4	86.3	97.5	69.7	72.3	71.0
Words selected by our iterative method (average over 30 runs)	56.1	66.5	91.0	95.5	69.4	78.4	73.9
Union of all words selected by our method (across 30 runs)	67.7	69.4	87.8	95.1	76.5	80.2	78.3

Table 5.1: Performance of SVM classifiers using different feature selection methods.

this leads to very good out-of-domain performance, the in-domain recall has dropped considerably. We trained TLDA 30 times, and selected the 50 most probable words in the health topic. The third row in Table 5.1 shows the average performance of SVM when using the tf-idf of these sets of words as the feature vector on in-domain and out-of-domain tasks. Using the 50 most probable words in health topic significantly improves average F-score (71%) across both tasks over using the full vocabulary (42.3%) but it is clear that this is only due to improvement in out-of-domain task. As table 4.1 shows, a set of the 50 most probable words in health topic (the result of one run of TLDA) contains many irrelevant words. This is the motivation for our iterative algorithm.

Next, we evaluate the performance of our iterative algorithm. The fourth row in Table 5.1 shows the average performance of SVM using expanded seed words that our algorithm suggested in 30 runs. Our algorithm improves the average F-score by 3% comparing to the standard TLDA. This is due to a 5% improvement in out-of-domain task as opposed to a 0.3% performance decrease in the in-domain task.

Since our algorithm has a probabilistic topic modeling component (*i.e.* TLDA), different runs lead to different sets of expanded seed words. We extract a union of all the words chosen over 30 runs and evaluate the performance of SVM using this union set. This improves the performance of our method further to achieve the best average F-score

of 78.3%, which is an 85% improvement over using the SVM with full vocabulary. The in-domain performance is still lower than the full-vocabulary baseline by less than 3% while the out-of-domain performance is the highest. Once again, we are more interested in the average F-measure because we need our algorithm to generalize well (work well on out-of-domain corpus) and to work well on natural private conversations (on the conversations similar to the on-domain corpus).

Another experiment we ran was to study the effect of different sets of initial seed words. Table 5.2 shows the results of this experiment. The first column shows the initial set of seed words; the second column shows the average of F-Measure of SVM over the two corpora; the third column shows the same measure,averaged over 20 runs; and the forth column presents the number of runs that actually added some word to the seed words. As we can see, it does matter what seed words we start with. When starting from doctors,nurse, and sick, all the runs of the algorithm added to the initial sets of seed words and the average F-Measure got improved considerably. Whereas only one or two runs starting from doctor, medication, and blood; or from doctors, nurse, and sick lead to expanded sets of seed words. Also, in the last two cases shown in Table 5.2 , the expansion algorithm is hurting the average F-Measure.

Initial Seed words	Initial Avg-F	Final Avg-F (Average over 20 runs)	No. of Expanded Sets (out of 20 runs)
doctors, nurse, sick	52.7	70.9	20
doctor, medication, blood	73.1	74.4	2
prescription, emergency, surgery	38.2	41.8	20
doctors, nurse, hospital	72.8	67.9	20
doctor, blood, sick	76.6	75.8	1

Table 5.2: The effect of different initial seed words .

To study the relevance of the words selected by the algorithm, we present the output of the expansion algorithm in Table 5.3. Iterations are separated by "/".

doctor, medicine, pain/ doctors/ medication, appointments, blood, located, emergency, patient
doctor, medicine, pain/ medication, doctors, appointments, emergency, patient
doctor, medicine, pain/ blood, doctors, medication, sick/ mother
doctor, medicine, pain/ blood, doctors

Table 5.3: Seed words selected by different runs of our algorithm.

Finally, we test the statistical relationship between health-related discussions in everyday telephone conversations, and actual medical events in older adults. As mentioned in Chapter 3, we have access to weekly self-reports on medical events for subjects' in everyday telephone conversations corpus. We used our best classifier, the SVM with union of expanded seed words, to classify all the conversations in our corpus into health-containing and health-free conversations. We then mark each conversation as temporally near a medical event if a reported medical event occurred within a 3-week time window. We chose a 3-week window to allow for one report before and after the event. The focus of this analysis is to examine whether there is a change (rise) in the conversation on health around medical events.

Table 5.4 shows the number of conversations in different categories. We can see that 90.9% of the conversations are classified as health-related but this percentage is slightly different for conversations near medical events: 91.5% vs. 89.1% for the other conversations. This slight difference is significant according to χ^2 test of independence ($\chi^2(df = 1, N = 47288) = 61.17, p < 0.001$).

near a medical event	Classified as health-related	Classified as non-health-related
yes	1348	11067
no	2964	31909

Table 5.4: Number of health-related or not-health-related telephone conversations close to vs. far from actual medical events.

