Applying Supervised & Unsupervised Machine Learning to Alzheimer's Caregiving Text

Andrew Bright ANBRIGHT@SEAS.UPENN.EDU
Timothy Lin TIMLIN@SEAS.UPENN.EDU
Michael Ciul CIULMIKE@GMAIL.COM

1. Project Overview

Family members caring for more than 5.5 million people with Alzheimer's disease require a great deal of specialized disease knowledge from bathing and hygiene techniques to medication administration. Unsurprisingly, most family caregivers utilize online forums, knowledge boards, and articles to find relevant information. However, most of these forums are legacy systems which display resources chronologically and lack modern features which helps users find relevant content like content tagging or recommendation. In addition, reading information about caregiving can be an overwhelming experience; caregivers we interviewed described wanting a balance between positive and negative content. By applying machine learning techniques to this legacy text, we can extract content tags and underlying emotions in order to create a more rich and efficient user experience. Our categorization program empowers family members and caregivers by giving them easy and quick access to relevant information they need to help their loved ones with Alzheimer's.

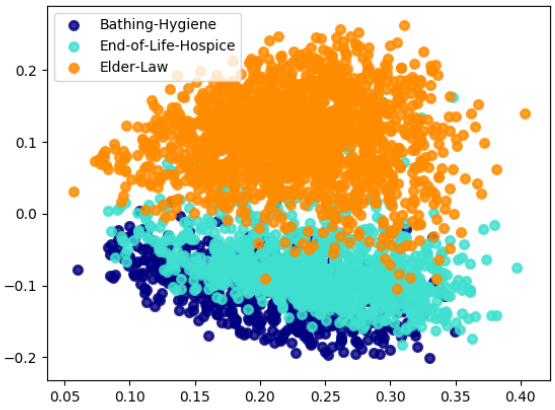
1.1. Data Summary

For this project, our group is analyzing forum posts from the AlzConnected forum associated with the Alzheimer's Association, a nonprofit whose mission is to "advances research to end Alzheimer's and dementia while enhancing care for those living with the disease." Based on discussions with the organization, they are very interested in using machine learning to improve the user experience and better utilize existing resources and posts. Their dataset includes 20k threads on a wide range of issues from "Elder Law" to "Assisted Living". Inside each thread is a set of comments, each of which is a text post.

28 labels have been preselected which describe the most common issues and topics related to Alzheimer's caregiving including *Family*, *Finances*, *Legal*, *Hospice*, *Bathing and Hygiene*, and *Medications* (Figure 1). Based on these labels, our group looked at publicly available articles and text posts for Alzheimer's under each label to understand the feature makeup and distribution of each category. After cleaning the training data, there were 29k labeled samples remaining.

There were also 20,193 forum text samples that were unlabeled (not categorized). These were not manually labeled by the user who posted them. AlzConnected would like these forum posts not categorized by the user to be categorized as well so other users can more easily find useful information from them.

Figure 1. Dimensionality Reduction on Subset of Labeled Data



1.2. Overview of Methods

To extract content tags from the text posts, our group employed both supervised and unsupervised techniques. Before applying any supervised techniques, we first preprocessed the data by removing stop words and punctuation and transforming the feature vectors into term frequency-inverse document frequency (tf-idf) vectors.

For the supervised approach, we tested two primary text classification techniques: Naive Bayes and Support Vector Classification. Naive Bayes, a probabilistic approach which assumes conditional independence, which is very efficient to train and is normally used as a baseline in text classification. Support Vector Classification involves fit-

ting a Support Vector Machine to the data and employing a One-vs-All approach for each label. Support Vector Machines are large-margin classifiers which means that they fit a hyperplane to the data maximizing the distance between the two classes.

For the unlabeled data, we used an unsupervised approach to see if machine learning algorithms could come up with forum topics that better classified each post. For this unsupervised approach, we used both Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF).

