3-4-2015

Mining and Integration of Structured and Unstructured Electronic Clinical Data for Dementia Detection

Joseph Bullard
jtb4478@rit.edu

Mining and Integration of Structured and Unstructured Electronic Clinical Data for Dementia Detection

Approved by Committee:

Dr. Cecilia Ovesdotter Alm, Chair

College of Liberal Arts

Dr. Xumin Liu, Co-chair

Deptartment of Computer Science

Dr. Qi Yu, Reader

College of Computing and Information Sciences

Dr. Richard Zanibbi, Observer

Department of Computer Science

Mining and Integration of Structured and Unstructured Electronic Clinical Data for Dementia Detection

by

Joseph Bullard

THESIS

Presented to the faculty of

the B. Thomas Golisano College of Computing and Information Sciences

Department of Computer Science

Rochester Institute of Technology

in partial fulfillment

of the requirements

for the degree of

Master of Science

Rochester Institute of Technology

March 4, 2015

Abstract

Dementia is an increasing problem for the aging population that incurs high medical costs, in part due to the lack of available treatment options. Accordingly, early detection is critical to potentially postpone symptoms and to prepare both healthcare providers and families for a patient's management needs. Current detection methods are typically costly or unreliable, and could greatly benefit from improved recognition of early dementia markers. Identification of such markers may be possible through computational analysis of patients' electronic clinical records. Prior work on has focused on structured data (e.g. test results), but these records often also contain natural language (text) data in the form of patient histories, visit summaries, or other notes, which may be valuable for disease prediction. This thesis has three main goals: to incorporate analysis of the aforementioned electronic medical texts into predictive models of dementia development, to explore the use of topic modeling as a form of interpretable dimensionality reduction to improve prediction and to characterize the texts, and to integrate these models with ones using structured data. This kind of computational modeling could be used in an automated screening system to identify and flag potentially problematic patients for assessment by clinicians. Results support the potential for unstructured clinical text data both as standalone predictors of dementia status when structured data are missing, and as complements to structured data.

Contents

List of Figures

List of Tables

1 Introduction

Dementia is an increasing problem for the aging population, and the 6th leading cause of death in the US [Alzheimer's Association, 2014]. Approximately 35 million people worldwide suffer from some form of dementia, and this number is expected to double by the year 2030 [Prince et al., 2013]. The most common form of dementia is Alzheimer's disease, which has no known cure and has limited treatment options. Thus 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. The cost of Alzheimer's disease care for the year 2014 is estimated at $214 billion dollars in the US [Alzheimer's Association, 2014]. 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. Improved understanding and recognition of early warning signs of dementia would greatly benefit detection and management of the disease.

With the advent of electronic clinical record-keeping comes the potential for large-scale computational analysis of patients' clinical data to understand or discover warning signs and development of medical conditions. At a coarse-grained level, data can be considered either *structured* or *unstructured*. The former refers to numerical or categorical data, such as test results or patient demographics, while the latter generally refers to text data, such as doctors' notes or summaries. Each of these data types provides its own set of challenges and

benefits, but most prior research has focused on structured data. Unstructured text, however, presents a potentially rich source of information that may be more easily interpretable by humans. The ability to predict dementia development based on either or both of these data sources in patients' electronic records would be useful for intelligent support systems which could automatically flag potentially problematic cases for further human evaluation, reducing the need for laborious manual inspection, as well as the risk of missed dementia cases. Furthermore, integration of these two data types may provide additional benefits for such intelligent support systems. From a computer science perspective, surmounting the technical challenge of data fusion contributes to predictive modeling in the medical domain.

The inclusion of unstructured text data is a critical contribution of this thesis. Structured data will often be absent from clinical records due to issues of cost or availability, whereas text notes will be present for nearly every visit of a patient. Moreover, text notes in medical records are a source of natural language which potentially more flexibly encode the diagnostic expertise and reasoning of the clinical professionals who write them. Processing and computationally analyzing natural language remains a formidable task, but insights gleaned from it may also be more intuitively interpretable by humans, and thus may translate better into actual clinical practice. In particular, this thesis models text in three stages: bag-of-words, tf-idf weighting, and topic modeling with Latent Dirichlet Allocation (LDA). In *bag-of-words*, a text document is represented in terms of the words it contains, along with their raw frequencies in that document. *Tf-idf*, or *term frequency - inverse document frequency*, is an extension on top of bag-of-words in which words (terms) are weighted based on a combination of their frequency within a given document and their frequency throughout the corpus, rewarding words which appear more times in fewer documents.[1] *Latent Dirichlet Allocation*, or *LDA*, is a topic modeling algorithm which attempts to infer groups (or *topics*) of statistically related words in a corpus of documents, and is explored here as a form of textually interpretable dimensionality reduction. Each of these processes is explained in detail later

[1]While tf-idf involves a form of bag-of-words model, in this thesis the term *bag-of-words* is used for the term frequency scenario.

in Section 6.1. This thesis addresses the question of whether or not these data and analyses can be useful for a supervised machine learning task of classifying the dementia-progression status of subjects in a study on Alzheimer's disease.

In addition to these contributions, this thesis also explores the *integration* of unstructured data with structured data through two different methods. The inputs for classification are vectors of feature values for every data point. The first, more straightforward integration method takes advantage of this common format by combining the vectors of features computed independently from structured and unstructured data. The second, more sophisticated integration method instead leverages probabilistic outputs of two classification models, one for each data type in isolation. The results of these experiments constitute methodological and practical contributions to data mining of electronic clinical records, as well as semantic data integration techniques in general.

This thesis is organized as follows. Chapter 2 elucidates the necessary background information about electronic medical data, domain-specific methodological considerations, and modeling techniques needed to comprehend the work presented in subsequent chapters. Chapter 3 orients the reader through a review of the literature on medical data mining, shedding light on unanswered research questions and further motivating this thesis. Chapter 4 describes the source and characteristics of the dataset, followed by an exploratory overview of the text corpus constructed from it. Chapter 5 formally outlines the problems to be addressed and hypothesizes their experimental outcomes. Chapter 6 lays out the complete details of all methodology, including data modeling, implementation decisions, and evaluation procedures. Chapter 7 presents and discusses the results of performed experiments. Finally, Chapter 8 summarizes the conclusions, contributions, limitations, and potential future work.

2 Background

This chapter provides an overview of concepts needed to understand the work presented in this thesis. Namely, it is important for the reader to understand what kinds of data are available in electronic clinical records, what distinctions can be made among them, and how they fit into the context of studying dementia and Alzheimer's disease. Additionally, some background on supervised machine learning and text modeling are also required.

2.1 Electronic Clinical Records

Electronic clinical records are digital collections of information obtained from clinical services received by patients. As with paper records, these are typically organized chronologically with metadata indicating the calendar dates of when each piece of information was entered. Data contained in these records are typically distinguished by the format of their stored representations, with numerical and categorical data termed *structured*, and free text termed *unstructured*. Examples of structured data include patient demographics, such as age, sex, or ethnic background, as well as medical test results, such as routine blood work, measurements from imaging scans, or scores on administered verbal exams and questionnaires. This straightforward numerical representation makes structured data attractive for computational studies. However, the rigid nature of this tabular structure may fail to accommodate important pieces of information when considering a specific and poorly understood medical interest such as dementia or Alzheimer's disease. Alternatively, unstructured text data consist of notes and summaries written by doctors or other clinical professionals who treat,

care for, or otherwise interact with a patient. From a computational perspective, handling text (i.e. natural language) data is a challenging and non-trivial task. Language use, even in the domain of medicine, can vary drastically between two users due to the ambiguous and expressive nature of language itself, as well as domain-specific factors, such as different medical school educations or backgrounds of the clinicians, recent vs. dated terminology, non-standard or personal abbreviations, preferences for ellipsis, and so on. Despite these difficulties, natural language is arguably a more interpretable format for humans and may contain information that is not feasibly encoded in a structured format (e.g. with respect to social context or behavioral health).

2.1.1 Data Considerations for Studying Dementia

When applying computational techniques to problems in the medical domain, it is important to consider the nature of the clinical condition being examined, as its specific characteristics may impact the usability or efficacy of certain data and methodology. This is particularly true in the case of dementia and Alzheimer's disease. As mentioned earlier, the distinction between structured and unstructured data is based on digital representation of the information, as either a number/category or as free text, respectively. However, this thesis is not merely concerned with data processing, but with the applicability and meaningfulness of the results obtained through computational means, which may be enhanced by finer-grained distinctions, particularly within the structured data. For example, dementia is typically diagnosed based on lab tests and/or scores on verbally administered cognitive exams. Both of these are considered structured data since they are both represented as metrics or scores, yet they are fundamentally different in how they are administered. A cognitive exam involves the intervention of another person's mind, experience, and expertise to evaluate a patient's cognition, as opposed to a typical lab test, which measures quantities in or of a patient's body. Benefits of this distinction might fail to be detected if data were only distinguished by their structured vs. unstructured format. This particular example comes into play later in

this thesis, where experimental work with integrating structured data will specifically report on including vs. excluding cognitive assessment scores.

Another potential issue that arises when studying dementia through clinical records is the availability of relevant structured data. Two major categories of dementia-related tests are cerebrospinal fluid markers and brain volume measurements obtained from imaging scans like Magnetic Resonance Imaging (MRI), which can be considered invasive and expensive. Not surprisingly, such test results are often missing from records of patients who are not already suspected of having dementia, and too infrequent among the patients who are. This is a major motivation for the inclusion of unstructured text data, which does not suffer the same availability issues, as note-taking and summarization of visits is standard practice in clinical settings, and is thus usually available for nearly all patient visits.

Finally, dementia is a cognitive condition also characterized by behavioral and social changes of an afflicted individual. Such changes may be addressed by the cognitive exams mentioned earlier, but those are typically not administered until later in disease progression. It is possible that such information may already be encoded or referenced in a patient's notes before the disease takes a turn for the worst. If so, then this would constitute another potential benefit of utilizing unstructured text data. Furthermore, dementia and Alzheimer's disease are not particularly well-understood, and it is possible that the broad and unbounded nature of clinical text may allow for the discovery of new meaningful warning signs.

2.2 Classification

In supervised machine learning, a classification task uses a set of data instances, each with a corresponding label from a pre-determined set of classes, to learn a model that can accurately predict those labels. Each instance is represented as a vector of values obtained from feature functions. Features are usually defined based on knowledge of the problem at hand to appropriately represent each data instance. These vectors serve as inputs for classification.

Formally, we have a collection

$$\{(x_i, y_i) \mid x_i \in \mathbb{R}^p \ , \ y_i \in C\}_{i=1}^n$$

where each data instance x_i is a p-dimensional feature vector and y_i is its associated class label. A classification algorithm is trained on a subset of this data to find boundaries that divide the data instances by class label. The trained model can then be used to predict the labels of another subset of the data instances, and evaluated based on how well it does so.

Examples of classification algorithms include the Support Vector Machine (SVM), decision tree, and k-nearest-neighbors. Some algorithms produce a probability distribution over the set of class labels, called a *posterior* distribution (as opposed to the *prior* distribution of the class labels before any modeling) and output the label with the highest probability. This is the case for logistic regression, which is the algorithm of choice for this thesis, for reasons explained in Chapter 6.

2.2.1 Text Modeling

Of particular interest to this thesis is the incorporation and integration of unstructured text data for classification purposes. In the field of natural language processing, a collection of text documents is referred to as a *corpus* (plural *corpora*). Transforming a corpus into a set of feature vectors for classification often involves representing each document in terms of the words they contain. It is common for these representations to make use of the *bag-of-words* assumption, which treats documents as collections of the unique word types contained within, irrespective of word-order. Weighting schemes such as *tf-idf* (*term frequency - inverse document frequency*) then extend this representation based on corpus-wide calculations. In any case, these representations are sparse, as documents will only contain a small subset of corpus-wide vocabulary. This thesis further makes use of the topic modeling algorithm *Latent Dirichlet Allocation* (*LDA*) to address this sparsity, both to explore performance in classification and to provide a more interpretable representation of the documents.

2.2.2 Topic Modeling

Topic modeling algorithms aim to discover groups of related words in a collection of documents. The idea behind topic modeling is rather intuitive: the probability of a word occurring may be higher or lower depending on what is being discussed or written about. For example, a topic about dermatology may be more likely to contain the words *macule* or *melanoma* than a topic about dentistry, which may be more likely to contain the words *enamel* or *gingivitis*. However, some words may have roughly the same probability of occurrence regardless of the topic. This includes not only function words, such as *the* or *and*, which will occur in nearly every English text, but also words that are shared between a set of topics, such as *treatment* or *visit* in the previous example. A document may have multiple prominent topics that are active at different times, and the probability of each word changes accordingly.

