Towards Early Dementia Detection: Fusing Linguistic and Non-Linguistic Clinical Data

Joseph Bullard[1], **Cecilia Ovesdotter Alm**[2], **Xumin Liu**[1], **Rubén A. Proaño**[3], **Qi Yu**[1]
[1]College of Computing & Information Sciences, [2]College of Liberal Arts, [3]College of Engineering
Rochester Institute of Technology, Rochester, NY
{jtb4478,coagla,xmlics,qyuvks,rpmeie}@rit.edu

Abstract

Dementia is an increasing problem for an aging population, with a lack of available treatment options, as well as expensive patient care. Early detection is critical to eventually postpone symptoms and to prepare health care providers and families for managing a patient's needs. Identification of diagnostic markers may be possible with patients' clinical records. Text portions of clinical records are integrated into predictive models of dementia development in order to gain insights towards automated identification of patients who may benefit from providers' early assessment. Results support the potential power of linguistic records for predicting dementia status, both in the absence of, and in complement to, corresponding structured non-linguistic data.

1 Introduction

Dementia is a problem for the aging population, and it is the 6th leading cause of death in the US (Alzheimer's Association, 2014). Around 35 million people worldwide suffer from some form of dementia, and this number is expected to double by 2030 (Prince et al., 2013). The most common form of dementia is Alzheimer's Disease, which has no known cure and limited treatment options. The clinical care for dementia focuses on prolonged symptom management, resulting in high personal and financial costs for patients and their families, straining the healthcare system in the process. Early detection is critical for potential postponement of symptoms, and for allowing families to adjust and adequately plan for the future. Despite this importance, current detection methods are costly, invasive, or unreliable, with most patients not being diagnosed until their symptoms have already progressed. Dementia diagnosis is a life-changing event not only for the patient but for the caretakers that have to adjust to the ensuing life changes. Improved understanding and recognition of early warning signs of dementia would greatly benefit the management of the disease, and enable long-term planning and logistics for healthcare providers, health systems, and caregivers.

With the advent of electronic clinical records comes the potential for large-scale analysis of patients' clinical data to understand or discover warning signs of dementia progression. The ability to follow the evolution of the disease based on patients' records would be key to develop intelligent support systems to assist medical decision-making and the provision of care. Current research using records mainly focuses on structured data, i.e. numerical or categorical data, such as test results or patient demographics (Himes et al., 2009). However, unstructured data, such as text notes taken during interactions between patients and doctors, presents a potentially rich source of information that may be both more straightforwardly interpretable for humans, as well as helpful for early dementia detection. Structured data from innovative diagnostic tests are often absent due to their cost and accessibility, text notes are generated for nearly every visit of a patient. Moreover, text notes in medical records are a source of natural language, and potentially more flexibly encode the diagnostic expertise and reasoning of the clinical professionals who write them.

Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 12–22,
San Diego, California, June 16, 2016.

Processing and computationally analyzing natural language remains a formidable task, but insights gleaned from it may translate particularly well into actual clinical practice, given its interpretable and accessible nature. Therefore, the ability to predict dementia development based on both structured and unstructured data would be useful for intelligent support systems which could automatically flag individuals who will benefit for further evaluation, reducing the impact of late diagnosis.

1.1 Related Work

Structured clinical data has been useful for identifying known disease markers (Himes et al., 2009). Procedural and diagnostic codes (e.g., ICD-9) can provide high specificity for identifying a disease, but may not provide sufficient sensitivity (Birman-Deych et al., 2005; Kern et al., 2006). A patient's history, however, is typically summarized by a clinician in text form, and can provide informative expressiveness and granularity not adequately captured by ICD-9 codes (Li et al., 2008). Interestingly,

Prior work has shown that natural language data can help synthesize details and discover trends in medical records. Natural language processing and text mining have been applied to the identification of various known medical conditions. One method maps specific conditions to relevant terms from ontologies (curated knowledge bases). For example, SNOMED-CT predicted post-operative patient complications (Murff et al., 2011), and MedLEE (Friedman et al., 1995) identified colorectal cancer cases (Xu et al., 2011), suspicious mammogram findings (Jain and Friedman, 1997), and adverse events related to central venous catheters (Penz et al., 2007). Similarly, the language analysis-based resource SymText (Haug et al., 1995) has been used for detecting bacterial pneumonia cases from descriptions of chest X-ray (Fiszman et al., 2000).

While such studies with medical knowledge bases are useful for disease identification, they mostly involve conditions with well known markers and known relationships between words and clinical concepts typically available once the patient is symptomatic. However, many cognitive conditions, such as dementia, as well as other illnesses of interest, are not well understood and their onsets gradually evolve over long periods of time. Furthermore, diagnosing such conditions is often primarily a function of experts' analysis, transcribed into notes. Thus, discovering lexical associations with the progression of these conditions could be tremendously beneficial, and could also help to validate and enhance the use of resources such as the Alzheimer's Disease Ontology (Malhotra et al., 2013).