LDA is a generative model of discovering topics that a text document contains and assigning a probability that a document is associated with each topic. The topic mixture of LDA is determined by a Dirichlet distribution over a fixed set of K topics (Blei). We used a count vectorizer with stop words to create a term-frequency (tf) matrix representation of the text data. We then tuned the hyperparameters of the model with gridsearch with 5 cross-fold validation using maximum likelihood per token as a scoring mechanism. With the best hyperparameters found, we used cross validation to evaluate the train and test maximum likelihood per token. We then ran the LDA on the whole dataset and received the top 10 keywords of each topic and the topic distribution across documents.

In Non-negative Matrix Factorization (NMF), a matrix V (with tf-idf features) is factorized into two matrices W and H through many iterations, with the property that all three matrices have no negative elements (Xu). We used a count vectorizer with stop words to create an term frequency-inverse document frequency (tfidf) matrix representation of the text data. We then tuned the hyperparameters of the model with gridsearch using reconstruction error as a scoring mechanism. With the best hyperparameters found, we used cross validation to evaluate the train and test reconstruction error. We then ran the NMF on the whole dataset and received the top 10 keywords of each topics and topic distribution across documents.

For Sentiment Analysis, we wanted to create a warning for authors on whether their post is too positive or negative to help eliminate posts the community may not want to see. The problem requires passing in a sequence of words and labeling that sequence (either positive or negative). There is a lot of empirical evidence that DNNs are very effective at these many-to-one problems. In particular, Recurrent Neural Networks (RNNs), networks with directed cycles, are effective at analyzing underlying semantic meaning in text, especially when they use Long Short Term Memory (LSTM) hidden units. LSTMs include four gates– a cell, an input gate, an output gate, and a forget gate– and are generally much less sensitive to large gaps in sequence data.

Table 1. Classification accuracies for Naive Bayes and SVC.

CLASSIFIER	ACCURACY	RECALL	PRECISION
NB	53.3	54.0	65.1
NB WITH PCA	70.2	-	-
SVC	84.2	84.1	85.5

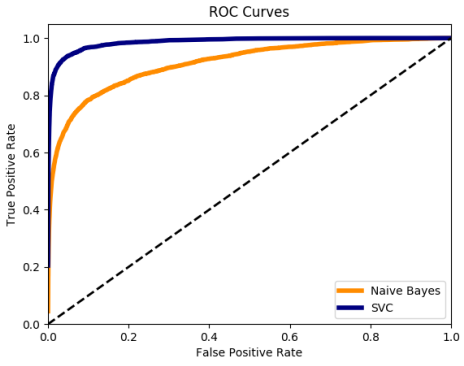
2. Analysis and Results

2.1. Supervised Classification

For both classifiers, we employed a 80-20 training to test data split and reported the testing accuracy in Table 1.

For the Naive Bayes model, our group selected a Naive Bayes model as a baseline. For the SVC, we used a One-vs-Rest wrapper to classify multiple labels. We conducted parameter search on both the NB and the SVC model. For the NB, we conducted a hyperparameter search using different prior distributions (e.g. Bernoulli) and different learning rates α from 0-1. Our final model was a Multinomial NB with a bag of words approach and $\alpha = 0.2$. We also experimented with dimensionality reduction using Principal Component Analysis (PCA) and found that by reducing the tf-idf vector from 47k to 500, we could significantly increase performance on the NB model. For SVC, the hyperparameters we used were the kernel method, C, and $gamma$. Our final model used the cosine similarity kernel with a $C = 10$ and $gamma = 1e - 3$ and achieved an 84% test score. The ROC curve for both classifiers is shown in Figure 3.

Figure 2. ROC Curve for Naive Bayes and SVM



However, this may not be an accurate estimate on the AlzConnected data population. Because of this, we tested these classifiers on unlabeled data from the AlzConnected data population by first labeling a small subset of the data ourselves, and then predicting it with the classifier (first 5 are shown in Table 2). This demonstrates that we can successfully transfer the classifier to this data population and

Table 2. Subset of Classifications on Labeled Forum Data

Human Label	NB Label	SVC Label
Guardianship	Family	Guardianship
Medications	Burnout	Medications
Emotions	Emotions	Emotions
Burnout	Family	Emotions
End of Life	Emotions	Hospice

generate accurate labels.