More formally, topic modeling is about discovering the hidden thematic structure of documents by working backwards from statistical observations of the words within a corpus. The algorithm employed in this thesis is Latent Dirichlet Allocation (LDA), described in detail in Section 6.1.4. In LDA, each *topic* is a probability distribution over a set of words (a vocabulary), and each document is a distribution over a set of topics. Topics are derived through statistical inference over the corpus of documents and reflect the language usage patterns found within. This contributes to facilitating the interpretability of topics, however, it does not necessarily mean that every single topic in its raw form (word distribution) will be as easily understandable as the examples above. Improving the human interpretability of topics can be accomplished through empirical selection of the number of topics to avoid extraneous granularity, as well as through text normalization techniques that reduce irrelevant noise in the corpus. However, the presence of some nonsensical topics is an expected and established outcome of LDA. For example, high-frequency general words may end up aggregating into non-specific meaningless topics, but this is actually beneficial

for the model as a whole because it allows the more interesting content words to form their own topics [Boyd-Graber et al., 2014]. Various quality metrics for individual topics have been defined in the literature, a few of which are selected for implementation in Section 6.3. It is also important to note that a lack of immediate intuition about a topic in a specific domain like medicine does not necessarily mean that it is faulty; it is entirely possible for new term relationships to be identified, but this type of outcome must be carefully evaluated. Importantly, LDA constitutes a form of dimensionality reduction for the sparse bag-of-words representation. The reduced topic representation of documents will be small and dense, as well as easy to visualize and inspect, making it potentially more interpretable to humans than other forms of dimensionality reduction.

3 Related Work

The potential of data mining in the medical domain has been recognized for some time. Utilizing structured clinical data, such as patient demographics and test results, is intuitive and may be useful for predicting certain disease cases based on known markers [Himes et al., 2009]. Most prominently, medical records often make use of standardized coding schemes, such as the International Classification of Diseases version 9 (ICD-9) codes, which can provide high specificity for a given disease, but may not provide sufficient sensitivity [Birman-Deych et al., 2005, Kern et al., 2006]. That is to say, the presence of a certain disease or symptom code may strongly indicate the presence of a particular condition, but the absence of that code may not necessarily indicate the absence of that condition. In such cases, a prediction based on merely that code could fail to identify many afflicted or at-risk patients. A patient's history often plays a critical role in diagnosis, but ICD codes are typically assigned upon admission or discharge, and thus may not be present for past conditions. However, historical information is typically summarized by a clinician in text form, especially when the patient has no existing electronic clinical health record. This natural language data provides a level of expressiveness and granularity not feasibly represented by ICD-9 codes [Li et al., 2008].

Natural language processing (NLP) and text mining techniques have been applied to texts from electronic clinical records in the past, with a focus on extraction of known disease markers obtained from medical knowledge sources. One such method is to identify relevant terms using an ontology (curated knowledge base), such as SNOMED-CT (Systematized Nomenclature of Medicine - Clinical Terms), which maps terms to their corresponding

medical concepts. This method was employed by Murff et al. [2011] for predicting post-operative complications for patients who recently underwent medical procedures, finding that it outperformed the health records' default Patient Safety Indicators system (tagging based on standardized ICD-9 diagnostic codes), in terms of both sensitivity and specificity performance metrics. Another common practice is to use a program like MedLEE (Medical Language Extraction and Encoding system) [Friedman et al., 1995] to automatically extract and codify medical terms and concepts in a text, then process that information as needed. MedLEE has provided impressive results for identifying cases of colorectal cancer [Xu et al., 2011], suspicious mammogram findings [Jain and Friedman, 1997], and adverse events related to central venous catheters [Penz et al., 2007]. SymText [Haug et al., 1995] is another example of a medical language extraction tool that has been used for similar purposes, such as detecting bacterial pneumonia cases from chest X-ray descriptions, achieving performance comparable to physician-evaluated gold standards [Fiszman et al., 2000].

In general, systems that utilize medical knowledge bases like these seem to be useful towards disease case identification. However, it is important to note that these studies mostly dealt with short-term and/or inspection-based medical events that are relatively easy for a human clinician to detect based on a single text report. Additionally, software like MedLEE and SymText rely on known word and clinical concept relations, which may be partly responsible for their high performance on more well-understood medical conditions, such as the ones above. A problem is that many diseases and conditions of interest, such as dementia (the focus of this thesis) and other cognitive or mental illnesses, are not necessarily as well-understood and thus may not be as adequately predicted by existing knowledge bases. Discovery of new terms associated with these afflictions could be more beneficial than simply attempting to predict them using the limited knowledge already available.

Text mining algorithms such as Latent Semantic Indexing (LSI) aim to discover statistical relationships between words in a corpus which can then further be related to disease states for a particular patient's clinical texts. Luther et al. [2011] used LSI to supplement the

development of a clinical vocabulary of terms associated with post-traumatic stress disorder, which was able to identify more unique terms than a model based on the SNOMED-CT vocabulary. To further specialize their results, the authors devised different term categories, including symptoms, medications, and traumas - resembling common divisions of text notes in many medical record systems. LSI can also be used for classification, as was done by McCart et al. [2013] to predict ambulatory falls in elderly patients.

Although LSI provides the benefit of identifying novel important terms in text corpora, it often requires around 300–500 dimensions to produce stable results [Bradford, 2008]. This is a considerable improvement over the dimensionality of the bag-of-words representation for classification purposes, but could potentially be improved further by using topic modeling algorithms, namely Latent Dirichlet Allocation (LDA). As discussed in previous chapters, the goal of topic modeling is to identify groups of related terms that also may be more intuitive for human interpretation than the latent dimensions produced from LSI. Additionally, representing each document by its topic distribution, as is typically done in the literature, will reduce the classification feature space to even fewer dimensions (i.e. if there are k topics, then each document has k features). In general, topic modeling has produced interesting results in medical and non-medical domains. Chan et al. [2013] used LDA on health record texts to find topics relating to genetic mutations. Resnik et al. [2013] used LDA to improve performance over other semantic category features when predicting neuroticism and depression in college students' self-reflective essays. More commonly, LDA is employed in modeling of social media data [McCallum et al., 2007, Paul and Dredze, 2011]. One relevant study by Hong and Davison [2010] using Twitter data demonstrated that document length can influence topic models, and that aggregating short documents on a per-author basis can result in improvement. This finding is useful for the present work due to the prevalence of similarly short text documents in the clinical dataset.

A theme in this thesis is interpretability, and the nature of these different feature types in this context is of interest for analysis. As will be described in later sections, the unstructured

and structured data are treated here as separate sources of information, and models based on each can be integrated[1] as distinct units. Ruta and Gabrys [2000] provides an overview of various techniques, some of which are more applicable here than others. For example, voting is the typical method of classifier integration, in which a final decision is determined by a majority or weighted tally of the outputs from a population of trained classifiers. This is common in ensemble methods such as random forests, but is not the best choice here, since one classifier each is considered for structured and unstructured data, making this kind of voting problematic. When using classifiers that rank class outputs by likelihood (i.e. they produce some kind of posterior probability distribution), the more interesting Borda Count voting method could be used. In this method, a class is re-ranked based on how many classes rank below it in each classifier. This is also not suitable for a two-classifier system because it would be easy for two classes to receive the same Borda Count (e.g. with $A > B > C$ and $B > A > C$, both A and B would have an equal Borda Count of 3). More importantly, the two classifiers utilize two different inputs (structured and unstructured data features), as opposed to multiple classifiers trained on the same inputs. The most appropriate solution under these circumstances is to leverage Bayesian probability and work with classifiers that produce posterior distributions [Bailer-Jones and Smith, 2011]. Assuming conditional independence of the two input types with respect to a given class allows for the probabilities to be combined, the exact equations for which are given in Section 6.1.6. This approach is less common in the literature because of uniqueness of this application; typical problems will use more traditional ensemble techniques like the ones mentioned above.

[1]Here, *integration* refers to techniques for utilizing both unstructured text data and structured data simultaneously for the purpose of supervised machine learning. It is unrelated to integration in the sense of creating databases or tools for storing or querying these data types together. The term *fusion* is sometimes used in the literature [Ruta and Gabrys, 2000].

4 Dataset

This chapter describes, characterizes, and explores the dataset utilized for this thesis, and details how it was used to create a text corpus. The dataset was chosen because it is openly available and approved for research purposes, and relates to the condition of interest.

4.1 Alzheimer's Disease Neuroimaging Initiative

The dataset used here was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (`adni.loni.usc.edu`). The methods employed here constitute a secondary use of the data for a purpose that is in line with the general goal of identifying dementia markers. The following two paragraphs are included verbatim, as required by the ADNI Data Use Agreement.

The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see ***www.adni-info.org***.

4.2 Characteristics of the ADNI Dataset

Most of the data fields present in the ADNI study contain structured data, such as measurements from brain imaging scans, and blood and cerebrospinal fluid biomarkers. These data are spread across many files, each containing related information for all subjects, with each line corresponding to one medical visit. Some files contain optional text fields in which the physicians or examiners could include notes or descriptions at their discretion. Section 4.3 describes how such text fields were used to construct the primary corpus used in this thesis. Each of the 1,783 subjects is assigned a diagnostic label upon entering the study. The original labeling scheme was modified in later phases of the study (ADNI-GO and ADNI-2), resulting in a total of six possible labels for each subject: Cognitively Normal (NL), Significant Memory Complaint (SMC), Early Mild Cognitive Impairment (EMCI), MCI, Late MCI (LMCI), and Alzheimer's Disease (AD). The SMC group is distinguished from MCI by self-reporting of their memory issues, as opposed to typical MCI sufferers whose problems are brought to attention by others. The original ADNI-1 study had only one MCI category,

with later phases introducing Early and Late MCI to increase the diagnostic granularity. However, some subjects from the original study did not return for the later ones and thus retained their original MCI label. Because of these changes to the label assignment procedure, the work presented in this thesis made use of a subset of the available ADNI subjects, which is explained in Section 4.2.1. A subject may also *convert* to another diagnostic label if their symptoms change during the course of the study. Of the nearly 1800 total subjects, only 19 converted, two of which appear to be corrections of the original diagnosis, i.e. the subject reverts back to a lower-level disease state, rather than progressing to a later one. These converted subjects have been studied in the past [Barnes et al., 2014], but are not examined in this thesis, given their rarity.

4.2.1 Inclusion and Exclusion of Subjects

As mentioned in the previous section, the diagnostic assignment procedures changed between phases of the ADNI collection, resulting in some subject having their labels updated to reflect the new rules, but other patients retaining the labels from the original study (as they did not return for the later ones). To deal with this, only subjects who joined the study under the most recent phase, *ADNI-2*, are included in this work. Of these, any subject with a label of *SMC* (Significant Memory Complaint) is excluded because of the ambiguity of the label (SMC is not a real diagnostic category outside of this study). Finally, a subject must have both unstructured text data and structured data in their record to be included. This restriction is necessary for the model integration experiments later on. This leaves 679 usable subjects. From this point on, this thesis will refer only to these subjects and their data.

4.2.2 Class Labels

The *ADNI-2* phase of the ADNI collection study used five labels to indicate the progression to Alzheimer's Disease: *NL* (Normal), *EMCI* (Early Mild Cognitive Impairment), *LMCI* (Late MCI), and *AD* (Alzheimer's Disease). Subjects labeled with *SMC* are excluded because it

is not a real diagnosis, but rather a special label used within the context of the ADNI study. The label (class) distribution of the remaining 679 subjects is shown in Figure 4.1. The distribution is relatively balanced between all classes.

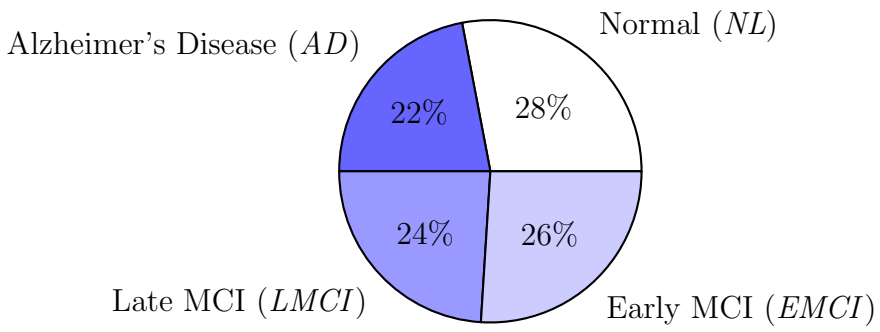


Figure 4.1: Distribution of diagnostic labels for the subjects used in this thesis ($n = 679$). MCI = Mild Cognitive Impairment.

4.3 Construction of a Text Corpus

As mentioned earlier, the ADNI dataset consists of many data files, most of which contain almost exclusively structured data. However, some files have an optional text field where physicians could add notes or descriptions at their discretion. Four files in particular were identified which contained considerable quantities of text data, as shown in Table 4.1. There is a different number of subjects present in each file because some subjects may not have text in all four files/categories for each visit. No single file contains entries for all subjects, and a subject without any text notes is excluded. However, all of the 679 subjects meeting the criteria established earlier in Section 4.2.1 possess text notes in at least one of these four files, and are thus usable for this thesis.

Each of the four files described above could be treated in isolation, considering all entries for each subject to be one document for that subject, as was shown in the previous section. Instead, entries from all of these files are aggregated by subject and concatenated to yield one text document per subject. Representing each subject as a single document is intuitive for a text mining study. This only requires a subject to have notes in at least one of the files, which was part of the inclusion criteria in the first place.