Topic models have produced interesting results across domains (Chan et al., 2013; Resnik et al., 2013; McCallum et al., 2007; Paul and Dredze, 2011). Latent Semantic Indexing (LSI) has been used in medicine to discover statistical relationships between lexical items in a corpus. LSI has been used to supplement the development of a clinical vocabulary associated with post-traumatic stress disorder (Luther et al., 2011), and for forecasting ambulatory falls in elderly patients (McCart et al., 2013). However, LSI often requires around 300–500 concepts or dimensions to produce stable results (Bradford, 2008). This limitation can be overcome by using LDA, whose identified groups of related terms are also more intuitive for human interpretation than LSI results. Additionally, representing documents by their LDA topic distribution reduces the dimensionality of the feature space. Furthermore, a study with microtext data demonstrated that document length influences topic models, and that aggregating short documents by author can be beneficial (Hong and Davison, 2010). This finding is relevant for this study due to the short nature of clinical texts.

This study is concerned with the fusion of linguistic data with structured non-linguistic data, as well as the integration of distinct models suitable for each. Approaches to the former case, have been studied (Ruta and Gabrys, 2000). For the latter case, integration of classifiers typically involves multiple models of the same data, e.g. ensemble methods such as random forests, and often utilizes voting algorithms to produce the final combined output. However, here we focus on the combination of two distinct models: one based on linguistic data and one on structured non-linguistic data. This setup complicates the use of typical voting methods, and thus we explore a less frequently studied solution that leverages Bayesian probability to produce posterior distributions (Bailer-Jones and Smith, 2011).

1.2 Our Contributions

(1) We compare performance of predictive modeling with linguistic vs. non-linguistic features, studying if linguistic features used alone as predictors yield performance comparable to that of non-linguistic record data – especially when the latter exclude cognitive assessment scores from expert-administered tests. Our results show the utility of linguistic data for dementia prediction, e.g., when relevant structured data are unavailable in the records, as is often the case. (2) We explore the use of Latent Dirichlet Allocation (LDA) (Blei et al., 2003) as textually interpretable dimensionality reduction of the lexical feature space into a topic space. We examine if LDA can transform the sparse term space into a reduced topic space that meaningfully characterizes the texts, and we discuss its practical value for classification. (3) We study the challenge of fusing linguistic and non-linguistic data from records in additional classification experiments. If fusion improves performance, this would strengthen the utility of records-based linguistic features for disease prediction. We explore two integration methods: combining feature vectors computed independently from structured and text data, or leveraging probabilistic outputs of their respective trained classifiers.

This paper is organized as follows. Section 2 describes the data for the dementia detection problem. Section 3 presents our framework and integration. Section 4 outlines experiments and results. We conclude with future directions in Section 5.

2 Dementia Detection Problem: Data

This study makes a secondary use of a data set from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (`adni.loni.usc.edu`). The ADNI study contains mostly structured data, such as measurements from brain imaging scans, blood, and cerebrospinal fluid biomarkers. The dataset also contains optional text fields in which examiners include notes or descriptions at their discretion.

Each ADNI subject[1] is labeled upon entering the study. ADNI's original labeling scheme was modified in later phases of the study, resulting in some subjects having updated labels, while others remain unchanged. Therefore, only subjects who joined the study under the most recent phase, *ADNI-2*, are included in this work. Subjects with a label of *SMC* (Significant Memory Complaint; reflecting a self-reported memory issue) are excluded as it is not a real diagnostic category outside of ADNI. A subject's record must have both unstructured text and structured data to be included, resulting in 679 usable subjects; from here on we refer to their data.

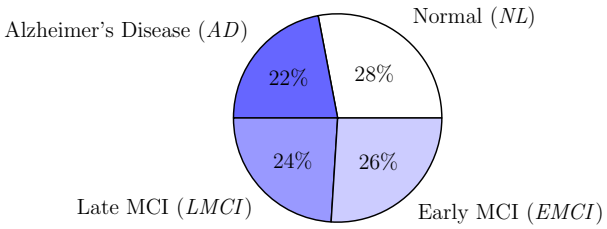

Figure 1: Distribution of subject diagnostic labels ($n = 679$). MCI = Mild Cognitive Impairment.

The *ADNI-2* phase of the ADNI collection uses several labels to indicate the progression to Alzheimer's Disease: *NL* (Normal), *EMCI* (Early Mild Cognitive Impairment), *LMCI* (Late MCI), and *AD* (Alzheimer's Disease). The label (class) distribution of the remaining 679 subjects is relatively balanced (see Figure 1). Moderately-sized data sets are common in clinical NLP contexts, where data is understandably more challenging to collate and access. For the text data, we considered text source files with considerable quantities of information.[2] All 679 subjects possess text notes in at least one of these four files. Entries from these files are aggregated by subject and concatenated to yield one text document per subject.

There are 22 structured data fields in this ADNI subset. The problem of missing values in the structured data was handled through multiple imputation (using the Amelia II package in `R`). This process uses log-likelihoods to generate probable complete datasets. Most structured data comes from either cerebrospinal fluid samples or brain imaging scans, while three fields correspond to scores on cognitive exam evaluations: the Clinical Dementia Rating (CDR), the Mini Mental State Examination (MMSE), and the Alzheimer's Disease Assessment Scale (ADAS13). Importantly, a meaningful distinction can be made between structured data from cognitive assessments versus those from biophys-

[1] The ADNI study refers to its participants as *subjects*.