2.2. Unsupervised Label Extraction

For unsupervised learning, LDA was ran over 20,193 training examples. The best hyperparameters found were 10 topics, batch learning method, learning decay of 0.6, and a batch size of 64. The training maximum-likelihood per token was -790 and the test one was -817.

Table 3: LDA Results

Topic #	Top 3 Keywords	Topic Name
0	care home facility	Facility
1	dementia doctor meds	Doctor/Meds
2	feeding food alcohol	Food and Drink
3	time day night	Time
4	dementia people think	Dementia
5	time life love	Other
6	disease alzheimer brain	Disease Science
7	essential oils disease	Disease Science
8	good time day	Other
9	mom mother dad	Parents

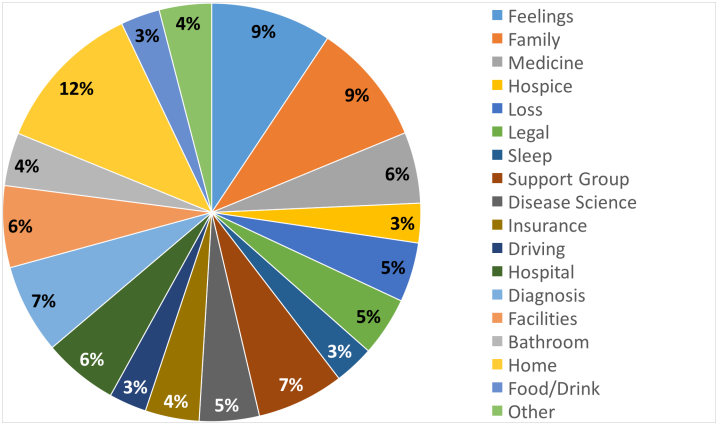
10 keywords related to each topic were printed and we manually assigned topic names based on those keywords. Because space in this document is limited, only the top 3 keywords were included in this table. Note that this does not totally represent the topic. Our LDA model ran into a few issues. First, topic 1 could be split into multiple topics. Its keywords included dementia, doctor, meds, hospice, hospital, and neurologist. This shows that the LDA may not have been able to differentiate between these posts. Second, some topics could be combined like topics 6 and 7. Third, some topics like topics 3, 5, and 8 had a mix of keywords that made it difficult to pinpoint it to one actual category. Overall, the topics given by the LDA model were not too helpful. The LDA model did pick up words that showed up a lot but picked up on words that were too general. It's possible that maximum-likelihood was not the best metric for scoring topics associated with the forum data. A disadvantage of LDA topics are soft-clusters and there is no perfect objective quantitative manner to gauge performance. The gridsearch may not have been able to pick up the best hyperparameters.

NMF was ran over 20,193 training examples as the second unsupervised approach. The best hyperparameters found were 20 topics, coordinate descent solver, frobenius beta loss, 0 alpha, and 0.01 l1 ratio. The minimum reconstruction error found was to be 130. The top 3 keywords associated with each topic along with a manually assigned topic name are displayed below.

Table 4: NMF Results

Topic #	Top 3 Keywords	Topic Name
0	feel love understand	Feelings
1	mom stage time	Other
2	mother sister brother	Family
3	meds aricept namenda	Medicine
4	hospice nurse palliative	Hospice
5	sorry loss prayers	Loss
6	poa attorney lawyer	Legal
7	dad parents brother	Family
8	sleep bed night	Sleep
9	support help group	Support Group
10	disease alzheimer amyloid	Disease Science
11	medicaid pay insurance	Insurance
12	driving car license	Driving
13	hospital surgery pain	Hospital
14	dementia diagnosis symptoms	Diagnosis
15	facility care home	Facilities
16	shower bathroom toilet	Bathroom
17	home house day	Home
18	eat food drink	Food/Drink
19	husband son wife	Family

Figure 3. NMF Topic Assignments to Documents Distribution NMF.png



From a manual inspection, NMF outperformed LDA. NMF had less of seemingly random keywords being assigned to

the same topic (topic 1 was the only one). However, overlapping topics is still an issue - family topics are repeated 3 times. In both NMF and LDA, a common theme is family. Keywords describing family members were picked up multiple times by both algorithms. Overall, NMF topics were also more specific and useful as categories. Topics like topic 6 (Legal), 9 (Support Group), and 12 (Driving) are useful categories that were added by the NMF.