File	Content description	Total # entries	# usable entries	# unique subjects
RECMHIST	Recent medical history	30,727	7,153	678
RECADV	Recent adverse events/hospitalizations	16,063	1,375	384
RECBLLOG	Symptoms at initial/baseline visit	12,768	2,156	585
BLCHANGE	Changes since initial/baseline visit	8,571	1,955	635

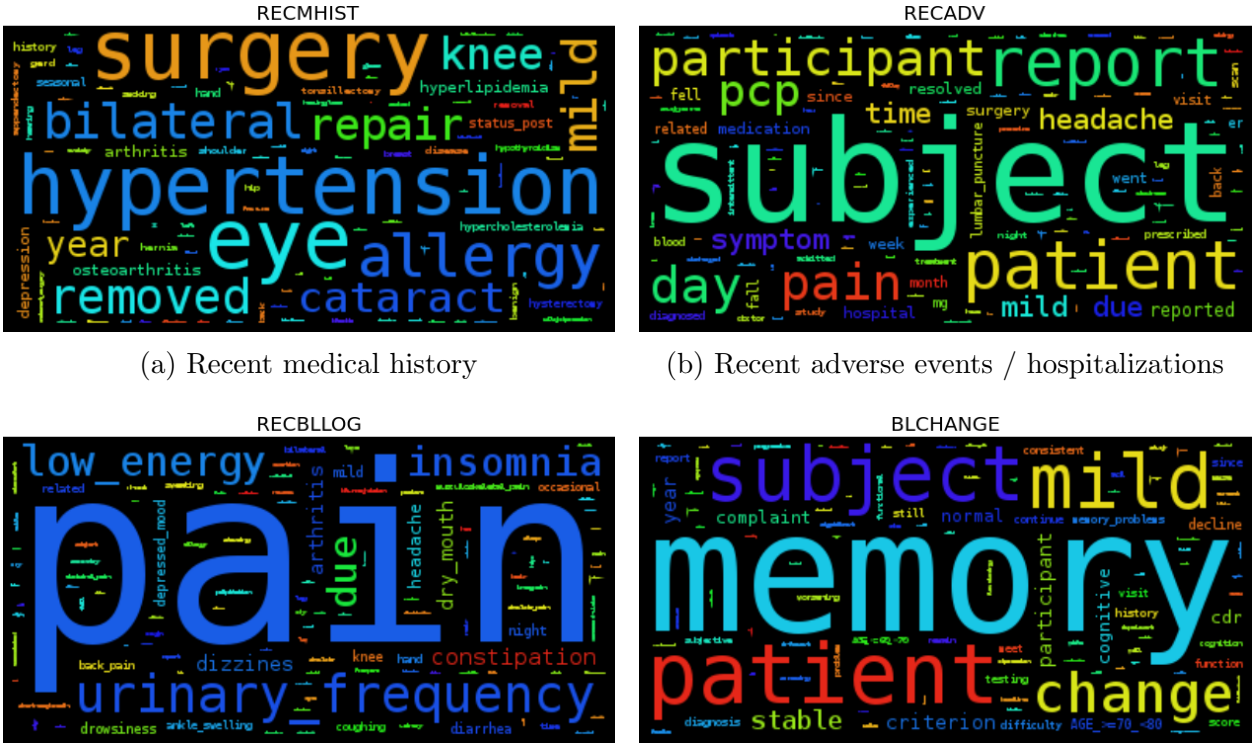
Table 4.1: Files chosen for corpus construction based on available text data. An *entry* refers to one medical visit. One subject may have multiple visits/entries. *Total # entries* refers to all entries, regardless of their inclusion in the corpus being constructed here. The last two columns identify the usable entries and subjects described in Section 4.2.1.

4.3.1 Exploration of Corpus Vocabulary

This section details exploration of the vocabulary content of the text corpus. Although text is considered unstructured, it could be argued that the categories of the files listed in the previous section (recent medical history, recent adverse events/hospitalizations, symptoms at initial/baseline visit, and changes since initial/baseline visit) enforce some form of structure by restricting the content. These data are still free in comparison to typical structured data, however. It is interesting to characterize and compare the four files in terms of their linguistic content.

One interesting way to view lexical differences in the files is through word clouds. A word cloud is a visualization in which the most common words in a corpus or text are arranged in a block with a font size corresponding to their relative frequency. Figure 4.2 shows word clouds of the most frequent 200 lexical word types in each of the four files (generated using the freely available `wordcloud` Python library, obtained from `https://github.com/amueller/word_cloud`). Note that for this analysis, the words were subjected to the preprocessing and normalization steps explained later in this thesis, in Section 6.1.1.

This high-level view of the lexical content of each text source shows obvious differences. Some of the large words in the recent medical history (Figure 4.2a) relate to visual/ocular (*eye*, *cataract*) and blood pressure (*hypertension*) problems, as well as surgery in general (*surgery*, *repair*, *removed*). This is not surprising given the older age range of the ADNI



(a) Recent medical history

(b) Recent adverse events / hospitalizations

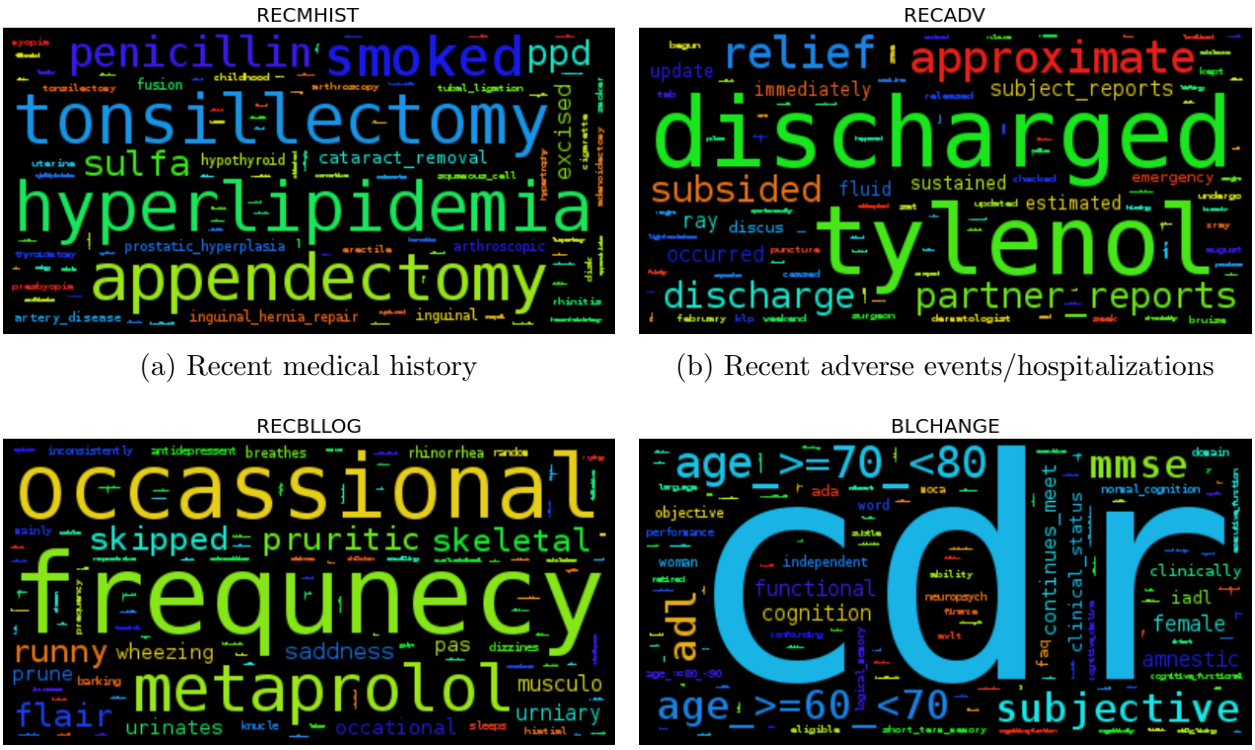
(c) Recent symptoms at initial/baseline visit

(d) Changes since initial/baseline visit

Figure 4.2: Word clouds of up to the top 200 word types (i.e. distinct lexical words) for each of the four ADNI files used. The size of a word corresponds to its frequency, where the biggest word is the most frequent in the documents of a given file. Color is only for visibility. Word clouds were generated using the freely available `wordcloud` Python library (`https://github.com/amueller/word_cloud`).

subjects. For recent adverse events and hospitalizations (Figure 4.2b), there appears to be a focus on the *subject*, *patient*, or *participant* (effectively synonyms here). Phrases such as *subject reports* are common in these documents. The word cloud for recent symptoms at the baseline visit (Figure 4.2c) is also not surprising, with *pain* appearing much larger than any other word in any of the four word clouds, followed by typical afflictions of the elderly, such as *urinary_frequency* and *low_energy* (see Section 6.1.1 for an explanation of why these contain underscores). Finally, the log of changes since the baseline visit (Figure 4.2d contains many instances of the word *memory* (as indicated by its large size), as well as similarly increased relative frequencies of *subject* or *patient* references when compared to the recent adverse events. It is important to notice that there are differences between the lexical content of

each of these text sources, showing the diversity of medical details contained within them. (This linguistic diversity would likely be more evident in regular electronic health records, as compared to this dataset, and would therefore likely make an even better basis for the work being presented here.)



(a) Recent medical history

(b) Recent adverse events/hospitalizations

(c) Recent symptoms at initial/baseline visit

(d) Changes since initial/baseline visit

Figure 4.3: Word clouds of top 200 word types (i.e. distinct lexical words) for each of the four ADNI files used, this time showing *only* the words which are unique to a given file (i.e. words that do not appear in any of the other three). The size of a word corresponds to its frequency, where the biggest word is the most frequent in the documents of a given file, excluding the words types found in other files. Color is only for visibility. Word clouds were generated using the `wordcloud` Python library (`https://github.com/amueller/word_cloud`).

Although there are some differences between the four text file sources, there were also some similarities in relative frequencies of words within each. It would be potential useful to look at the words that differ between them in a similar way. Figure 4.3 shows the same kind of word clouds as before, but this time only words unique to a given file's documents are included. For example, the word *subject* appears in all four files, and therefore is excluded from any word cloud, whereas the word *cdr* (Clinical Dementia Rating, a cognitive test) appears only in BLCHANGE and is thus included in the BLCHANGE word cloud. This version of the word clouds allows us to see the relative frequency of words that may be confined to a particular text file and reveals an advantage of each. That is to say, each file captures something slightly different, and thus using all of them grants us a bigger picture of subjects' health conditions.

A limitation of this word cloud analysis is that it only shows the *relative* frequencies in the documents of each file. This is acceptable in Figure 4.2 because the number of word tokens is high, but in Figure 4.3, the frequency of the unique words in each file may be quite low, possibly only a single occurrence. Yet, visualization is rather meant to be a creative way to explore and understand the data on a high level as a complement to other experimentation and development.

4.4 Source and Preparation of Structured Data

The collection of structured data used for comparison and integration in this thesis was prepared in the past by a colleague, Rohan Murde, active in the same research group as the thesis author as part of a collaborative effort towards a larger project. These methods are summarized in Bullard et al. [2015], although the set of subjects considered in that paper differs slightly from the set used here. The relevant preparation and modeling procedures for the structured data are explained in this section, rather than in the Methods chapter, which deals with those of the unstructured data and other work done solely by the thesis author.

There are a total of 22 structured data fields originally obtained from a subset of the ADNI database files, which are shown in Table 4.2. Nineteen of these fields are measurements obtained from either cerebrospinal fluid samples or brain imaging scans. The remaining three are scores from cognitive exams: the Clinical Dementia Rating (CDR), the Mini Mental State Exam (MMSE), and the Alzheimer's Disease Assessment Scale (ADAS13). These three features will be distinguished later in Chapter 6. The early work published in Bullard et al. [2015] made use of a different subset of the ADNI subjects, and the structured data here was matched to the new subject set used in this study. The previously mentioned problem of missing values in the structured data was handled through multiple imputation, a statistical process that uses log-likelihoods to generates probable complete datasets, averaging the values to get an estimate for the missing values. This had been accomplished using the Amelia II package in the `R` programming language. The outcome is that each subject has one value for each of the 22 structured data fields.

File	Content description
baimrinmrc	Brain volume atrophy
cdr	Clinical Dementia Rating (CDR) scores
upennbiomk5	Cerebrospinal fluid (CSF) biomarkers
upennbiomk6	Cerebrospinal fluid (CSF) biomarkers
upennplasma	Plasma biomarkers
ucberkeleyav45	PET scan with florbetapir

Table 4.2: ADNI files from which the structured data collection was obtained.