[2] See Table in the supplementary documentation.

ical tests/markers. A cognitive assessment is administered by a clinical professional, and thus is a reflection of that person's opinion and expertise. Essentially, cognitive assessment scores are outputs of professional interpretation, whereas other structured data are inputs for future interpretation. Cognitive assessments are also usually administered when providers already suspect dementia, and thus can be regarded as post-symptomatic. Patients, providers, and families will benefit from early detection, and such automated detection can also help prioritize the scheduling of expert-based cognitive assessments in resource-strained healthcare environments.

3 Modeling of Linguistic Data

There are three main feature representations for the linguistic data: *bag-of-words* (BOW), *term-frequency inverse-document-frequency* (*tf-idf*) on top of BOW, and *topics* from LDA.

Preprocessing and text normalization were performed in Python and NLTK, involving lowercasing, punctuation removal, stop-listing, and number removal (with exception of age mentions). Besides regular stop-listing, words or phrases revealing a subject's diagnostic state (for example *MCI*) were removed. Words in a document were lemmatized to merge inflections (removing distinctions between for instance *cataracts* and *cataract*). Abbreviation expansion used lexical lists. The 200 most frequent lexical content bigrams and trigrams were extracted and concatenated (*breast cancer* → *breast_cancer*). Lastly, while dates were removed, age expressions were kept after conversion and binning (*AGE_>=70_<80*), as they may be important for this problem. Ages below 40 were represented as *AGE_<40* and ages at or above 90 as *AGE_>=90*.

BOW and tf-idf were implemented using `gensim` (Řehůřek and Sojka, 2010). The standard BOW representation is very sparse, since any document only contains a small subset of the vocabulary. An extension weights the terms based on their distribution in the corpus using *tf-idf*. Thus higher weights are assigned to terms which appear more times in fewer documents, and lower weights to terms which appear fewer times and/or in more documents. The feature space of tf-idf corresponds to standard BOW, but the values are the weights.

LDA is a generative model for identifying latent topics of related terms in a text corpus, D, which consists of M documents and is assumed to contain K topics. Each topic k is essentially follows a multinomial distribution over the corpus vocabulary, parameterized by ϕ_k, which is drawn from a Dirichlet distribution, i.e., $\phi_k \sim \mathrm{Dir}(\beta)$. Similarly, each *document* follows a multinomial distribution over the set of topics in the corpus, also assumed to have a Dirichlet probability, denoted $\theta_i \sim \mathrm{Dir}(\alpha)$. Working backwards, the probability of each term in a document is determined by the term distribution of its topic, which is in turn determined by the topic distribution of the document (Blei et al., 2003).

Under LDA, a document is modeled as a probabilistic distribution over topics, learned from the occurrence of terms through Collapsed Variational Bayesian (CVB) inference methods using the Stanford Topic Modeling Toolbox (Teh et al., 2007).[3] Since topics are determined based on statistical relationships of terms, the effectiveness of the model can be hampered by extremely frequent or infrequent terms. For these reasons, we filter out the vocabulary (Boyd-Graber et al., 2014, p. 9) for terms appearing less than 3 times and the 30 most common terms.[4]

3.1 Integration with Structured Data Models

Integration is performed on the results of each unstructured modeling experiment (BOW, tf-idf, and LDA) and those of each structured ones–with vs. without cognitive assessment features. For LDA, only the parameters with the highest performance are used in integration. The most intuitive form of integration is concatenation of the feature vectors for structured and unstructured data. Hence, *concatenation* refers to joining two vectors of length n and m into a single new vector of length $n + m$. This concatenated feature vector is used in classification.

The second approach of integration leverages posterior probabilities from the individual (linguistic vs. non-linguistic) classification models. For each input, a classifier produces a posterior probability of each class label and selects the most probable as its output. One classifier is trained on structured data

[3] Compared to Gibbs sampling (also explored initially), CVB converged on more sensible topics and performed better in model development.

[4] Other cutoff values were explored initially.

features X_s, and another on unstructured data features X_u, resulting in two posterior distributions. The probability of a class C_k is then denoted as $p(C_k \mid X_s, X_u)$. If these distributions are assumed to be conditionally independent with respect to their class labels, then by Bayes' theorem:

$$p(C_k \mid X_s, X_u) \propto \frac{p(C_k \mid X_s)\, p(C_k \mid X_u)}{p(C_k)} \quad (1)$$

From here, the class label with the highest probability is selected as the output; for details see Bailer-Jones and Smith (Bailer-Jones and Smith, 2011).