Chapter 6

Conclusion

Having a dataset of private natural telephone conversations of older adults, we proposed a method to expand a set of initial seed words for an arbitrary topic of interest (in our case health) using a small labeled subset of the conversations. Then we used the tf-idf of this expanded set of seed words as features of a support vector machine and showed that these feature sets work almost as well as the tf-idf of all words in the vocabulary on our dataset and also can generalize better to an out-of-domain corpus that does not participate in training.

The generalization is very important in our task for two reasons. First, the labelled set of conversations we use for training is very small and consists of only business calls; therefore it might not be a good representative of the whole dataset. Also, the labels can be noisy since they are obtained heuristically based on the phone number of the other conversant; thus, evaluation on another labeled set where the labels can be verified is of interest. These two reasons are the justification for evaluating our method on a public corpus (Fisher [12]).

The topic of interest in our work is health. The ultimate goal of this work, Stark and colleagues' work, and future work on this dataset is to come up with new tools for home-based physical or mental health monitoring of older adults. We used the extra information we had on the subjects of this study to show that there is a significant relation between the telephone conversations containing health-related discussions and actual medical events in the lives of the subjects.

6.1 Limitations of the conclusion

There are a few things that we should take into account about how reliable the conclusions are. Firstly, the population in this study have been selected based on a health-status criteria. This means that our subjects are probably healthier from the average and thus, might not be a good representative of the whole elderly population. Furthermore, all the conversations in our corpus were landline telephone conversations. Obviously, with the increasing use of cell-phones among older adults, we might have missed a noticeable part of our subjects' conversations. Finally, although the χ^2 test shows a significant relationship between the predicted label of conversations and their proximity to medical events in the lives of the subjects, the effect size is small.

Chapter 7

Future Work

Here we mention a few ideas for improving our method, for applying this method to another health-related problem and also for applying topic modeling methods or other techniques on this type of data to answer other health related questions.

One problem with our method is that we look at what both conversants are saying. While hearing about others' health-related issues might influence an individual's mental state (causing anxiety for example), these discussions are not likely to be related to the individual's medical events like sickness or going to the emergency room. Thus, it might be a good idea to perform speaker identification on our telephone conversations and stick to what the subject is saying.

Another idea could be using the words "forget" or "remember" on conversational data from patients with Alzheimer's disease or mild cognitive impairment and expand them to see if we can find Alzheimer's related words. This is something that Dr. Jeffery Kay (the director of the Layton Aging and Alzheimer's Disease Center, as well as director of the Oregon Center for Aging and Technology) mentioned in several meetings we had with him. There were two reasons we didn't try this. Firstly, none of our subjects had Alzheimer's disease and only one was diagnosed with mild cognitive impairment. Secondly, we did not have access to even a small labeled subset of the conversations with Alzheimer's related discussions. However, if we had more such patients, using topic modeling on the conversations with some health-related discussion might give us another small heuristically labeled set.

Another wealth of data that we have access to for the subjects of this study, is their answers to some social engagement assessment questionnaires before and after the data

recording. We could use this data to associate their social engagement with factors like number and duration of telephone conversations (specially the social ones as opposed to business ones based on Stark and colleagues' work [30]), or the number of topics they talk about. For obtaining the number of topics we could apply HLDA [15]. After applying HLDA we can also look at the coverage of topics in conversations to specify which conversations contain more specific topics (the topics in the lower parts of the hierarchy) and use the number of those specific conversations as another factor that can associate with the subjects' social engagement.

This is also worth mentioning that, if more data or other pieces of information are required for other questions, our colleagues in Layton Aging and Alzheimer's Disease Center have promised to provide them for us. So, in one sentence, the opportunities for working on this unique dataset seem endless.

Bibliography

[1] AlSumait, L., Barbará, D., and Domeniconi, C. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on* (2008), IEEE, pp. 3–12.

[2] Andrzejewski, D., and Zhu, X. Latent dirichlet allocation with topic-in-set knowledge. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing* (2009), SemiSupLearn '09, pp. 43–48.

[3] Blei, D., and McAuliffe, J. Supervised topic models. In *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds., vol. 20. MIT Press, Cambridge, MA, 2008, pp. 121–128.

[4] Blei, D. M. Probabilistic topic models. *Communications of the ACM 55*, 4 (2012), 77–84.

[5] Blei, D. M., and Lafferty, J. D. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning(ICML2006)* (Pittsburgh, Pennsylvania, USA, 2006), ACM, pp. 113–120.