5 Problems, Goals, and Hypotheses

The previous chapters have introduced and motivated this thesis by briefly presenting its practical applications and technical contributions. These, along with the description of the dataset and corpus construction, provide the context necessary to formally outline the problems addressed by this thesis and to hypothesize the experimental outcomes. Detailed descriptions of the experimental methods are given later in Chapter 6. In this work, each subject from the ADNI dataset is considered an instance for the supervised machine learning task of correctly classifying the diagnostic label assigned to that subject in the ADNI study. Features of patients' structured and unstructured data are represented separately and treated as distinct sources of information. There are three primary goals as outlined below.

5.1 Incorporation of Unstructured Text Data

The first goal is to compare the performance of predictive modeling based on structured and unstructured data features separately. Processing of structured data from the ADNI have been performed in the past by a colleague, Rohan Murde, in this thesis author's research group, and is contrasted against models of the subjects' corresponding unstructured data. The hypothesis is that the unstructured data features alone will yield performance comparable to that of the structured data, especially when excluding the cognitive assessment scores. This result would show the practical utility of natural language data for disease prediction when relevant structured data are unavailable, as is often the case for conditions like dementia. Success of this goal is determined by classification performance metrics, explained in detail in Chapter 6.

5.2 Topic Modeling for Dimensionality Reduction

As part of the second goal, the topic modeling algorithm Latent Dirichlet Allocation (LDA) is explored as a form of textually interpretable dimensionality reduction for the bag-of-words feature space. The hypothesis is that LDA will improve classification performance with the unstructured data by reducing the sparse feature space to a dense representation, and furthermore, that this reduced topic space will provide a meaningful and interpretable characterization of subjects' text documents. The latter result is desirable for this human-centered medical application, where intuitive interpretation is favorable. The success of this goal is evaluated through computational metrics described in Chapter 6.

5.3 Integration of Structured and Unstructured Data

Finally, features and classification models based on structured and unstructured data are integrated for additional classification experiments. The hypothesis is that combining the power of these two feature types will improve performance over either in isolation. This is evaluated in the same ways as the previous classification tasks. This result would further strengthen the previously hypothesized utility of unstructured data features for disease prediction by showing that they not only provide benefit on their own, but also complement established work with structured data.

6 Methods

This chapter begins by explaining the modeling (Section 6.1) of unstructured text data, as well as the techniques for integration with the structured modeling, all of which produce features to be used in classification of the subjects' diagnostic class label. The classification experiments and evaluation procedures are then described in Section 6.2, followed by a section (6.3) describing the evaluation of LDA topics.

6.1 Data Modeling

There are three main feature representations for the unstructured data: *bag-of-words*, *tf-idf*, and *topic modeling* with Latent Dirichlet Allocation (LDA). A prerequisite to this modeling is text normalization, which is described first, below, followed by subsections dedicated to establishing the nuances and implementation decisions of each of the text modeling representations. The source and preparation of the structured data were already covered earlier in Section 4.4, but the experimental choices regarding its incorporation are described here. Finally, the integration techniques are also defined. Each of these modeling stages is meant to introduce new functionality and provide benchmarks of performance improvement in subsequent stages. The entire modeling process is visualized in Figure 6.1 below. As already stated, all processing, modeling, and experimental design using unstructured data, and the integration of structured and unstructured models, constitute the independent work of the author of this thesis. Modeling of the structured data was performed in the past by a colleague, Rohan Murde, within the same research group as the author of this thesis, the final results of which are used here for the integration experiments.

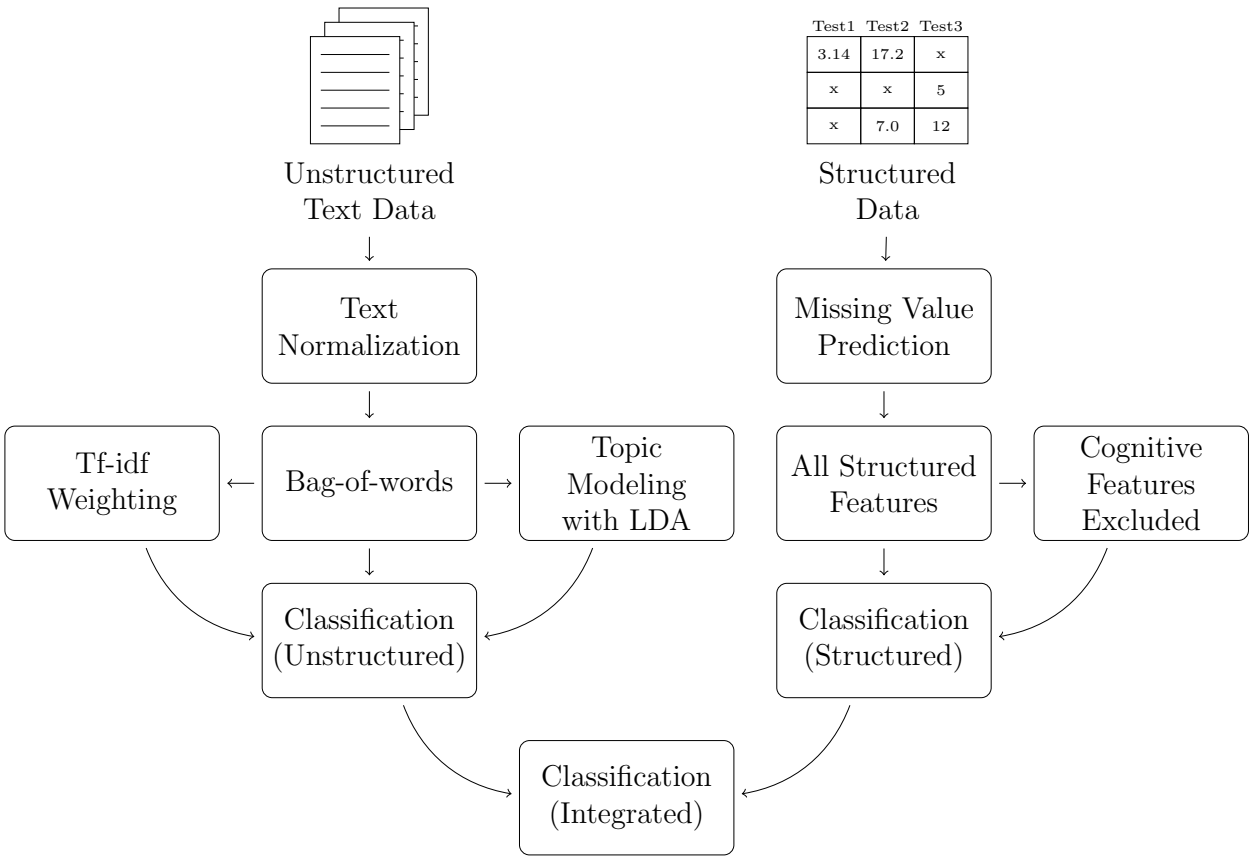


Figure 6.1: Box diagram illustrating the stages of modeling implementation. Arrows indicate where the output of one stage feeds into the next. The items along the bottom path (*Structured Data* → *Missing Value Prediction* → *Classification (Structured)*) are the work of a colleague, Rohan Murde, in the same research group, but all work pertaining to unstructured data and model integration constitute independent work for this thesis (see Section 6.1.6)

6.1.1 Text Preprocessing and Text Normalization

The first input in this model is text (i.e. natural language) data from electronic medical records. Natural language introduces a number of irregularities, ambiguities, or otherwise non-standard words that are accounted for before proceeding with further computational modeling [Sproat et al., 2001]. A central idea of text normalization is to identify and convert multiple forms of the same thing and into one common form for more complete and effective processing. For example, one calendar date can be written in many different formats, but statistical and linguistic models may not be able to reconcile them adequately. Converting

all dates into one format would allow the same date to be recognized in all contexts in which it appears. Text normalization is important in medicine due to the high frequency of domain-specific lexicon and shorthand (i.e. jargon). This section outlines the various stages of pre-processing and normalization implemented in this thesis. A full example of these procedures is given in Figure 6.2 on page 32. All preprocessing and text normalization procedures are performed in Python, with help from the Natural Language Toolkit (NLTK).

Preprocessing

Preprocessing differs from text normalization in that it is not concerned with unifying various forms of semantically identical linguistic units, but rather it deals mostly with converting text data into a usable format for normalization (hence *pre*processing). Standard preprocessing techniques implemented here include lowercasing, punctuation removal, and stop-listing (removal of common function words, pronouns, prepositions, e.g. *the*, *I*, *around*). The stop-list used consists of the English stop-words from the Natural Language Toolkit (NLTK) in Python. In addition to typical stop-listing, words or phrases revealing a subject's diagnostic state (e.g. *MCI*), as discovered during data exploration, are removed from all documents. This is done because the goal of this work is to identify features of and predict dementia based on information in a medical record *before* the diagnosis. Finally, words in a document are lemmatized using the Natural Language Toolkit (NLTK) WordNet lemmatizer to reduce inflections of the same word into one word token (e.g. *cataracts* and *cataract*). This helps reduce sparsity without altering the meaning of the text. The medical domain uses may abbreviations, some of which end with the letter *s*, confusing the lemmatizer. Unit testing of lemmatization produced a list of all altered words, which was inspected to identify errors. Any token in this list is ignored during lemmatization. The result of preprocessing can be seen in the full example in Figure 6.2 on page 32, along with the rest of the text normalization procedures explained below.

Date and Age Expressions

References to specific dates or ages are also quite common in medical texts, and can be represented in many different ways. Such expressions are converted into uniform representations that can be easily identified and extracted. For date expression, the tag format is *DATE_yyyy_mm_dd*, where *yyyy* is a four-digit year, *mm* is two-digit month with leading zeros, and *dd* is a two digit day with leading zero. For age expressions (e.g. *72 y/o*), the tag format is *AGE_yyy_mm*, where *yyy* is the number of years as three digits with leading zeros, and *mm* is the number of months as two digits with leading zeros. Age expressions with a number of months exceeding 12 are carried over into the years, and any unit of time smaller than a month (e.g. days, weeks, etc.) is ignored. Table 6.1 shows a full list of all date and age formats which are handled, with examples and explanations.

Date			**Age**		
2/3/2013	→	*DATE_2013_02_03*	*75 y/o*	→	*AGE_075_00*
12/31/78	→	*DATE_1978_12_31*	*75 yo*	→	*AGE_075_00*
feb/03/13	→	*DATE_2013_02_13*	*1 year 1-month old*	→	*AGE_001_01*
03/february/13	→	*DATE_2013_02_13*	*10-1/2 year-old*	→	*AGE_010_06*
- -/03/2013	→	*DATE_2013_00_31*	*16 half year-old*	→	*AGE_016_06*
02/- -/2013	→	*DATE_2013_02_00*	*5 month 3 day old*	→	*AGE_000_05*
- -/- -/2013	→	*DATE_2013_00_00*			
- -/- -/- - - -	→	*DATE_0000_00_00*			
7/1991	→	*DATE_1991_07_00*			
january 2 2001	→	*DATE_2001_01_02*			
2 january 2001	→	*DATE_2001_01_02*			
january 2	→	*DATE_0000_01_02*			
january 2001	→	*DATE_2001_01_00*			

Table 6.1: Examples of date and age expression tagging.

A problem with date expression is that they represent the continuous spectrum of time, and therefore it is entirely possible that none of them exist more than once in a given corpus. For this reason, they are simply removed here after tagging. The described tagging procedure could be useful in future work involving temporal modeling. Age expressions have a similar

issue, but rather than removing them completely, they are binned into decades starting at 40 and going up to 90. The resulting tag has the format *AGE_>=mm_<MM*, where *mm* and *MM* are the lower and upper bounds of the bin, respectively (e.g. *AGE_>=70_<80*). Ages below 40 are represented as *AGE_<40* and ages at or above 90 are represented as *AGE_>=90*. This grouping allows for certain ranges of ages to show up in the modeling, which may be important in the context of dementia.

Numbers and Numerals

Numbers and numeral representations can create difficulty in NLP and text mining because there are potentially infinite possibilities, or at the very least an infeasibly large, set of possibilities, depending on the context, but the numerical value itself may not actually be of much importance. Initially, any Roman numerals (e.g. *II*) and spelled numbers (e.g. *twenty*) were all converted to Arabic (e.g. *II* $\rightarrow$ *2*, *twenty-two* $\rightarrow$ *22*). However, it was decided in later experimentation that any number outside of temporal date or age expressions should be removed from the text, as the actual numerical value is of little importance to the type of modeling being done.

Abbreviations and Acronyms

Language used in specialized domains like medicine often contains abbreviations for single words (e.g. *patient* $\rightarrow$ *pt*) or acronyms for phrases and multi-word expressions (e.g. *status post* $\rightarrow$ *sp*). In the former case, a find-and-replace method is used, based on a word-list of common medical abbreviations. The latter case of multi-word expressions (MWEs) can present more difficulty, which is dealt with later in Section 6.1.1. For this normalization step, common medical acronyms are expanded into their constituent words (e.g. *bph* $\rightarrow$ *benign prostate hyperplasia*). The list of abbreviations was mostly built during the data exploration phase, and is based on this dataset. More examples are shown in Table 6.2.