For integration purposes, we use logistic regression for all classification experiments, implemented in `scikit-learn` (Pedregosa et al., 2011) to compute the posterior probabilities of all classes. We adopt a regularized logistic regression model to further improve the predictive accuracy. By incorporating a regularization term into the basic logistic regression model, regularized logistic regression is able to reach a good bias-variance trade-off and hence achieve a better generalization capability. The regularization term is comprised of two parameters, which are C, the inverse of regularization strength,[5] and the penalty function (either the L^1 or L^2 vector norm). A smaller C corresponds to harsher penalties for large coefficients. The values of these parameters are selected through a grid search of possible values, evaluated by accuracy in cross validation. The process is repeated for each labeling scheme.

4 Experimental Study

Each subject is annotated with a dementia status class label. Each subject's linguistic and structured non-linguistic data are used separately or integrated, as instances for classification.Two different classification problems are reported on. One involves all four classes (*NL*, *EMCI*, *LMCI*, *AD*). This 4-class problem is henceforth referred to as *Standard*. As discussed, early detection of dementia is critical. Accordingly, *EMCI* subjects are of particular interest, as they represent the beginning of the disease's progression. In the second experiment, we use 367 subjects having one of these two class labels (187 *NL*, 180 *EMCI*). While this does not perfectly match the reality of diagnosis, as it excludes the later dementia stages, it could be argued that those later stages are in less need of automatic analysis since they are more readily observable.The resulting binary problem is referred to here as *Early Risk*.

The results and discussions presented later in this paper include a comparison to a majority class baseline, however, this is included merely as a standard comparison, while the actual comparison of interest is between integration of non-linguistic (with vs. without cognitive assessment scores) and linguistic features compared to those groups in isolation.

Held-out Data The data set is randomly split into 80% ($n = 544$ subjects) for model development (*dev* set), and 20% ($n = 135$ subjects) for final evaluation (*held-out* set). Models are only exposed to the *held-out* set after satisfactory performance is achieved using the *dev* set. Class distributions are preserved in the *dev* and *held-out* sets.

LOO Cross-Validation Although the *dev* and *held-out* sets have similar class distributions, overfitting is still a potential issue. For this reason, after the held-out evaluation is complete, a leave-one-out cross-validation (LOO or LOOCV) procedure is run on the entire merged dataset to serve as an additional evaluation, to either confirm or call into question the trends from held-out testing, which may be evident through differences in performance of the same features and models. LOOCV is a case of k-fold cross-validation where k is equal to the number of training instances, resulting in one fold for every data point in which all other data points are used for training.

4.1 Topic Exploration and Evaluation

Tuning of the topic number parameter is essential to finding an appropriate LDA model. This process is performed by iteratively measuring classification accuracy at values of K ranging from 5 to 100, in multiples of 5, using the training data from the held-out evaluation split. LDA is being used here with two goals in mind: to improve classification performance as a form of dimensionality reduction, as well as to provide human-interpretable topics. The former is more convenient and appropriate in the context of this work, but does not necessarily imply good results for the latter. A clinical expert view-

[5] It is common in other sources to use λ for the regularization strength, but the employed `scikit-learn` library instead uses $C = 1/\lambda$, i.e. the *inverse* of regularization strength.

ing the output of such a model would likely prefer fewer topics, each with higher interpretability. Accordingly, LDA models in classification are examined with various per-topic metrics known to correlate well with human evaluation. Thus, the best-performing reduced topic-feature space is selected for classification results and then additionally analyzed using the *topic coherence* metric (Mimno et al., 2011), which measures how often the most probable words of a topic appear together in documents, and has been shown to match well with human evaluation of topic quality (Boyd-Graber et al., 2014).

4.2 Classification of *Standard* Labels

The upper part of Table 1 shows the results of structured vs. linguistic features in isolation for the *Standard* problem, while the rest of the table shows results of integration techniques. Overall, performance improved in LOOCV, with a few exceptions (e.g. *tf-idf*), which is likely due to the greater number of available training instances in this evaluation.

The performance of structured data alone is substantially higher than the majority class baseline, and more so when cognitive assessment features were included (+*cognitive*), as expected. Importantly, the BOW representation for text data achieved similar performance compared to the structured data without cognitive assessment scores, showing that simple text modeling can be useful in the common event that structured data are missing.

The benefit of tf-idf appears inconsistent between held-out and LOOCV evaluations, possibly attributable to differences in document frequency of important terms in the different training data (*dev* vs. *dev+held-out*, respectively).

For LDA, performance was dependent on the number of topics, as seen in Figure 2, with two performance peaks (at $K = 60$ and $K = 85$) surpassing BOW. This supports that dimensionality reduction by LDA can improve performance, but data size may influence results. This is a limitation of using an unsupervised algorithm for a supervised task. Performance differences between held-out and LOOCV indicate overfitting to the *dev* set in particular.