Comparing NMF topics to topics picked by AlzConnected, 14 topics are not directly covered by NMF topics (eg. Tough Issues and Depression). These topics may be encompassed in other topics (eg. depression under "feelings"). 17 topics are directly covered (eg. Bathing is under "bathroom"). All NMF topics were covered by the AlzConnected topics, which there were more of.

2.3. Sentiment Analysis

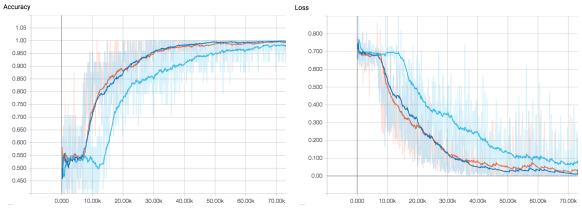
In order to train our model, we could not find free sentiment data relating to caregiving or healthcare, so we instead used the most comprehensive training data we could find– the IMDB movie review database. The database contains 15,000 reviews with the text of the review and a label, either positive or negative.

To train a sentiment analysis model, we first converted the training data into a word embedding representation. A word embedding model is a high dimension feature space that is trained on a lexicon. The purpose of the embedding model is to capture the syntactical richness associated with certain words. For instance, a close distance between words in an embedding model can captures similar meaning (synonyms). We used the Stanford GloVe model.

After converting the reviews to the GloVe feature space, we trained a Tensorflow LSTM using a single layer of LSTM cells. We performed a Grid Search using dropout (0.25, 0.50, 0.75) and the number of LSTM units (64, 124, 256) as hyperparameters, training first on a 67% training set and selected based on a 33% test set. Based on this search, we selected a final model with $dropout = 0.50$ and $units = 64$ (Figure 3) which achieved a training accuracy of 99% and a test accuracy of 85%.

By applying this analyzer to the Alzheimer's forum data, we found that it labeled roughly 70% as negative and 30% as positive. Inspecting a sample of this, we found the extreme ones in either direction to be plausible. For instance, a post titled "We Are Meeting With Hospice Tomorrow" was correctly labeled negative whereas a post titled "Wishing everyone could be a caregiver for 1 month!" was correctly labeled positive. However, the analyzer had difficulty on topics that it had little exposure; for example, it incorrectly labeled a post on Gardening as negative. To improve the accuracy of this classifier, it would be helpful to crowdsource labeling sentiment from the Alzheimer's forum data and train the LSTM on that data population.

Figure 4. Hyperparameter Search for Dropout on 64 Unit LSTM



3. Next Steps

3.1. Labeling Legacy Data

Through the trained text classifier, we have the ability to retroactively label all posts. By doing this, we can tag and group content together making it much easier to find useful and relevant posts. In addition, if a user is interested in a particular topic, such as End of Life, they can browse all posts tagged with that category.

3.2. Optimal Category Choices

Using unsupervised clustering, we discovered 20 optimal topics and top keywords for each. We suggest that the organization adopt the categories outlined in the above table. Going forward, by providing these tags and relevant examples, users can correctly label their own posts.

3.3. Balancing Emotional Attributes

Caregiving is a very difficult and emotional experience. Through the course of our project, several caregivers interviewed described having to take a break from reading posts because it was too emotionally overwhelming. By using the sentiment analyzer, posts with extremely negative or positive content can be labeled. By properly balancing these labels, a feed system can be developed so that the user experience does not become too overwhelming.

4. Conclusion

Through this project, we demonstrated the feasibility of using supervised classification, unsupervised label extraction, and sentiment analysis on legacy Alzheimer's caregiving data. Despite different data populations, these results are promising and suggest that increased accuracy can be achieved with more resources (e.g. crowdsourcing labels). We hope these results can be used as a starting point for further research in caregiving and have a positive impact in the lives of those afflicted by Alzheimer's disease.