Abbreviations			Acronyms		
dx	→	*diagnosis*	*ad*	→	*alzheimers disease*
ha	→	*headache*	*bp*	→	*blood pressure*
htn	→	*hypertension*	*bph*	→	*benign prostate hyperplasia*
pt	→	*patient*	*cdr*	→	*clinical dementia rating*
r/x	→	*prescription*	*dm*	→	*diabetes mellitus*
w/	→	*with*	*lp*	→	*lumbar puncture*

Table 6.2: Examples of expansion of abbreviations and acronyms. Full word-lists are available in the CD archive submitted with this document.

Multi-word Expressions

Multi-word expressions (MWEs) are sequences of words whose meaning as a whole is not simply the semantic composite of each individual word (e.g. *hot dog* is a semantic multi-word unit that is not simply a combination of *hot* and *dog*). In such cases, it is more reasonable to treat the fully-expanded sequence of words as one unit to preserve its meaning, which may be distorted by unigram-based linguistic models, such as bag-of-words and LDA, that do not account for word order. In the domain of medicine, there are many expressions that do not strictly meet the criteria for MWEs, but whose meaning may still be lost in such models. For example, *clinical dementia rating* is essentially a *rating* of *dementia* used in a *clinical* setting, but the order of the words is still very important; a medical record containing these three words in different positions would not be distinguished from one containing them in sequence. It is beneficial to relax the definition of MWE to include such cases. This is done by concatenating constituent word with underscores (_), so that it will be treated as one unit during modeling (e.g. *breast cancer* → *breast_cancer*). This is common practice in pre-processing for Latent Dirichlet Allocation (LDA) [Boyd-Graber et al., 2014, p. 9–10].

Many such MWEs can be identified during data exploration, but a more thorough list is generated through analysis of n-grams (word sequences of length n) in the corpus. The 200 most frequent bigrams and trigrams (sequences of 2 and 3 words, respectively) are extracted into a word-list, which is manually inspected for validity. This is done after all

previously described text normalization procedures, thus these are lexical content n-grams. This method appears to do well at finding common MWEs, but also produces a number of false positives for coincidentally common n-grams. For example, *insomnia depressed mood* is a common content trigram in this dataset, presumably because those symptoms are often listed together in patient symptoms or history, but the three words together do not constitute one semantic concept. Thus this trigram is removed from the list during inspection is not concatenated for modeling. In the case that a bigram is a subsequence of another trigram (e.g. *restless leg* and *leg syndrome* are actually part of *restless leg syndrome*), the bigrams are removed so that the trigram can be properly recognized. Finally, there are some cases in which an MWE may be synonymously represented by only a subset of its constituent words (another form of abbreviation), such as referring to *diabetes mellitus* as simply *diabetes*. For these kinds of MWEs, the words are not concatenated, as doing so would actually separate the two expressions, while leaving them as individual words would allow the meaning to be identified by the important word in both cases. The final list contains 112 expressions, a subset of which are shown in Table 6.3.

Multi-word Expressions from n-grams		
Concatenated	**Ignored**	
blurred_vision	*cdr scores*	$*$
breast_cancer	*diabetes mellitus*	$*$
chest_pain	*history depression*	$*$
clinical_dementia_rating	*activities daily living*	$\times$
cognitive_decline	*hypertension hyperlipidemia*	$\times$
daily_living	*insomnia depressed mood*	$\times$
high_blood_pressure	*male no*	$\times$
kidney_stone	*bowel syndrome*	$\circ$
memory_problems	*high blood*	$\circ$
short_term_memory	*leg syndrome*	$\circ$

Table 6.3: Examples of multi-word expressions (MWEs) in the corpus extracted from n-gram analysis (see Section 6.1.1). An identified n-gram is concatenated with underscores (_) unless it is synonymously expressed by one of its constituent words alone (indicated by $*$), it is a sub-sequence of a true MWE (indicated by $\circ$), or it is not truly an MWE (indicated by $\times$). Full word-lists are available in the data archive submitted with this document.

Full Process

Figure 6.2 shows a full example of all preprocessing and normalization procedures to one subject's text. The procedures explained and performed above appear to cover most of the potential issues in this dataset, and were unit-tested and inspected during implementation. However, text normalization is tricky, and it is expected that that some linguistic inconsistencies remain in the texts after normalization.

Mild depressive symptoms. Treated with SSRI sertraline Hypertension Hypercholesterolemia History of constipation Participant accidently fell on kitchen floor on 9/15/12. No injuries reported. Husband had scrubbed the floor so the floor was wet when SMT entered the room. 71 year old woman first evaluated for memory complaints in April 2008, noted to have mild deficits at that time. Has had progression gradually over time with increasing impairment in memory and executive function with no commensurate changes in general health. She presently has a clinical diagnosis of mild dementia, probable Alzheimer's type. 72 year old in excellent general health with notable memory changes over the last two years or so. Memory function and CDR consistent with AD. Participant has had no change in medical condition, no evidence for delirium, but she has had a subjstantial decline in memory and functional status over the last six months such that she has likely crossed the threshold to moderate dementia. She is displaying decline in cognition and functional skills that is consistent with AD and not attributable to other factors.

⇓

mild depressive symptom treated ssri sertraline hypertension hypercholesterolemia history constipation participant accidently fell kitchen floor no injury reported husband scrubbed floor floor wet smt entered room AGE_>=70_<80 woman first evaluated memory complaint noted mild deficit time progression gradually time increasing impairment memory executive_function no commensurate change general health presently clinical diagnosis mild probable type AGE_>=70_<80 excellent general health notable memory change last year memory function cdr consistent participant medical condition no_evidence delirium subjstantial memory_problems functional status last month likely crossed threshold moderate displaying decline cognition functional skill consistent attributable other factor

Figure 6.2: Full example of preprocessing and text normalization procedures. Note that this document is shown verbatim, including typographical errors in the dataset. Each document here is the concatenation of all text entries for one subject.

6.1.2 Bag-of-Words

In a bag-of-words model, a *dictionary* is created by assigning an index to every distinct term in a corpus. Each document is then represented as a list of these indexes, along with their frequency in the document. Figure 6.3 shows a toy example using a corpus of two documents to clearly illustrate this concept. The dictionary constructed on a corpus can also be used to represent documents outside of the corpus. It is possible that a new document may contain an out-of-vocabulary term (i.e. one that did not appear in the corpus used to construct the dictionary). One way to handle this is by ignoring such terms, as is seen in the example in Figure 6.3. Ideally, the size and vocabulary of a training corpus would capture enough that this method would not present a problem, as is expected with a large and broad medical dataset. This limitation is acceptable, considering that bag-of-words is overall a simple model, not taking into account relationships between words. Bag-of-words models are implemented using the `gensim` Python library [Řehůřek and Sojka, 2010].

Dictionary

`0 :` *complain*
`1 :` *headache*
`2 :` *history*
`3 :` *improve*
`4 :` *recent*
`5 :` *status*
`6 :` *subject*

Doc 1 *history headache subject complain recent headache*
`[(2,1),(1,2),(6,1),(0,1),(4,1)]`

Doc 2 *subject status improve*
`[(6,1),(5,1),(3,1)]`

Doc 3 *history hypertension*
`[(2,1)]`

Figure 6.3: Example of a bag-of-words model on two documents. This is meant to illustrate the concept, which is done more clearly with this toy example than with a full example from the dataset (which contains thousands of word types). These sample documents resemble the final product of all preprocessing and text normalization techniques described earlier. The `Dictionary` is an index of the unique word types in the corpus of `Doc 1` and `Doc 2`. Below each document is its bag-of-words representation - a list of indexes and counts. For example, the tuple `(1,2)` in `Doc 1` indicates that word `1` (*headache*) appears 2 times. `Doc 3` was not used to construct the dictionary and contains an out-of-vocabulary word (*hypertension*), which is ignored in its bag-of-words representation using this dictionary.

Classification using Bag-of-Words

The standard bag-of-words representation is already suitable for use as a feature vector, since each term can be treated as a feature dimension with a value equal to the term frequency within a given document. Such features were shown to be promising with this dataset in preliminary work on this project, as seen in Bullard et al. [2015]. This representation is very sparse, since any document will only contain a small subset of the corpus-wide vocabulary, resulting in many zero-valued features. Sparse storage formats are available and usable for classification, however, dimensionality reduction is common practice to help improve classification performance when operating in a sparse feature space. This is explored here through topic modeling, explained below in Section 6.1.4.

6.1.3 Tf-idf

An extension of the standard bag-of-words representation is to weight the terms based on their distribution in the corpus using *tf-idf*, or *term-frequency inverse-document-frequency*. The idea of tf-idf is that words which appear many times in fewer documents may be more meaningful than words which appear across many documents. Using tf-idf, a term w_i in document d_j is weighted by

$$\texttt{tfidf}(w_i, d_j) = \texttt{term\_freq}(w_i, d_j) \times \log_2 \left(\frac{D}{\texttt{doc\_freq}(w_i)} \right) \tag{6.1}$$

where $\texttt{term\_freq}(w_i,\ d_i)$ is the frequency of w_i in d_j, $\texttt{doc\_freq}(w_i)$ is the number of documents containing w_i, and D is the number of documents. 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 is identical to that of standard bag-of-words, but the values for each feature are equal to the weights, as defined above. Tf-idf is often used to achieve performance improvements over the standard bag-of-words representation, as is the goal of its application here.

6.1.4 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) [Blei et al., 2003] is a generative model for identifying topics of related terms in a text corpus[1]. Under LDA, a corpus D consisting of M documents is assumed to contain a fixed number of topics K. The value of K is a parameter for the model. The initial proportions of the topics in the corpus are assumed to be drawn from a Dirichlet distribution (hence the name of the algorithm). A Dirichlet distribution is parametrized by a vector α of real numbers, and its probability density function returns a multinomial distribution. For a simple illustration of this concept, consider a typical six-sided die - essentially a multinomial distribution over six possible outcomes. If there was a bag full of dice, each weighted differently, and one was pulled out at random, this would be sampling from a Dirichlet distribution (i.e. the bag is the Dirichlet distribution that yields a die, which is multinomial). Dirichlet distributions are a common choice for priors in Bayesian statistical models such as LDA. In the case of LDA, each *topic* is a multinomial distribution over the vocabulary of the corpus, drawn from a Dirichlet distribution, denoted $\phi_k \sim \text{Dir}(\beta)$. Similarly, each *document* is a multinomial distribution over the set of topics in the corpus, also assumed to have a Dirichlet prior, denoted $\theta_i \sim \text{Dir}(\alpha)$. This process is written formally in Figure 6.4 and a visualization in plate notation is shown in Figure 6.5. 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. This can be written formally as:

$$P(w_j \mid d_i; \theta, \phi) = \sum_{k=1}^{K} P(w_j \mid z_k; \phi_k) P(z_k \mid d_i; \theta_i) \tag{6.2}$$

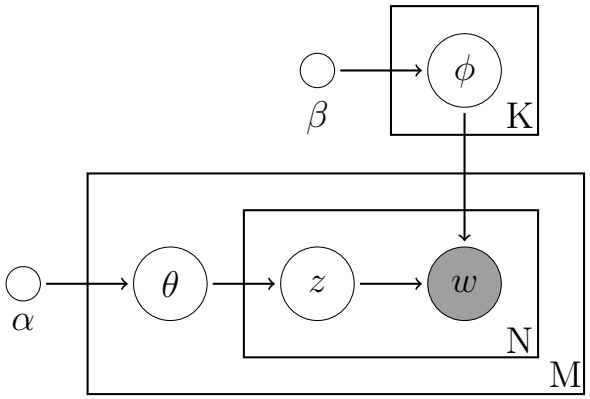
[1]LDA has been applied to image data as well [Wang et al., 2009]; the description here considers text.

Let D be a corpus of M documents with N_i terms

Let K be the pre-determined number of topics

1. Draw $\theta_i \sim \text{Dir}(\alpha)$, $i \in \{1, \ldots, M\}$
2. Draw $\phi_k \sim \text{Dir}(\beta)$, where $k \in \{1, \ldots, K\}$
3. For $i \in \{1, \ldots, M\}$
   (a) For $j \in \{1, \ldots, N_i\}$
      i. Draw topic $z_{ij} \sim \text{Mult}(\theta_i)$
      ii. Draw word $w_{ij} \sim \text{Mult}(\phi_{z_{ij}})$

Figure 6.4: Generative process of Latent Dirichlet Allocation (LDA).



α, β	parameters for Dirichlet priors
θ	topic distribution for current document
ϕ	word distribution for current topic
z	topic of current word
K	number of topics
M	number of documents
N	number of words in current document

Figure 6.5: Plate notation for Latent Dirichlet Allocation (LDA). Arrows indicate a generative process and rectangles represent a repetition of the contained nodes. Each document is a distribution (θ) over K topics, and each topic is a distribution (ϕ) over the vocabulary of the corpus. The probability of a word (w) is based on the current topic (z) in the current document.

Computing the actual distributions is intractable and is approximated through Bayesian inference methods. Blei et al. [2003] used variational Bayes approximation in the original paper, but Gibb's sampling is also commonly used. This thesis performs LDA using the Stanford Topic Modeling Toolkit (TMT) with collapsed variational Bayes (CVB) [Teh et al., 2007]. The Stanford TMT can also implement Gibbs sampling, but CVB converged on more sensible topics and performed better in classification during development.