Table 2a shows 5 of the top 10 topics from the 60 topic model, based on topic coherence. This metric appears to aid in identifying interpretable topics. For example, Topic 2 is about cognitive assessment,

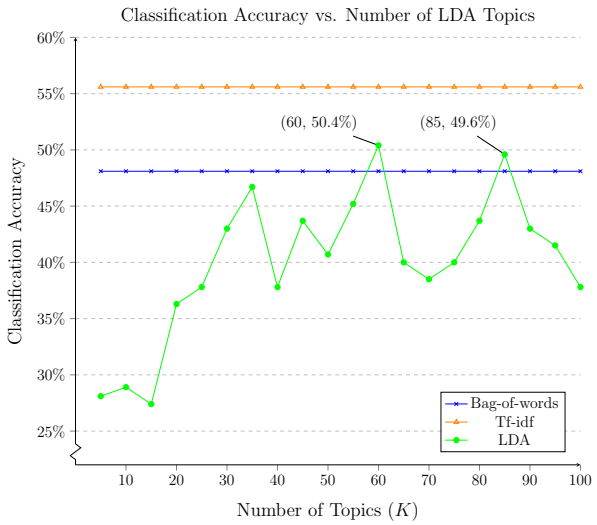

Figure 2: Classification accuracy of LDA features on the *held-out* set, with increasing number of topics K (increments of 5).

referencing people in their 60's. Topic 45 pertains to regular medical visits (*PCP* is *primary care physician*), with some common concerns of elderly patients (*back*, *heart*). Topic 25 captures heart disease (*cardiac*, *stent*, *chest_pain*) and related visits (*hospitalization*, *admitted*, *discharged*).

Linguistic and non-linguistic models are integrated to improve classification performance. Table 1 shows results for 16 integrated models (2 non-linguistic models $\times$ 4 linguistic models $\times$ 2 integration methods). Similar trends were observed for BOW and tf-idf in most cases. Interestingly, integrating with BOW is better than including cognitive assessment scores for held-out. The LDA-reduced features are again less consistent than other text features, but still comparatively improved performance in many cases. LDA integration experiments appear more robust between held-out and LOOCV than when LDA features were used alone, likely due to structured features taking the brunt of the decision.

It was predicted that the posterior probability composition method would yield better results than vector concatenation. Interestingly, this is not apparent, with many cases revealing the opposite. Yet overall, the best performing cases include results where integration is done by this method. One potential limitation of the posterior probability composition is that a stronger decision is made when each of the underlying classifiers produces an asymmetric posterior class distribution. A limitation of

	Held-out Evaluation					Leave-one-out Cross-validation				
		NL	*EMCI*	*LMCI*	*AD*		*NL*	*EMCI*	*LMCI*	*AD*
Features	**Acc.**	**P / R**	**P / R**	**P / R**	**P / R**	**Acc.**	**P / R**	**P / R**	**P / R**	**P / R**
Baseline (majority class)	32.6%	33 / 100	– / 0	– / 0	– / 0	27.5%	28 / 100	– / 0	– / 0	– / 0
Structured (−cognitive)	51.9%	68 / 73	27 / 28	47 / 23	55 / 88	53.9%	57 / 77	43 / 34	40 / 26	66 / 81
Structured (+cognitive)	55.6%	80 / 84	35 / 38	33 / 20	56 / 79	62.7%	70 / 86	52 / 48	50 / 33	71 / 85
Bag-of-words	48.1%	67 / 55	32 / 38	52 / 46	43 / 54	50.2%	59 / 67	40 / 39	43 / 42	60 / 51
Tf-idf	55.6%	61 / 61	37 / 63	78 / 40	74 / 58	48.9%	49 / 75	39 / 43	49 / 32	73 / 42
LDA($K = 85$)	49.6%	57 / 48	39 / 72	65 / 37	53 / 42	39.3%	39 / 62	34 / 32	39 / 29	52 / 32
LDA($K = 60$)	50.4%	64 / 61	37 / 66	53 / 23	57 / 50	37.4%	39 / 54	32 / 33	35 / 28	48 / 33
S_{-cog} ∪ Bag-of-words	61.5%	77 / 68	41 / 50	70 / 46	62 / 88	59.8%	69 / 79	48 / 44	45 / 41	73 / 76
S_{-cog} ⊕ Bag-of-words	57.0%	90 / 59	41 / 53	47 / 40	59 / 83	58.3%	69 / 73	46 / 46	45 / 44	75 / 71
S_{+cog} ∪ Bag-of-words	58.5%	78 / 71	35 / 41	59 / 46	61 / 79	61.3%	72 / 79	48 / 43	47 / 44	74 / 80
S_{+cog} ⊕ Bag-of-words	59.3%	88 / 64	39 / 53	56 / 43	63 / 83	61.9%	74 / 80	48 / 48	48 / 45	77 / 75
S_{-cog} ∪ Tf-idf	53.3%	74 / 71	31 / 34	45 / 26	55 / 88	58.0%	62 / 83	49 / 38	45 / 31	68 / 81
S_{-cog} ⊕ Tf-idf	51.1%	83 / 55	37 / 59	39 / 14	51 / 88	59.6%	63 / 83	52 / 43	46 / 30	70 / 82
S_{+cog} ∪ Tf-idf	59.3%	79 / 86	41 / 44	45 / 26	58 / 79	64.7%	73 / 88	53 / 53	52 / 34	72 / 84
S_{+cog} ⊕ Tf-idf	61.5%	95 / 80	45 / 72	42 / 14	57 / 83	65.4%	73 / 89	55 / 53	54 / 35	73 / 85
S_{-cog} ∪ LDA($K = 85$)	54.8%	73 / 73	31 / 34	56 / 29	55 / 88	56.4%	60 / 82	46 / 33	42 / 31	70 / 80
S_{-cog} ⊕ LDA($K = 85$)	44.4%	80 / 46	28 / 50	27 / 09	50 / 88	56.3%	60 / 79	45 / 36	42 / 30	71 / 81
S_{+cog} ∪ LDA($K = 85$)	58.5%	84 / 86	39 / 44	38 / 23	58 / 79	62.0%	70 / 86	49 / 45	46 / 34	74 / 84
S_{+cog} ⊕ LDA($K = 85$)	58.5%	90 / 77	44 / 69	36 / 11	53 / 79	63.6%	71 / 87	52 / 48	49 / 35	75 / 85
S_{-cog} ∪ LDA($K = 60$)	51.1%	69 / 61	30 / 34	48 / 29	55 / 88	55.7%	59 / 78	47 / 37	40 / 28	69 / 82
S_{-cog} ⊕ LDA($K = 60$)	45.9%	78 / 48	30 / 53	33 / 09	49 / 88	56.4%	60 / 78	47 / 37	40 / 30	71 / 82
S_{+cog} ∪ LDA($K = 60$)	60.0%	88 / 86	44 / 53	37 / 20	56 / 79	62.4%	72 / 86	50 / 47	45 / 33	74 / 85
S_{+cog} ⊕ LDA($K = 60$)	59.3%	92 / 77	45 / 72	33 / 11	54 / 79	62.9%	74 / 86	51 / 48	44 / 34	75 / 84

