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

Golnar Sheikhshab, Izhak Shafran, Jeffrey Kaye
Oregon Health & Science University
sheikhsh,shafrani,kaye@ohsu.edu

Abstract

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 discussion of health related conversations is statistically associated with actual medical events obtained through weekly self-reports.

1 Introduction

There has been considerable interest in understanding, promoting, and monitoring healthy lifestyles among older adults while minimizing the frequency of clinical visits. Longitudinal studies on large cohorts are necessary, for example, to understand the association between social networks, depression, dementia, and general health. In this context, detecting discussions of health are important as indicators of under-reported health events in daily lives as well as for studying healthy social support networks. The detection of medical events such as higher levels of pain or discomfort may also be useful in providing timely clinical intervention for managing chronic illness and thus promoting healthy independent living among older adults.

Motivated by this larger goal, we develop and investigate techniques for identifying conversations containing any health related discussion. We are interested in detecting discussions about medication with doctors, as well as conversations with others, where among all different topics being discussed, subjects may also be complaining about pain or changes in health status.

The privacy concerns of recording and analyzing everyday telephone conversation prevents us from manually transcribing and annotating conversations. So, we automatically transcribe the conversations using an automatic speech recognition system and look-up the telephone number corresponding to each conversation as a heuristic means of deriving labels. This technique is suitable for labeling a small subset of the conversations that are only sufficient for developing semi-supervised algorithms and for evaluating the methods for analysis.

Before delving into our approach, we discuss a few relevant and related studies in Section 2 and describe our unique naturalistic corpus in Section 3. Given the restrictive nature of our labeled in-domain data set, we are interested in a classifier that generalizes to the unlabeled data. We evaluate the generalizability of the classifiers using an out-of-domain corpus. We adopt a semi-supervised topic modeling approach to address our task, and develop an iterative feature selection method to improve our classifier, as described in Section 4. We evaluate the efficacy of our approach empirically, on the in-domain as well as an out-of-domain corpus, and report results in Section 5.

2 Related Work

The task of identifying conversations where health is mentioned differs from many other tasks in topic

Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP 2014), pages 38–44,
Baltimore, Maryland USA, June 26-27 2014.

modeling because in this task we are interested in one particular topic. A similar study is the work of Prier and colleagues (Prier et al., 2011). They use a set of predefined seed words as queries to gather tweets related to tobacco or marijuana usage, and then use LDA to discover related subtopics. Thus, their method is sensitive to the seed words chosen.

One way to reduce the sensitivity to the manually specified seed words is to expand the set using WordNet. Researchers have investigated this approach in sentiment analysis (Kim and Hovy, 2004; Yu and Hatzivassiloglou, 2003). However, when expanding the seed word set using WordNet, we need to be careful to avoid antonyms and words that have high degree of linkage with many words in the vocabulary. Furthermore, we can not apply such an approach for languages with poor resources, where manually curated knowledge is unavailable. The other drawback of this approach is that we can not use characteristics of the end task, in our case health-related conversation retrieval, to select the words. As an alternative method, Han and colleagues developed an interactive system where users selected the most relevant words from a set, proposed by an automated system (Han et al., 2009).

Another idea for expanding the seed words is using the statistical information. Among statistical methods, the simplest approach is to compute pairwise co-occurrence with the seed words. Li and Yamanishi ranked the words co-occurring with the seed words according to information theoretic costs, and used the highest ranked words as the expanded set (Li and Yamanishi, 2003). This idea can be more effective when the co-occurrence is performed over subsets instead, as in Hisamitsu and Niwa's work (Hisamitsu and Niwa, 2001). However, it is computationally expensive to search over subsets of words. Depending on the language and task, heuristics might be applicable. An example of this kind of approach is Zagibalov and Carroll's work on sentiment analysis in Chinese (Zagibalov and Carroll, 2008).

Alternatively, we can treat the task of identifying words associated with seed words as a clustering problem with the intuition that the seed words are in the same cluster. An effective strategy to cluster words into topics, is Latent Dirichlet Allocation (LDA) (Blei et al., 2003) . However, LDA is an unsupervised algorithm and the clustered topics are not guaranteed to include the topic of interest. The Seeded LDA, a variant of LDA, attempts to address this problem by incorporating the seed words as priors over the topics (Jagarlamudi et al., 2012). However, the estimation procedure is more complicated. Alternatively, in Topic LDA (TLDA), a clever extension to LDA, Andrzejewski and Zhu address this problem by fixing the membership of the words to valid topics (Andrzejewski and Zhu, 2009). When the focus is on detecting just one topic, as in our task, we can expand the seed words more selectively using the small set of labeled data and that is the approach adopted in this paper.

3 Data

One interesting aspect of our study is the uniqueness of our corpus, which is both naturalistic and exhaustive. We recorded about 41,000 land-line everyday telephone conversations from 56 volunteers, 65 years or older, over a period of approximately 6 to 12 months. Since these everyday telephone conversations are private conversations, and might include private information such as names, telephone numbers, or banking information, we assured the subjects that no one would listen to the recorded conversations. Thus, we couldn't manually transcribe the conversations; instead, we used an Automatic Speech Recognition (ASR) system that we describe here.