Additional Preprocessing

Since topics are determined based on statistical relationships of words, extremely infrequent terms are unlikely to correspond to anything at all. Similarly, terms that appear in too many documents will end up being related to too many other terms. The effectiveness of modeling can be hampered in either case. For these reasons, it is common practice to filter the vocabulary [Boyd-Graber et al., 2014, p. 9]. This is addressed here by filtering out terms appearing fewer than 3 times, as well as filtering out the 30 most common terms. Other values were explored initially, but these appeared to suit the data.

Dimensionality Reduction and Classification using LDA

Although LDA is typically employed for unsupervised exploration of large corpora, it can also be thought of as a form of dimensionality reduction for bag-of-words. As mentioned earlier, bag-of-words is a sparse representation, with thousands of word features, the majority of which are absent for any given document. In LDA, each document is defined as a probability distribution over K topics, with K typically being relatively small. Each topic can be considered a feature whose value is equal to the probability of that topic within a given document, thus the feature space is K-dimensional. This reduction in dimensionality may be beneficial for classification, provided that the trained LDA model adequately represents the linguistic relationships in the data. A commonly used algorithm specifically for dimensionality reduction is Principal Component Analysis (PCA), which finds linear combinations of the feature dimensions that best explain the variance in the data. An issue is that the latent variables produced by PCA are not easily interpretable by humans, which is one of the key considerations of this thesis. One advantage of LDA as a form of dimensionality reduction is that the resulting output is a collection of topics which can be recognized or understood through human intuition, in addition to potentially improving the performance over simpler text features.

6.1.5 Structured Features

The source and preparation of the structured data were described already in Section 4.4. As was briefly mentioned in that section, a potentially meaningful distinction can be made between structured data which comes from cognitive assessments and those which come from other biophysical tests or markers. Namely, cognitive assessments are verbally administered by a clinical professional, and thus include another person's mind and expertise in order to reach those structured data values present in the utilized dataset. This differs from other major sources of structured data from the ADNI, which consist of cerebrospinal fluid markers and brain volume measurements, all of which are measurements to be interpreted by a physician after their collection. Essentially, the cognitive assessment scores in the dataset are the outputs of professional interpretation, whereas the other structured data are inputs for future interpretation. This distinction was initially explored in collaborative work [Bullard et al., 2015] and is replicated here by experimenting with the inclusion and exclusion of the three cognitive assessment score features.

6.1.6 Integration with Structured Data Models

The resulting feature vectors and trained models of the structured data analysis described above are used in conjunction with the those of the unstructured for the integration experiments. Integration is performed on each unstructured modeling experiment (bag-of-words, tf-idf, and LDA) and each structured (with and without cognitive assessment features). In the case of LDA, only the models/parameters with the highest performance are used in integration. Two integration techniques are described below: one which focuses on integrating at the feature level, and one which focuses on integrating the outputs of the two different models.

Vector Concatenation

The most intuitive way of integrating the features is to simply concatenate the feature vectors for structured and unstructured data. The term *concatenation* here refers to treating two vectors of length n and m as lists, and joining them to form a new vector/list of length $n+m$. This concatenated feature factor can be used in classification to determine new class outputs for each subject.

Posterior Probability Composition

A more sophisticated method of integration is to take advantage of posterior probabilities from the individual classification models. For each input, a logistic regression classifier produces a posterior probability of each class label (i.e. a distribution over the class labels), selecting the most probable as its output. One classifier is trained on structured data features X_s, and a second on unstructured data features X_u, resulting in two posterior distributions. The probability of a particular 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 class label, then Bayes' theorem can be leveraged as follows:

$$\begin{aligned} p(C_k \mid X_s, X_u) &\propto p(X_s, X_u \mid C_k)\, p(C_k) \\ &\propto p(X_s \mid C_k)\, p(X_u \mid C_k)\, p(C_k) \\ &\propto \frac{p(C_k \mid X_s)\, p(C_k \mid X_u)}{p(C_k)} \end{aligned} \tag{6.3}$$

From here, the class label with the highest probability is selected as the output. This methodology is explained in Bailer-Jones and Smith [2011], and it was implemented with bag-of-words modeling on a different subset of the ADNI data in Bullard et al. [2015].

6.2 Classification Experiments

Each subject in the dataset is annotated with one of four class labels indicating their dementia status - Normal (*NL*), Early MCI (*EMCI*), Late MCI (*LMCI*), or Alzheimer's disease (*AD*) - of its corresponding subject. Each of these subjects has unstructured and structured data, which are used separately, and later integrated, as instances for classification where the goal is to assign the correct class label. Two different classification problems are reported on: one using the standard labeling scheme of the dataset, and another designed to address an alternative interpretation of the original problem. These problems are explained below and their results are given in separate tables later in Section 7.

6.2.1 Labeling Schemes

Standard ADNI classes The first problem uses the four class labels as they appear in the ADNI study: *NL*, *EMCI*, *LMCI*, and *AD*. The distinction between early and late mild cognitive impairment (*EMCI* and *LMCI*) may be imprecise, and the resulting class confusion will hurt classification performance, but they were explicitly assigned by the clinical professionals in the study and therefore it makes sense to leave them as-is, rather than combine them. This 4-class problem is henceforth referred to as *Standard*.

Early Risk As discussed in Chapter 1, early detection of dementia is critical. It follows that a group of particular interest would be the early mild cognitive impairment (*EMCI*) subjects, as they represent the beginning of the disease progression. It would be useful to be able to distinguish those two groups in particular. There are 367 subjects having one of these two class labels (187 *NL*, 180 *EMCI*), and only this subpopulation can be used for this experiment. While this does not perfectly match the reality of diagnosis because it excludes the later stages, it could be argued that those later stages are in less need of automatic analysis since they are more easily observable than the earlier ones. This binary problem is henceforth referred to as *Early Risk*.

6.2.2 Logistic Regression Classifier

The primary integration method described in Section 6.1.6 depends on posterior probabilities being directly computed by the classification models. For this reason, the popular logistic regression was chosen as the algorithm for all classification experiments performed for this thesis. Logistic regression is a linear classifier, meaning that its decision function is an equation of one variable for each feature with a coefficient. The magnitude of the coefficient corresponds to the feature's relative importance to the decision, and the sign indicates which class is favored by a larger value for that feature. Logistic regression is a binary classifier, but is extended to multi-class problems through a *1 vs. all* approach, in which each class label is tested against the collection of all other class instances (e.g. *NL* vs. *not-NL*, *EMCI* vs. *not-EMCI*, etc.). This is is implemented in the `scikit-learn` library used here.

Parameter Tuning

Parameters of a classification algorithm can have a great deal of impact on its performance. For the logistic regression, the two parameters of interest are C, the inverse of regularization strength[2], and the penalty function, either the L^1 or L^2 vector norm. A smaller C correspond 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 on the bag-of-words *dev* data only, as the unstructured features are the main focus of the thesis.[3] The process is, however, repeated for each labeling scheme. Table 6.4 summarizes the selected parameters.

Labeling Scheme	C	Penalty
Standard	1.0	L^2
Early Risk	10.0	L^1

Table 6.4: Logistic regression parameters selected through cross-validated grid search on the training data. The C parameter was tested at powers of 10 from -5 to $+3$.

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

[3]Repeating this process on every feature input type could improve performance, but is not done here in the attempt to keep this experimental condition stable.

6.2.3 Evaluation

This section provides an explanation of the underlying implementation decisions for classification experiments, and how those decisions address potential experimental issues. There are two main evaluation procedures for classification performance, explained below.

Held-out Data

The dataset is randomly split into 80% ($n = 544$ documents) for model development (*dev* set), and 20% ($n = 135$ documents) for final evaluation (*held-out* set). Models are only exposed to the *held-out* set after satisfactory performance is achieved using only the *dev* set. This is standard practice in supervised machine learning. The roughly even class distribution in the corpus and randomness of the data split produces approximately class distributions in the *dev* and *held-out* sets. Results from final *held-out* testing of each modeling stage are explained in detail in the corresponding subsections of Section 7.

Leave-one-out Cross-validation (LOOCV)

All development of classification models on the *dev* set uses leave-one-out cross-validation (LOO or LOOCV), a variation of k-fold cross-validation. In k-fold cross-validation, the data is split into k segments, each containing $1/k$ data points, so that one segment can be used for testing a model trained on the other $k - 1$ segments. This process is repeated k times (folds) such that each segment has been used for testing once, and no segment is used for both training and testing within the same fold. Leave-one-out cross-validation is a special case where k is equal to the number of training instances, n, resulting in n folds each testing on exactly one data point. This version of cross-validation is used because it minimizes bias, and it additionally provides clearly interpretable performance metrics with one confusion matrix generated from the whole process. This is implemented using the `scikit-learn` machine learning library [Pedregosa et al., 2011] in Python.

Although the *dev* and *held-out* sets have similar class distributions, overfitting is still a potential issue. For this reason, after the previously described held-out evaluation is complete, a LOOCV procedure is run on the entire dataset to serve as an additional evaluation.

6.3 Topic Exploration and Evaluation

The most influential parameter in LDA is the number of topics. Tuning of this parameter is essential to finding an appropriate model. 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. The interpretability of the best LDA models in classification are examined with various per-topic metrics known to correlate well with human evaluations (see below). Tuning of the topic number K is performed by iteratively measuring classification accuracy at values of K ranging from 5 to 100, at multiples of 5. The best-performing reduced topic-feature space is selected for classification results and additionally analyzed using the following metrics:

Topic Coherence Topic coherence, defined by Mimno et al. [2011], measures how often the most probable words of a topic appear together in documents. The coherence of topic t is defined as $C(t, V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)})+1}{D(v_l^{(t)})}$, where $V^{(t)}$ is the top M words from topic t, $D(v_l^{(t)})$ is the document frequency of word $v_l^{(t)}$, and $D(v_m^{(t)}, v_l^{(t)})$ is the co-document frequency of words $v_m^{(t)}$ and $v_l^{(t)}$ (the number of documents in which both terms occur). This metric captures how often the most probable words of a topic co-occur in the corpus, and was shown to match well with human evaluation of topic quality in the original paper.

Topic Size The size of a topic t is equal to the number of word tokens in the corpus which have been assigned topic t. Larger topic size typically correlates with human-perceived topic quality [Boyd-Graber et al., 2014].

These metrics are useful because some topics in a given LDA model will be meaningless or nonsensical, as was described earlier in Section 2.2.2. The metrics above are employed to filter out those topics, with topic coherence being of particular interest. Other topic quality metrics exist, but the two above appeared more interesting for this application and thus were included for evaluative topic exploration.

7 Results and Discussion

This chapter presents experimental results for the labeling schemes outlined in Section 6.2.1, each of which has its own section discussing performance and tuning of each feature representation and data modeling technique. The previous chapter described the reasoning, implementation, and design of all modeling and experimentation, and this chapter therefore centers on the presentation and discussion of results.

7.1 Classification of *Standard* Labels

Classification results for the *Standard* labeling scheme (4 classes) are shown in Table 7.2 on page 52. The upper parts of the table shows the results with structured vs. unstructured data features in isolation, while the rest of the table shows results of integration techniques. Keep in mind that the *held-out* evaluation was performed on data which were not used in any development, while the leave-one-out cross-validation (LOOCV) makes use of the entire merged *dev* and *held-out* sets to either confirm or call into question the trends seen in held-out testing. It can be noted in many places in the table that the 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 method. Importantly, the relative performance differences between each modeling stage are comparable.

7.1.1 Performance of Structured vs. Unstructured Features

The performance of the structured data alone was substantially higher than the majority class baseline, more so when cognitive assessment score features were included (+*cognitive*

in Table 7.2). The performance of the bag-of-words representation for unstructured data approached that of the structured data sans cognitive assessment scores, falling short by a few percentage points, but was not as close when cognitive scores were included in the structured data modeling. Importantly, bag-of-words features still considerably improved upon the baseline, showing that even simple modeling of unstructured text data can be useful in its own right in the common event that structured data are missing. These observations are expected, given the early work presented in Bullard et al. [2015] on a similar subset of the ADNI data, and the fact that structured data have been regularly used for similar tasks in the past, as discussed earlier in Chapter 3.

The tf-idf representation improved on bag-of-words in the held-out evaluation, as well as matching the performance of structured data with cognitive features (and exceeding them without), but becoming seemingly worse in the LOOCV evaluation. One possible explanation for the lack of stability of tf-idf in these two cases is that, given the influence of document frequency, some important terms may have been quite different after merging the *dev* and *held-out* sets. This issue would not be observed in regular bag-of-words. Also, there are a number of variations of tf-idf weighting, which may affect its utility in text classification.