Table 1: Results on *Standard* problem (4-classes). Integration by *vector concatenation* is indicated by ∪, and *posterior probability composition* by ⊕. Structured (−cognitive) = S_{-cog} , Structured (+cognitive) = S_{+cog}.

ID	Top 10 Words
3	*corroborated, subjective, continues_meet, score, factor, other, SP, AGE_>=60_<70, controlled_medication, unremarkable*
2	*impression, CDR, MMSE, ADLS, AGE_>=60_<70, cog, amnestic, global, function, score*
17	*medical, consistent, status, function, continues, health, occasional, active, daily, functional*
45	*blood, pressure, month, visit, PCP, diagnosed, dizziness, back, doctor, heart*
25	*hospital, admitted, discharged, stent, cardiac, went, chest_pain, AE, anxiety, total*

Table 2a: Five high-ranked topics from the *Standard* problem with $K = 60$ (ranked by *topic coherence*).

ID	Top 10 Words
38	*completed, visit, reported, mg, performed, protocol, testing, study_partner, blood, year*
25	*criterion, subjective, corroborated, factor, other, AGE_>=60_<70, continues_meet, score, memory_problems, confounding*
55	*hip, left, right, removed, normal, arthritis, cataract, eye, allergy, hand*
36	*year, smoked, ago, pack, o, quit, per_day, c, urinary_frequency, memory_problems*
56	*work, up, valve, cardiac, aortic, ER, heart, x, cardiologist, visit*

Table 2b: Five high-ranked topics from the *Early Risk* problem with $K = 100$ (ranked by *topic coherence*).

this method is its dependence on strong or accurate decisions from the underlying models. Vector concatenation is not subject to this limitation, but has the drawback of potentially overwhelming a smaller feature set with a larger sparse one. As for class-specific differences, the *NL* (normal) and *AD* (Alzheimer's disease) subjects were classified with higher precision and recall scores than were the MCI classes in nearly all integration experiments, pointing to the challenge of subtler disease stages.

4.3 Classification of *Early Risk*

In addition to the experiments above, the more specific problem of distinguishing normal (*NL*) subjects from those with early mild cognitive impairment (*EMCI*) was also explored. Only LOOCV is performed because the subsampling of *NL* and *EMCI* subjects slightly distorts the class distributions in the original held-out set. Results are given in Table 3.

As in the *Standard* problem, all non-linguistic and linguistic feature types perform well above the majority class baseline. One major difference here is that all linguistic data types outperform the structured features when cognitive assessments are excluded. This may suggest a potential linguistic difference in clinical notes at the onset of MCI.

The number of LDA topics is selected as before (but using the whole *Early Risk* subsample, as opposed to the *Standard dev* set). Two peaks found at $K = 65$ and $K = 100$ achieve the same classification accuracy, but do not outperform BOW and tf-idf. The difficulties LDA faced in the *Standard* problem are also faced here, and thus similar performance shortcomings are observed. The ability to approximately match tf-idf performance is still noteworthy since the LDA features are a smaller and denser representation than tf-idf, which may be more easily interpretable by clinical professionals.

Table 2b shows 5 of the top 10 topics from the 100 topic model trained on the *Early Risk* subset, based on the topic coherence metric. A consequence of a smaller sample of subjects is a smaller vocabulary and thus weaker statistical judgments, Topics 38, 25, 36, and 56 appear to be about routine visits/tests, cognitive evaluations, smoking habits, and cardiac issues, respectively. Topic 55 is an example of a *chained* topic (Boyd-Graber et al., 2014, p. 17), where unrelated words are linked together through shared co-occurring words, in this case with *left* and *right* seeming to link *eye* and *hand*, along with their associated terms *cataract* and *arthritis*.