[6] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of Machine Learning Research 3* (2003), 993–1022.

[7] Boyd-Graber, J. L., and Blei, D. Syntactic topic models. In *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., vol. 21. MIT Press, Cambridge, MA, 2009, pp. 185–192.

[8] Cai, J. F., Lee, W. S., and Teh, Y. W. Nus-ml: Improving word sense disambiguation using topic features. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (2007), Association for Computational Linguistics, pp. 249–252.

[9] Carrivick, L., Prabhu, S., Goddard, P., and Rossiter, J. Unsupervised learning in radiology using novel latent variable models. In *Computer Vision and Pattern*

Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (2005), vol. 2, IEEE, pp. 854–859.

[10] Chandola, V., Sukumar, S. R., and Schryver, J. C. Knowledge discovery from massive healthcare claims data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (2013), ACM, pp. 1312–1320.

[11] Chang, C.-C., and Lin, C.-J. Libsvm : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2* (2011).

[12] Cieri, C., Miller, D., and Walker, K. The fisher corpus: a resource for the next generations of speech-to-text. In *LREC* (2004), vol. 4, pp. 69–71.

[13] De Choudhury, M., and De, S. Mental health discourse on reddit: Self-disclosure, social support, and anonymity.

[14] Godfrey, J. J., Holliman, E. C., and McDaniel, J. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on* (1992), vol. 1, IEEE, pp. 517–520.

[15] Griffiths, T., Jordan, M., Tenenbaum, J., and Blei, D. M. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems*, S. Thrun, L. Saul, and B. Schölkopf, Eds., vol. 16. MIT Press, Cambridge, MA, 2004, pp. 106–114.

[16] Griffiths, T. L., and Steyvers, M. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America 101*, Suppl 1 (2004), 5228–5235.

[17] Griffiths, T. L., Steyvers, M., Blei, D. M., and Tenenbaum, J. B. Integrating topics and syntax. In *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. MIT Press, Cambridge, MA, 2005, pp. 537–544.

[18] Gupta, S. Modeling and detecting anomalous topic access in emr audit logs.

[19] Hassanali, K.-n., Liu, Y., and Solorio, T. Using latent dirichlet allocation for child narrative analysis. *ACL 2013* (2013), 111.

[20] Huang, Z., Dong, W., Duan, H., and Li, H. Similarity measure between patient traces for clinical pathway analysis: problem, method, and applications.

[21] Kandula, S., Curtis, D., Hill, B., and Zeng-Treitler, Q. Use of topic modeling for recommending relevant education material to diabetic patients. In *AMIA Annual Symposium Proceedings* (2011), vol. 2011, American Medical Informatics Association, p. 674.

[22] Liu, B., Liu, L., Tsykin, A., Goodall, G. J., Green, J. E., Zhu, M., Kim, C. H., and Li, J. Identifying functional mirna–mrna regulatory modules with correspondence latent dirichlet allocation. *Bioinformatics 26*, 24 (2010), 3105–3111.

[23] Minka, T., and Lafferty, J. Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence* (2002), Morgan Kaufmann Publishers Inc., pp. 352–359.

[24] Paul, M. J., and Dredze, M. You are what you tweet: Analyzing twitter for public health. In *ICWSM* (2011).

[25] Prier, K. W., Smith, M. S., Giraud-Carrier, C., and Hanson, C. L. Identifying health-related topics on twitter: an exploration of tobacco-related tweets as a test topic. In *Proceedings of the 4th international conference on Social computing, behavioral-cultural modeling and prediction* (2011), pp. 18–25.

[26] Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing(EMNLP 2009)* (Singapore, 2009), Association for Computational Linguistics, pp. 248–256.

[27] Resnik, P., Garron, A., and Resnik, R. Using topic modeling to improve prediction of neuroticism and depression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural* (2013), Association for Computational Linguistics, pp. 1348–1353.

[28] Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence* (2004), AUAI Press, pp. 487–494.

[29] Soltau, H., Kingsbury, B., Mangu, L., Povey, D., Saon, G., and Zweig, G. The ibm 2004 conversational telephony system for rich transcription. In *Proc. ICASSP* (2005), vol. 1, pp. 205–208.

[30] Stark, A., Shafran, I., and Kaye, J. Inferring social nature of conversations from words: Experiments on a corpus of everyday telephone conversations. *Computer speech & language 28*, 1 (2014), 224–239.

[31] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology 61*, 12 (2010), 2544–2558.

[32] Tuarob, S., Tucker, C. S., Salathe, M., and Ram, N. Discovering health-related knowledge in social media using ensembles of heterogeneous features. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management* (2013), ACM, pp. 1685–1690.