5. Acknowledgements

We would like to thank Professor Eric Eaton for insight on project design and unsupervised learning techniques.

5.1. Citations and References

Alzheimer's Association. (2011). 2011 Alzheimer's disease facts and figures. Alzheimer's dementia: the journal of the Alzheimer's Association, 7(2), 208.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." Journal of Machine Learning Research, vol. 3, Jan. 2003.

Xu, Wei, Xin Liu, and Yihong Gong. "Document Clustering Based On Non-negative Matrix Factorization." ACM, July 2003.

Zhang, H. (2004). The optimality of naive Bayes. AA, 1(2), 3.

Zhang, H., Li, D. (2007, November). Nave Bayes text classifier. In Granular Computing, 2007. GRC 2007. IEEE International Conference on (pp. 708-708). IEEE.

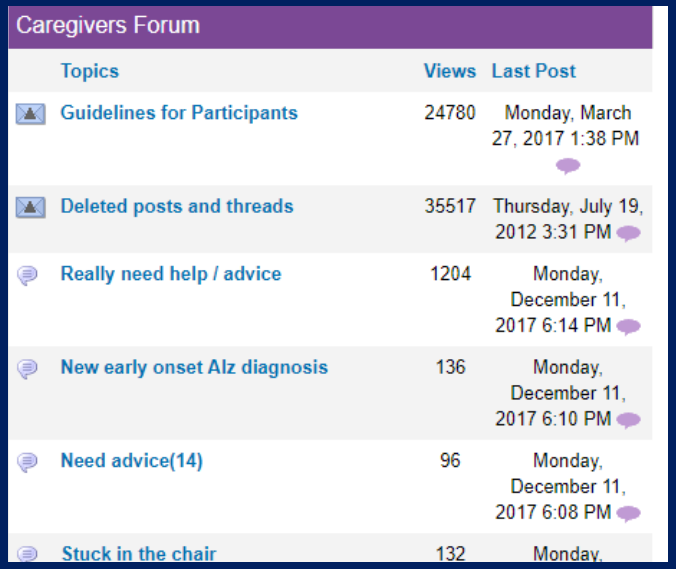
Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. Machine learning: ECML-98, 137-142.

Gers, F. A., Schmidhuber, J., Cummins, F. (1999). Learning to forget: Continual prediction with LSTM.

Dos Santos, C. N., Gatti, M. (2014, August). Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In COLING (pp. 69-78).

The Problem

Alzheimer's caregiving requires a great deal of specialized knowledge. However, these forums are legacy and lack the modern features like category tags which help users find relevant content. Our group analyzed the Alzheimer's Association forum to build a better user experience by 1) categorizing existing content, 2) clustering topics, and 3) analyzing sentiment to ensure the experience is not overwhelming.



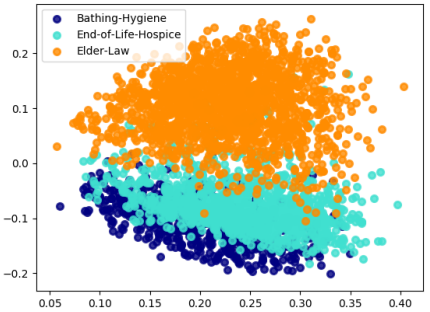
Caregivers Forum

Topics	Views	Last Post
Guidelines for Participants	24780	Monday, March 27, 2017 1:38 PM
Deleted posts and threads	35517	Thursday, July 19, 2012 3:31 PM
Really need help / advice	1204	Monday, December 11, 2017 6:14 PM
New early onset Alz diagnosis	136	Monday, December 11, 2017 6:10 PM
Need advice(14)	96	Monday, December 11, 2017 6:08 PM
Stuck in the chair	132	Monday,