	LOOCV		
		NL	*EMCI*
Features	**Acc.**	**P / R**	**P / R**
Baseline	51.0%	51 / 100	– / 0
Structured (−cognitive)	67.6%	67 / 73	69 / 62
Structured (+cognitive)	79.8%	78 / 84	82 / 76
Bag-of-words	70.8%	71 / 73	71 / 69
Tf-idf	69.2%	68 / 75	71 / 63
LDA($K = 65$)	68.9%	67 / 76	71 / 62
LDA($K = 100$)	68.9%	68 / 74	70 / 63
$S_{-cog} \cup$ Bag-of-words	76.8%	76 / 79	78 / 74
$S_{-cog} \oplus$ Bag-of-words	76.0%	77 / 77	76 / 76
$S_{+cog} \cup$ Bag-of-words	77.1%	76 / 80	78 / 74
$S_{+cog} \oplus$ Bag-of-words	80.7%	80 / 82	81 / 79
$S_{-cog} \cup$ Tf-idf	72.2%	71 / 78	74 / 66
$S_{-cog} \oplus$ Tf-idf	72.8%	71 / 79	75 / 66
$S_{+cog} \cup$ Tf-idf	80.7%	79 / 85	83 / 77
$S_{+cog} \oplus$ Tf-idf	83.1%	82 / 86	84 / 81
$S_{-cog} \cup$ LDA($K = 65$)	72.2%	71 / 78	74 / 67
$S_{-cog} \oplus$ LDA($K = 65$)	72.5%	71 / 78	74 / 67
$S_{+cog} \cup$ LDA($K = 65$)	79.0%	78 / 82	81 / 76
$S_{+cog} \oplus$ LDA($K = 65$)	79.3%	78 / 83	81 / 76
$S_{-cog} \cup$ LDA($K = 100$)	71.4%	70 / 77	73 / 66
$S_{-cog} \oplus$ LDA($K = 100$)	71.9%	70 / 76	74 / 66
$S_{+cog} \cup$ LDA($K = 100$)	80.4%	80 / 82	81 / 78
$S_{+cog} \oplus$ LDA($K = 100$)	80.9%	80 / 83	82 / 79

Table 3: Classification performance on *Early Risk* (2 classes). *Vector concatenation* is indicated by $\cup$, and *posterior probability composition* by $\oplus$. Structured with (S_{+cog}) and without (S_{-cog}) cognitive.

The performance trends for the integrated models are slightly more consistent for the *Early Risk* problem than they were for the *Standard* problem. When excluding cognitive assessment scores, all integration experiments result in a modest improvement, although there is little to no difference between the two integration methods employed. This may suggest that results can be achieved without extra sophistication provided by posterior probability composition, or that further sophistication is needed beyond either of these techniques. In general, our results further justify the integration of linguistic and non-linguistic features and/or models.

5 Conclusion and Future Work

We explored classification of dementia progression status of subjects from a study on Alzheimer's disease, and the integration of text data models with those of structured data, with vs. without cognitive assessment scores. Experiments support texts' viability as a useful source for dementia classification, as an important complement to structured data, or alone when structured data are missing. LDA was also studied as interpretable dimensionality reduction. With a larger sample size, the LDA model may converge to a more stable set of topics, but other appropriate public datasets (with both linguistic and non-linguistic data) are presently not available. An alternative is to apply supervised versions of LDA (Blei and McAuliffe, 2007; Ramage et al., 2009). Furthermore, with access to a pool of clinical specialists, it would be useful to integrate experts in evaluating the latent topics. Chang et al. (2009) proposed various such human evaluation techniques, such as the *word intrusion* task, in which human evaluators are presented with a list of n high probability terms of a randomly chosen topic, and one additional low probability term from that topic, and asked to identify the former. A drawback is that it would require access to a large enough pool of dementia specialists.

Other avenues of future work would include the incorporation of lexical similarity measures from sources like WordNet.

Acknowledgement

This study was supported by a Kodak Endowed Chair award form the Golisano College of Computing and Information Sciences at the Rochester Institute of Technology.

We are thankful for being able to use the ADNI projects dataset in this work. As per the data-use agreement, information about the ADNI data are included verbatim. Data collection and sharing for this project was funded by the Alzheimers Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimers Association; Alzheimers Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimers Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Alzheimer's Association. 2014. 2014 Alzheimer's Disease Facts and Figures. *Alzheimer's & Dementia*, 10.

C.A.L. Bailer-Jones and K. Smith. 2011. Combining probabilities. GAIA-C8-TN-MPIA-CBJ-053, July.

Elena Birman-Deych, Amy D. Waterman, Yan Yan, David S. Nilasema, Martha J. Radford, and Brian F. Gage. 2005. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43:480–485.

David M. Blei and Jon D. McAuliffe. 2007. Supervised topic models. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems*, pages 121–128, Vancouver, B.C., Canada.