Automatic Speech Recognition System Conversations in our corpus were automatically transcribed using an ASR system, which is structured after IBM's conversation telephony system (Soltau et al., 2005). The acoustic models were trained on about 2000 hours of telephone speech from Switchboard and Fisher corpora (Godfrey et al., 1992). 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. We are unable to measure the performance of this recognizer on our corpus due to the stringent privacy

requirements mentioned earlier. Since both corpora are conversational telephone speech and the training data contains large number of conversations (2000 hours), we expect the performance of our recognizer to be relatively close to results on NIST benchmark.

Heuristically labeling a small subset of conversations For training and evaluation purposes, we need a labeled set of conversations; that is, a set of conversations where we know whether or not they contain health-related discussions. Since the privacy concerns do not allow for manually labeling the conversations, we used reverse look-up service in `www.whitepages.com`. We sent the phone number corresponding to each conversation (when available) to this website to obtain information about the other end of the conversation. Based of the information we got back from this website, we labeled a small subset of the conversations which fell into unambiguous business categories. For example, we labeled the calls to "hospital" and "pharmacy" as health-related, and those to "car repair" and "real estate" as non-health-related.

The limitations of the labeled set The labeled set we obtained is small and restricted in type of conversations. Since phone numbers are not available for many of the conversations we recorded, and also because `www.whitepages.com` does not return unambiguous information for many of available phone numbers, we managed to label only 681 conversations – 275 health-related and 406 non-health-related. This labeled set has another limitation: it contains conversations to business numbers only. In reality however, we are interested in the much larger set of conversations between friends, relatives, and other members of subjects' social support network. Thus, the generalizability of the classifier we train is very important.

Fisher Corpus To explicitly test the generalizability of our classifier, we use a second evaluation set from Fisher corpus (Cieri et al., 2004). Fisher corpus contains telephone conversations with preassigned topics. There are 40 topics and only one of them, illness, is health-related. We identified 338 conversations on illness, and sampled 702 conversations from the other 39 non-health topics. Since we do not train on Fisher corpus, we call it the out-of-domain task to apply our method on Fisher corpus; as opposed to the in-domain task which is to apply our method on the everyday telephone conversations.

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 injury or going to emergency room. We will use these pieces of information to relate the health-related conversations to actual medical events in the subjects' lives.

4 Method

4.1 Overview

As we explained in Section 3, we can label a small set of conversations in the everyday telephone conversations corpus as health-related vs. non-health related. Using this labeled set we can train a support vector machine (SVM) to classify the conversations. In absence of feature selection, the conversations are represented by a vector of tf-idf scores for every word in the vocabulary where tf-idf is a score for measuring the importance of a word in one document of a corpus. As we see in Section 5, such a classifier doesn't generalize to the out-of-domain Fisher task (*i.e.* when we test the classifier on Fisher data set, we do not get good precision and recalls). 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. We propose a new way for feature selection for retrieving documents containing information about a specific topic when there is only a limited set of labeled documents available. The idea is to pick a few words highly related to the topic of interest as seed words and to use TLDA (Andrzejewski and Zhu, 2009) to force those seed words into one (for example, the first) topic. 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 dif-

ferent 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.

Figure 1 shows the proposed iterative algorithm. The algorithm starts with initializing the seed words to *doctor, medicine, and pain*. Then, in each iteration, TLDA performs semi-supervised topic modeling and returns the 50 most probable candidate words in the health topic. We select a subset of these candidate words which, if added to the seed words, would maximize the average of precision and recall on the train set for a simple classifier. This simple classifier marks a conversion as health related if, and only if, it contains at least one of the seed words. The algorithm terminates when the subset selection is unable to add a new word contributing to the average of precision and recall. The tf-idf vector for the expanded set represents the conversations in the classification process.

It is worth mentioning that we train TLDA using all 41000 unlabeled conversations, and chose the number of topics, K, to be 20.

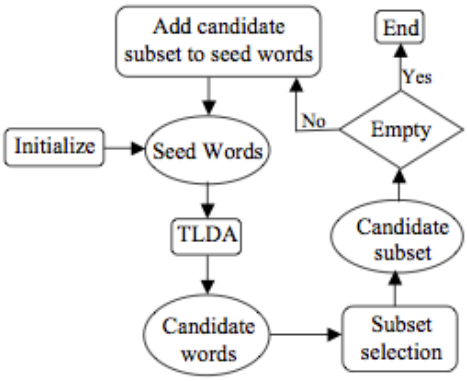

Figure 1: *Expanding the set of seed words:* in each iteration, the current seed words are forced into the topic of health to guide TLDA towards finding more health related words. The candidate set consists of the 50 most probable words of the topic of health in TLDA. We investigate the gain of adding each word of the candidate set to the seed words by temporarily adding it to the seed words and looking at the average of precision and recall on the training set for a classifier that classifies a conversation as health-related if and only if it contains at least one of the seed words. We select the words that maximize this objective and add them to the seed words until no other words contributes to the average precision and recall.