Our Approach

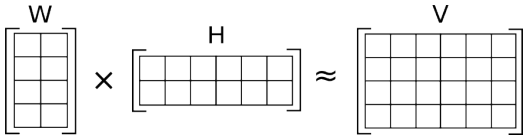
Categorizing Content

- ~29k labeled examples
- Train NB and SVM



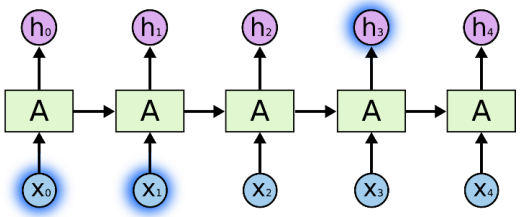
Clustering Topics

- LDA and NMF
- Maximum likelihood and reconstruction error as scoring metrics



Sentiment Analysis

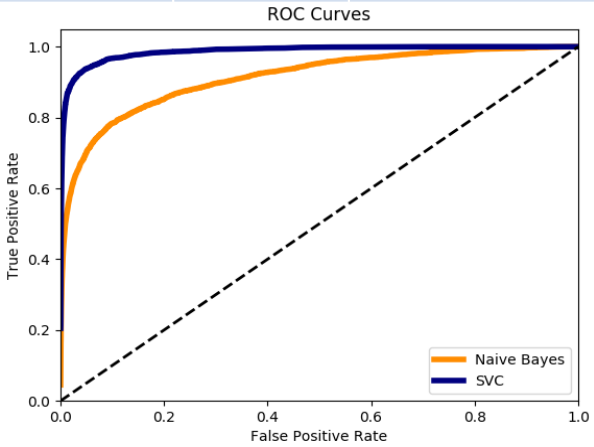
- LSTM RNN
- Trained on IMDB movie reviews



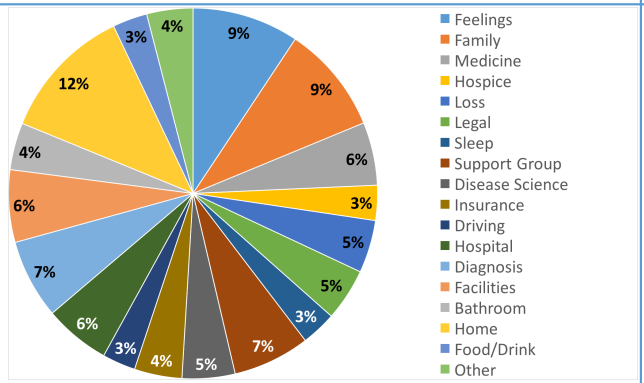
Analysis and Results

Categorizing Content

Classifier	Accuracy	Optimal Parameters
NB	53.3	**Prior:**Bernoulli **α:** 0.2
NB w/ PCA	70.2	**Dimensions:** 500
SVM	84.2	**Kernel:** cosine **C:** 10 **γ:** 1e-3



Clustering Topics



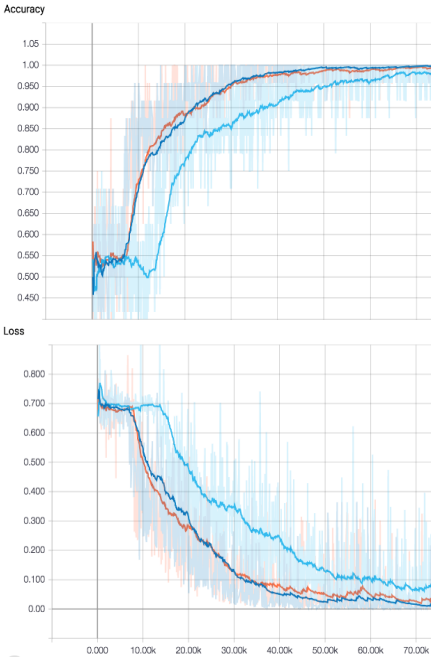
NMF outperformed LDA. These are NMF Results

Topic Name	Topic Keywords
Medicine	Meds, aricept, namenda
Hospice	Hospice, nurse, palliative
Legal	Poa, attorney, lawyer
Food/Drink	Eat, food, drink
Sleep	Sleep, bed, night

Sentiment Analysis

Applied to Alzheimer's Data

- 69.6% labeled negative
- 30.4% labeled positive



One layer LSTM

- 64 hidden units
- dropout = 0.5