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

Jordan Boyd-Graber, David Mimno, and David Newman, 2014. *Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements*. CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, Florida.

Roger B Bradford. 2008. An empirical study of required dimensionality for large-scale latent semantic indexing applications. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 153–162.

Katherine Redfield Chan, Xinghua Lou, Theofanis Karaletsos, Christopher Crosbie, Stuart Gardos, David Artz, and Gunnar Rätsch. 2013. An empirical analysis of topic modeling for mining cancer clinical notes. In *13th IEEE International Conference on Data Mining Workshops*, pages 56–63, Dallas, Texas, December 7–10.

Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*, Vancouver, British Columbia.

Marcelo Fiszman, Wendy Webber Chapman, Dominik Aronsky, R. Scott Evans, and Peter J. Haug. 2000. Automatic detection of acute bacterial pneumonia from chest x-ray reports. *Journal of the American Medical Informatics Association*, 7(6):593–604.

Carol Friedman, Stephen B. Johnson, Bruce Forman, and Justin Starren. 1995. Architectural requirements for a multipurpose natural language processor in the clinical environment. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 347–351.

Peter J. Haug, Spence Koehler, Lee Min Lau, Ping Wang, Roberto Rocha, and Stanley M. Huff. 1995. Experience with a mixed semantic/syntactic parser. In *Proceedings of the Annual Symposium on Computational Application in Medical Care*, pages 284–288.

Blanca E. Himes, Yi Dai, Isaac S. Kohane, Scott T. Weiss, and Marco F. Ramoni. 2009. Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records. *Journal of the American Medical Informatics Association*, 16:371–379.

Liangjie Hong and Brian D. Davison. 2010. Empirical study of topic modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88, Washington DC, USA.

Nilesh L. Jain and Carol Friedman. 1997. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. In *Proceedings of the American Medical Informatics Association Annual Fall Symposium: American Medical Informatics Association*, pages 829–833.

Elizabeth F. O. Kern, Miriam Maney, Donald R. Miller, Chin-Lin Tseng, Anjali Tiwari, Mangala Rajan, David Aron, and Leonard Pogach. 2006. Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. *Health Services Research*, 41(2):564–580.

Li Li, Herbert S. Chase, Chintan O. Patel, Carol Friedman, and Chunhua Weng. 2008. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: A case study. In *American Medical Informatics Association Annual Symposium Proceedings 2008*, pages 404–408.

Stephen Luther, Donald Berndt, Dezon Finch, Michael Richardson, Edward Hickling, and David Hickam. 2011. Using statistical text mining to supplement the development of an ontology. *Journal of Biomedical Informatics*, 44:S86–S93.

Ashutosh Malhotra, Erfan Younesi, Michaela Gündel, Müller, Michael T. Heneka, and Martin Hofmann-Apitius. 2013. ADO: A disease ontology representing the domain knowledge specific to Alzheimer's disease. *Alzheimer's and Dementia*, 10:238–246.

Andrew McCallum, Xuerui Wang, and Andrés Corrada-Emmanuel. 2007. Topic and role discovery in social networks with experiments on Enron and academic email. *Journal of Artificial Intelligence Research*, 30(1):249–272, Oct.

James A. McCart, Donald J. Berndt, Jay Jarman, Dezon K. Finch, and Stephen Luther. 2013. Finding falls in ambulatory care clinical documents using statistical text mining. *The Journal of American Medical Informatics Association*, 20(5):906–914.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 262–272, Edinburgh, United Kingdom.

Harvey J. Murff, Fern FitzHenry, Michael E. Matheny, Nancy Gentry, Kristen L. Kotter, Kimberly Crimin, S. Dittus, Robert, Amy K. Rosen, Peter L. Elkin, Steven H. Brown, and Theodore Speroff. 2011. Automated identification of postoperative complications within an electronic medical record using natural language processing. *The American Journal of Medicine*, 306(8):848–855, August.

Michael J. Paul and Mark Dredze. 2011. You are what you tweet: Analyzing Twitter for public health. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, pages 265–272, Barcelona, Catalonia, Spain, July 17–21.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Ma-

chine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Janet F. E. Penz, Adam B. Wilcox, and John F. Hurdle. 2007. Automated identification of adverse events related to central venous catheters. *Journal of Biomedical Informatics*, 40(2):174–182, April.

Martin Prince, Matthew Prina, and Maëlenn Guerchet. 2013. *World Alzheimer Report 2013*. Alzheimer's Disease International (ADI), London, September.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. 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: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, pages 1348–1353, Seattle, Washington, USA, 18–21 October.

Dymitr Ruta and Bogdan Gabrys. 2000. An overview of classifier fusion methods. *Computing and Information Systems*, 7:1–10.

Yee Whye Teh, David Newman, and Max Welling. 2007. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 19.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

Hua Xu, Zhenming Fu, Anushi Shah, Yukun Chen, Neeraja B. Peterson, Qingxia Chen, Subramani Mani, Mia A. Levy, Qi Dai, and Josh C. Denny. 2011. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. In *American Medical Informatics Association Annual Symposium Proceedings 2011*, pages 1564–1572, Washington, DC, USA.