In the case of latent Dirichlet allocation (LDA), the number of topics greatly influenced the performance of classification, as can be seen in Figure 7.1, which shows the change in classification accuracy on the *held-out* set at multiples of five topics from 5 to 100. The figure indicates the performance of bag-of-words and tf-idf as lines for comparison. Two close peaks are noted at $K = 60$ and $K = 85$, both of which outperform bag-of-words, but neither of which matches tf-idf. There are various reasons why LDA was bested by tf-idf, namely that the reduction in dimensionality may also be a reduction in information, which is likely the case for the values of K which are below even the regular bag-of-words line in the figure. This does demonstrate, however, that the dimensionality reduction of bag-of-words provided by LDA can improve performance. It may be more useful if dealing with very large datasets. The full classification results for these LDA features from these two models

are given in Table 7.2. Comparing the held-out and LOOCV results in Table 7.2, it can be seen that this choice of topic number is likely tied to the *dev* set in particular, as their classification accuracy drops by over 10% in LOOCV. This is a limitation of using LDA, an unsupervised algorithm, for a supervised task; the class labels are introduced after the topic model is built, and thus any supervised metric based on it will depend heavily on the training data.

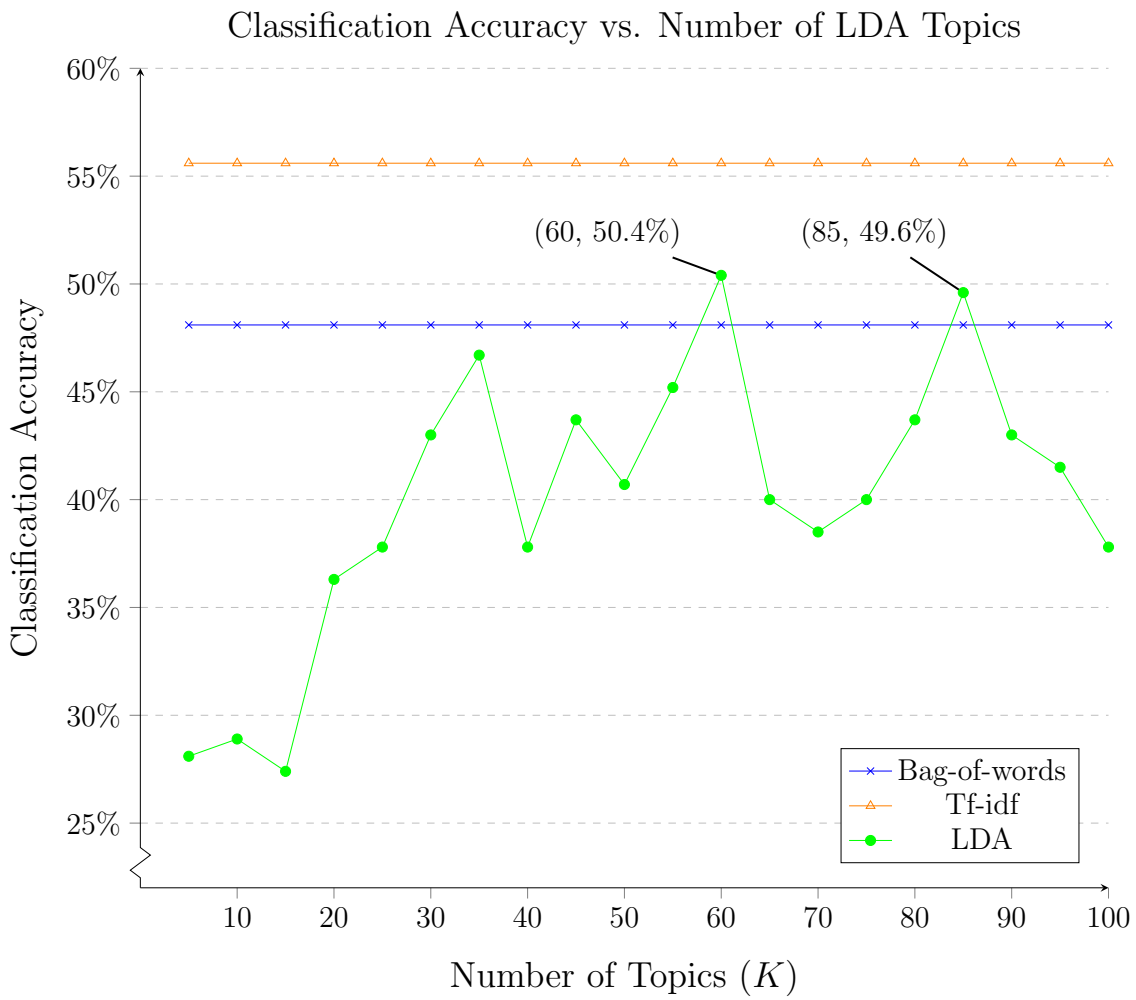


Figure 7.1: Accuracy of LDA-reduced features on the *held-out* set as the number of topics K increases (by multiples of 5). Lines indicating the performance of bag-of-words and tf-idf are shown for comparison and obviously do not change with the topic number of LDA. The two local maxima for LDA which surpass the performance of bag-of-words are labeled ($K = 60$ and $K = 85$).

The result above was to be expected, but recall that the argument for the use of LDA in this thesis is that the latent topics should additionally be interpretable. This can be evaluated through various metrics from the literature described earlier in Section 6.3, but it was noted during experimentation that the *topic coherence* and *topic size* metrics appeared to be the most useful. Topics from the 60-topic LDA model were ranked separately by each of these two metrics, and the top ten topics based on each are presented in Tables 7.1a and 7.1b on page 49. The coherence metric appears to live up to its name. For example, topics 2 and 33 are both about cognitive assessments and cognitive problems, respectively. The former also includes an age reference to people in their 60's, a time that is often associated with dementia. Topics 45 and 16 both pertain to regular medical visits (*PCP* is *primary care physician*), with the former seeming to be about typical problems or concerns associated with age (*back*, *heart*, *dizziness*), while the latter contains references to medications (*taking*, OTC = over the counter, *medication*). Another good example is topic 25, which is clearly about heart attacks (*cardiac*, *stent*, *chest_pain*, AE = adverse event) and their accompanying emergency visits (*hospitalization*, *admitted*, *discharged*). Some of the most coherent topics also rank in the top ten based on topic size (Table 7.1b), namely topics 2, 33, 16, and 25. This is interesting, as it indicates that the topics contain frequently co-occurring words that together account for a larger fraction of the overall word tokens in the corpus.

Topic ID	Coherence	Top Words
3	-72.448	*corroborated, subjective, continues_meet, score, factor, other, SP, AGE_>=60_<70, controlled_medication, unremarkable*
2	-79.621	*impression, CDR, MMSE, ADLS, AGE_>=60_<70, cog, amnestic, global, function, score*
20	-90.202	*effect, study_partner, side, PCP, approximately, cause, caused, resolved, diarrhea, unknown*
17	-90.496	*medical, consistent, status, function, continues, health, occasional, active, daily, functional*
45	-91.259	*blood, pressure, month, visit, PCP, diagnosed, dizziness, back, doctor, heart*
50	-92.003	*time, onset, approximate, ADNI, pleasant, exam, physical, estimated, woman, week*
29	-92.394	*family, month, mg, increased, Dr, per, wife, possible, however, reported*
25	-95.864	*hospital, admitted, discharged, stent, cardiac, went, chest_pain, AE, anxiety, total*
16	-95.967	*taking, low_energy, PCP, may, beginning, OTC, feeling, medication, lot, take*
33	-97.504	*decline, recall, informant, testing, subjective, logical_memory, delayed, MMSE, functional, AVLT*

(a) Top 10 topics based on *topic coherence*

Topic ID	Size	Top Words
52	921	*completed, visit, testing, study_partner, reported, per, mg, MRI, added, performed*
58	880	*diagnosis, remains, intermittent, probable, presbyopia, well, excised, chronic, episode, resolved*
42	866	*today, feel, screening, x, well, PCP, worse, cognition, month, ED*
17	840	*medical, consistent, status, function, continues, health, occasional, active, daily, functional*
3	824	*corroborated, subjective, continues_meet, score, factor, other, SP, AGE_>=60_<70, controlled_medication, unremarkable*
25	819	*hospital, admitted, discharged, stent, cardiac, went, chest_pain, AE, anxiety, total*
2	800	*impression, CDR, MMSE, ADLS, AGE_>=60_<70, cog, amnestic, global, function, score*
33	793	*decline, recall, informant, testing, subjective, logical_memory, delayed, MMSE, functional, AVLT*
20	785	*effect, study_partner, side, PCP, approximately, cause, caused, resolved, diarrhea, unknown*
16	761	*taking, low_energy, PCP, may, beginning, OTC, feeling, medication, lot, take*

(b) Top 10 topics based on *topic size*, or the number of word tokens assigned to a given topic.

Table 7.1: Top 10 topics from the 60 topic model based on two metrics, showing the 10 most probable words from each topic. See Section 6.3 for definitions of these metrics.

7.1.2 Performance of Integration

The goal of integrating the unstructured features and models with those of the structured ones is to improve classification performance over either in isolation. The results of two integration methods, vector concatenation and posterior probability composition, are presented using combinations of each of the four unstructured models identified and described in the previous section, along with the structured features, with and without cognitive assessment scores included, yielding 16 integrated models (See Table 7.2) For clarity, these results are discussed in the same order of unstructured feature groups as in the previous section: bag-of-words, tf-idf, then LDA.

In held-out testing, integration with bag-of-words features improved performance over both the bag-of-words and corresponding structured features in all cases. Interestingly, integrating bag-of-words with the cognitive assessment scores excluded actually outperformed the structured features when they are included, further strengthening the argument in favor of unstructured text modeling. In LOOCV, however, this did not hold for the cases where cognitive features were included, with the performance falling slightly below that of the structured in isolation. There was a similar pattern for tf-idf integration, with all but two cases successfully outperforming its constituent feature types: held-out with cognitive features excluded for both integration methods. Examples such as this are likely attributable to random divisions of the *dev* and *held-out* sets.

The LDA-reduced features are again less consistent than the other unstructured features, just as discussed in the previous section. In five out of the eight cases for held-out testing and six out of eight cases in LOOCV, integration with LDA produced performance gains. The most noticeable exceptions are with posterior probability composition without cognitive features in the *held-out* evaluation, for which the accuracy dropped by around 5%. This is not seen in cross-validation for these experiments, however, so it may too be a product of initial set divisions. In general, the LDA integration experiments seem to be somewhat

more robust between held-out and cross-validation than they were when LDA features were used alone. A possible explanation for this is that the structured features may be taking on the brunt of the classification work (i.e. being weighted more heavily in the case of vector concatenation, or consistently assigning high enough probabilities to the correct class labels to overpower the LDA models), and thus increasing the stability of those experiments.

It was predicted that the more sophisticated posterior probability composition method would yield better results than vector concatenation. The outcome appears to be less consistent, with many cases being the opposite of that prediction, namely when excluding cognitive assessment features (except for some in LOOCV). Yet overall, the best performing cases include results where integration is done by this method, cognitive features included. One potential limitation of posterior probability composition is that a stronger decision is made when all of the underlying classifiers produce an asymmetric posterior class distribution. Models which do not make strong or accurate decisions themselves may hurt the performance in integration using this method. Vector concatenation is not subject to this limitation, although it has the drawback of potentially overwhelming a smaller dense feature set with a larger sparse one.

7.1.3 Class-specific Performance

It is also interesting to examine the performance on a per-class basis using the precision and recall values presented in Table 7.2. In nearly all integration experiments, the *NL* (normal) and *AD* (Alzheimer's disease) patients had higher precision and recall scores than the two MCI classes. This is not surprising, given the stark contrast between patients on the two opposite ends of the disease spectrum (*NL* and *AD*), and the relative lack of diagnostic ambiguity when compared to *EMCI* and *LMCI*. This pattern is less pronounced in some of the isolated unstructured feature experiments.

	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	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 7.2: Held-out set results for classification of *Standard* labeling scheme (4 classes) using each modeling technique and feature group. Structured features with and without cognitive features are referenced in the lower sections of the table using the shorthand S_{+cog} and S_{-cog}, respectively. Integration by *vector concatenation* is indicated by ∪, and *posterior probability composition* by ⊕.

7.2 Classification of *Early Risk*

In addition to the experiments with the original labeling scheme, the sub-problem of distinguishing the normal (*NL*) subjects from the early mild cognitive impairment subjects (*EMCI*) was also explored, as this represents the earliest point in the progression to Alzheimer's disease (*AD*). The same classification approach described in the previous section is also used here, except that only leave-one-out cross-validation (LOOCV) is performed because the subsampling of *NL* and *EMCI* subjects slightly distorts the class distributions in the original held-out set. LOOCV may inflate the performance metrics slightly, but the relative performance between feature and model types is ultimately of more interest for this analysis. The results for the *Early Risk* classification are given in Table 7.4 on page 57 and discussed below.

7.2.1 Performance of Structured vs. Unstructured Features

As with the *Standard* labeling scheme, all structured and unstructured feature types are well above the majority class baseline (51% in this case). Similarly, including cognitive features produced performance gains. One major difference here is that all unstructured data types outperformed the structured features when cognitive assessments were excluded, as opposed to only the bag-of-words in the *Standard* problem. This is interesting because it suggests a potential linguistic differences in clinical notes at the onset of mild cognitive impairment.