5 Experiments

In all of our experiments, we trained SVM classifiers, with different features, to detect the conversations on health using the popular libSVM (Chang and Lin, 2011) implementation. We chose the parameters of the SVM using a 30-fold cross-validated (CV) grid search over the training data. We also used 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 in-domain task we always report the average performance across the folds.

Table 1 shows the results of our experiments using different input features. We report on recall, precision and F-measure in in-domain and out-of-domain (Fisher) task as well as on average F-measure of the two. 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 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 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

	Recall		Precision		F-measure		
Feature Words	In-Domain	Fisher	In-Domain	Fisher	In-Domain	Fisher	Average
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* (average over 30 runs)	58.4	57.4	86.3	97.5	69.7	72.3	71.0
Words selected by our method (average over 30 runs)	56.1	66.5	91.0	95.5	69.4	78.4	73.9
Union of all selected words (across 30 runs)	67.7	69.4	87.8	95.1	76.5	80.2	78.3

Table 1: Performance of SVM classifiers using different feature selection methods. The In-Domain task involves the everyday telephone conversations corpus. We call Fisher corpus out of domain, because no example of this corpus was used in training.

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. Table 2 shows one set of the 50 most probable words in health topic,the result of one run of TLDA. Evidently, these words contain 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 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 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. It is important to notice that the in-domain performance is still lower than the full-vocabulary baseline by less than 3% while the out-of-domain performance is the best obtained. 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).

Our last experiment tests statistical association between health-related discussions in everyday telephone conversations, and actual medical events in older adults. As mentioned in Section 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.

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**

Table 2: *50 most probable words in the topic of health returned by one run of TLDA. The bold words are the ones are hand-picked.*

Table 3 shows the number of conversations in different categories. At first glance it might seem like the number of false positives or false negatives is quite large but we should notice that being near a medical event is not the ground truth here. We just want to see if there is any association between occurrence of health-related conversations and occurrence of an actual medical event in lives of our subjects. 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. for the other conversations (89.1). 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 3: Number of telephone conversations in different categories. Each conversation is considered near a medical even if and only if there is at least one self-report in a window of 3 weeks around its date. Being near a medical event does not reveal the true nature of the conversation and thus is not the ground truth. So, there are no false positive, true positive, etc. in this table.

6 Conclusions

In this paper, we investigated the problem of identifying conversations with any mention of health. The private nature of our everyday telephone conversations corpus poses constraints on manual transcription and annotation. Looking up phone numbers associated with business calls, we labeled a small set of conversations when the other end was a business clearly related or unrelated to the health industry. However, the labeled set is not large enough for training a robust classifier. We developed a semi-supervised iterative method for selecting features, where we learn a distribution of words on health topic using TLDA, and subsequently filter irrelevant words iteratively. We demonstrate that our method generalizes well and improves the average F-score on in-domain and out-of-domain tasks over two baselines, using full vocabulary without feature selection or feature selection using TLDA alone. In our task, the generalization of the classifier is important since we are interested in detecting not only conversations on health with business (the annotated examples) but also with others in subjects' social network. Using our classifier, we find a significant statistical association between the occurrence of conversations about health and the occurrence of self-reported medical events.

Acknowledgments

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. We thank Nicole Larimer for help in collecting the data, Maider Lehr for testing the data collection devices and Katherine Wild for early discussions on this project. We are grateful to Brian Kingsbury and his colleagues for providing us access to IBM's *attila* software tools.

References

David Andrzejewski and Xiaojin Zhu. 2009. Latent dirichlet allocation with topic-in-set knowledge. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, SemiSupLearn '09, pages 43–48.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

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

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Hong-qi Han, Dong-Hua Zhu, and Xue-feng Wang. 2009. Semi-supervised text classification from unlabeled documents using class associated words. In *Computers & Industrial Engineering, 2009. CIE 2009. International Conference on*, pages 1255–1260. IEEE.

Toru Hisamitsu and Yoshiki Niwa. 2001. Topic-word selection based on combinatorial probability. In *NLPRS*, volume 1, page 289.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213. Association for Computational Linguistics.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics.

Hang Li and Kenji Yamanishi. 2003. Topic analysis using a finite mixture model. *Information processing & management*, 39(4):521–541.

Kyle W. Prier, Matthew S. Smith, Christophe Giraud-Carrier, and Carl L. Hanson. 2011. 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*, pages 18–25.

Hagen Soltau, Brian Kingsbury, Lidia Mangu, Daniel Povey, George Saon, and Geoffrey Zweig. 2005. The ibm 2004 conversational telephony system for rich transcription. In *Proc. ICASSP*, volume 1, pages 205–208.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics.

Taras Zagibalov and John Carroll. 2008. Automatic seed word selection for unsupervised sentiment classification of chinese text. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 1073–1080. Association for Computational Linguistics.