The number of LDA topics was selected the same way as before (except using the whole *Early Risk* subsample, as opposed to the *Standard dev* set), as seen in Figure 7.2. The figure shows lines for the structured feature models as well, as some of the metrics are more comparable here than they were in the previous experiment. The two peaks at $K = 65$ and $K = 100$, each achieving the same classification accuracy, managed to almost match the tf-idf features. However, bag-of-words was the winner for the unstructured feature types in isolation. It is important to note that any difficulties LDA faced in the *Standard* problem

are also faced here, i.e. small sample size and small vocabulary, 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 who would be using such models and might derive information from them.

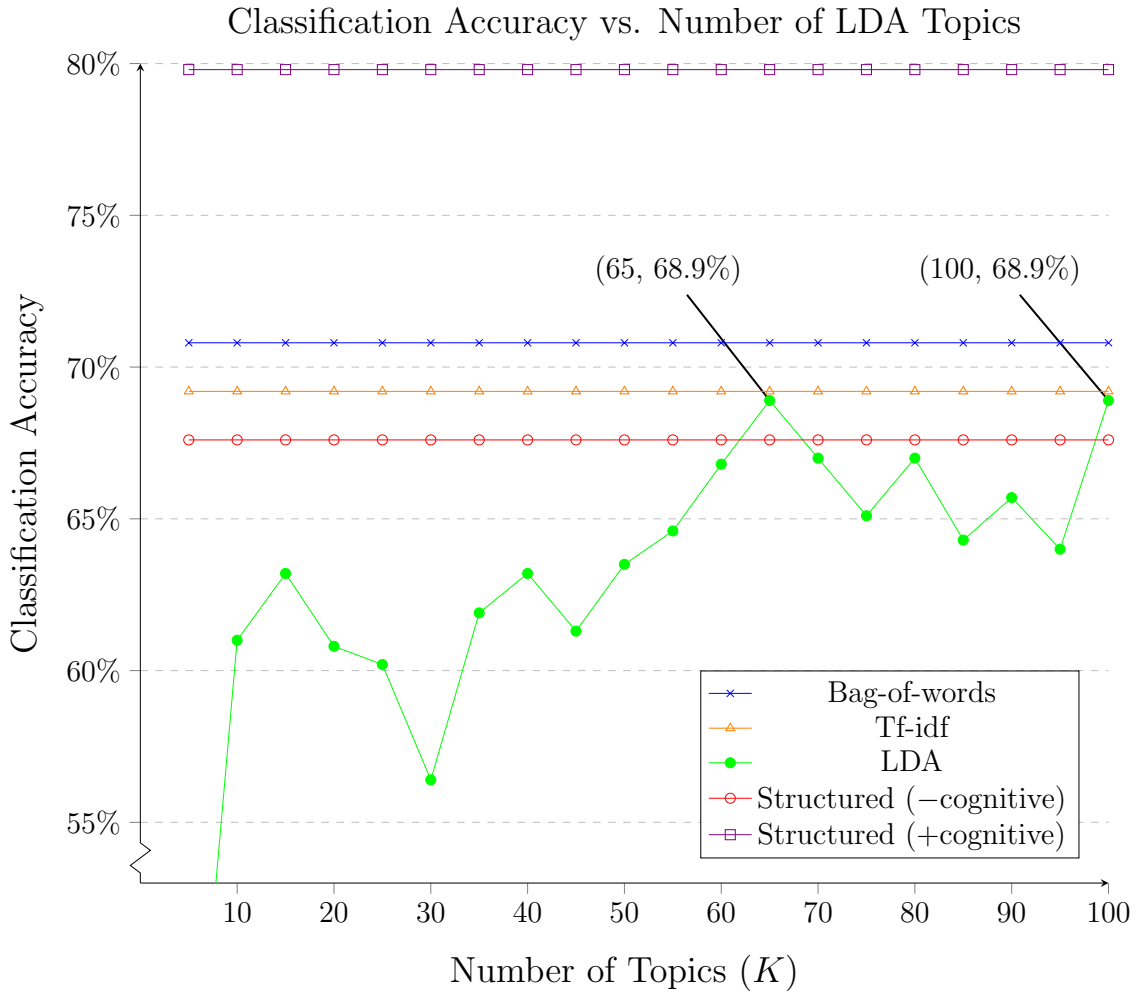


Figure 7.2: Accuracy of LDA-reduced features as the number of topics K increases (by multiples of 5). Lines indicating the performance of bag-of-words and tf-idf are shown for comparison and obviously do not change with the topic number of LDA. Structured feature performance is also shown here. The two local maxima for LDA which slightly surpass the performance of bag-of-words are labeled ($K = 65$ and $K = 100$).

Tables 7.3a and 7.3b show the top 10 topics from the 100 topic model trained on the *Early Risk* subset, based on topic coherence and topic size metrics, respectively. One consequence of the smaller sample of subjects is that the vocabulary becomes smaller and the strength of statistical judgments weakens, resulting in topics that are less interesting, despite their performance as classification features. In this case, only a few of the most coherent topics seem to make sense, namely Topics 25, 36, 38, and 56, which appear to be about cognitive evaluations, smoking habits, routine visits/tests, and cardiac issues, respectively. Topic 55 is a good 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*. Ranking the topics by the topic size metric produces an almost entirely different list, with only Topic 25 in common. In this case, Topic 94 seems to be mostly about depression or related symptoms, which can commonly occur with dementia. Overall, these results do not quite match what was seen in the LDA models trained on the whole dataset in the previous section, presumably due to the smaller and more distinct sample size and vocabulary.

7.2.2 Performance of Integration

The performance trends for the integrated models are slightly more consistent here than they were for the *Standard* classification problem. When cognitive assessment scores are excluded, all integration experiments result in an at least modest improvement, although there is little to no difference between the vector concatenation and posterior probability composition methods. This may suggest that results can be achieved without extra sophistication provided by the latter, or that more sophistication is needed beyond either of these techniques. This is discussed later in Section 8.1. When all structured features are included, accuracy improves by about 1% in most cases, with tf-idf being the best, around 3% above structured alone. This is in line with the *Standard* problem, in which the best performance in LOOCV was achieved with tf-idf and all structured features. In general, these results further justify the integration of unstructured and structured features and/or models.

Topic ID	Coherence	Top Words
7	-71.926	*pressure, blood, AE, hospital, once, visit, change, total, back, went*
57	-72.590	*criterion, score, meet, AGE_>=70_<80, change, complaint, CDR, significant, hypercholesterolemia, cognitive*
38	-73.222	*completed, visit, reported, mg, performed, protocol, testing, study_partner, blood, year*
25	-73.340	*criterion, subjective, corroborated, factor, other, AGE_>=60_<70, continues_meet, score, memory_problems, confounding*
55	-77.251	*hip, left, right, removed, normal, arthritis, cataract, eye, allergy, hand*
64	-79.150	*resolved, antibiotic, prescribed, went, AE, back, asleep, urinary_frequency, difficulty, intermittent*
27	-79.867	*blood, small, time, pressure, visit, lower, leg, month, repair, increased*
33	-81.314	*time, onset, approximate, dizziness, estimated, ADNI, AGE_>=70_<80, reported, fall, pleasant*
36	-85.387	*year, smoked, ago, pack, o, quit, per_day, c, urinary_frequency, memory_problems*
56	-85.392	*work, up, valve, cardiac, aortic, ER, heart, x, cardiologist, visit*

(a) Top 10 topics based on *topic coherence*

Topic ID	Size	Top Words
54	571	*since, screening, PCP, change, knee, decline, memory_problems, still, worse, last_visit*
25	496	*corroborated, subjective, factor, other, criterion, continues_meet, AGE_>=60_<70, memory_problems, score, confounding*
78	488	*criterion, meet, hypercholesterolemia, CDR, assessment, controlled_medication, complaint, score, visit, worsening*
17	391	*year, symptom, approximately, per_day, PCP, pack, day, effect, due, caused*
85	371	*since, OTC, medication, hand, see, problem, think, lot, off, taking*
63	360	*year, difficulty, wife, male, man, history, AGE_>=70_<80, diagnosis, study, functioning*
84	356	*removed, change, time, other, osteoarthritis, clinically, normal, relevant, worsening, per*
30	354	*change, still, function, daily, consistent, memory_problems, complaint, prostate, cognitive_functional, basal_cell_carcinoma*
95	352	*hand, went, diagnosed, PCP, leg, worse, given, running, began, med*
94	350	*last, depressed_mood, low_energy, insomnia, depression, month, memory_problems, anxiety, mg, abdominal_pain*

(b) Top 10 topics based on *topic size*, or the number of word tokens assigned to a given topic.

Table 7.3: Top 10 topics from the 100 topic model, trained on the *Early Risk* subset, based on two metrics. For each topic, the 10 most probable words are shown. See Section 6.3 for definitions of these metrics.

	Leave-one-out Cross-validation		
		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 7.4: Classification performance on *Early Risk* labeling scheme (2 classes) for each modeling technique and feature group. Structured features with and without cognitive features are referenced in the lower sections of the table using the shorthand S_{+cog} and S_{-cog}, respectively. Integration by *vector concatenation* is indicated by $\cup$, and *posterior probability composition* by $\oplus$.

8 Conclusion

This thesis examined the incorporation of unstructured text (natural language) data from electronic clinical records for the task of classifying dementia progression status of subjects in a study on Alzheimer's disease, and additionally explored integration of these data and models with those of structured data. This predictive modeling approach would be beneficial for intelligent diagnostic support systems for automatic screening of patients' electronic data. Results and experiments with unstructured data indicated its viability as a source of useful features for dementia classification, either as a complement to available structured data, or in isolation when structured data are missing, as may often be the case for a condition like dementia. The topic modeling algorithm latent Dirichlet allocation (LDA) was also explored as a form of interpretable dimensionality reduction, for both classification performance evaluation and corpus characterization. The LDA results appear promising in some circumstances, but their instability in classification would need to be further examined in future work.

8.1 Limitations and Future Work

Processing and Modeling

One text processing task that was not performed here was word-sense disambiguation (WSD), which aims to resolve ambiguities at the meaning-level (*senses* of a word) rather than the lexical level (written form of a word). For example, the word *patient* has a different sense

in *the patient has dementia* than it does in *he was a patient man*. This process could be beneficial for topic modeling in larger datasets with less restricted vocabulary, but the corpus used here is relatively small and contains highly specific language. WSD relies on part-of-speech tagging, which could add challenges as well, given the specialized medical lexicon and shorthand grammar. This is compounded by the relatively small size of the dataset, which would likely only be made sparser by WSD. For these reasons, it was omitted here, but may be explored in future work.

Classifier Choice

The choice of logistic regression as the classification algorithm for all experiments was based on its direct computation of a posterior probability over the class labels, as this was a requirement for one for the integration techniques. Logistic regression is not the only example of such an algorithm, but it is commonly used and accepted. One its main drawbacks is the need to tune parameters of the model. An alternative for future work is the Relevance Vector Machine (RVM) [Tipping, 2001], a sparse Bayesian model which eliminates the need for parameter selection and additionally produces sparse probabilistic output. Mainstream software packages are not currently available, and while it is outside the scope of this thesis, it appears a potentially interesting direction for future work.

Topic Modeling for Classification

The standard form of latent Dirichlet allocation (LDA) used here is an unsupervised algorithm meant for exploration and characterization of large text corpora. This thesis attempted to use it for an alternative application of dimensionality reduction of sparse bag-of-words models for use in a supervised machine learning task. There are, however, supervised revisions of the original LDA algorithm which incorporate class labels into the generative model, and may be applicable in subsequent experimentation. Popular examples include supervised LDA (sLDA) [Blei and McAuliffe, 2007] and labeled LDA (L-LDA) [Ramage et al., 2009].

Although sLDA seems appealing, it incorporates a continuous response variable rather than a discrete class label. Another version of sLDA was implemented by Wang et al. [2009] for image labeling and annotating images, but this requires both a class label and a list of annotations, as would be found on tagged images, but is not applicable to the ADNI dataset. Finally, L-LDA expects multiply labeled documents, such as articles with keywords or tagged web pages, which is also not applicable to this dataset. As explained by Ramage et al. [2009], using L-LDA with a corpus containing only singly labeled documents is equivalent to multinomial naive Bayes. Thus it would not be appropriate for the goal of this thesis.

Evaluation of Latent Topics

The computational metrics used in this study (see descriptions in Section 6.3) have been correlated to human evaluation in the past, but it would be interesting in the future to perform human evaluation experiments directly. Such procedures are described by Chang et al. [2009] and are popularly used to judge latent topics produced by LDA. The *word intrusion* task presents human evaluators with high probability terms of a randomly chosen topic, with an additional low probability word from that topic, which is to be identified by the evaluator. The intruding word should be easily identifiable in a good topic, as high probability terms should relate to each other more strongly to they do to the intruder, whereas in a bad topic, the relationships will be harder to determine, and thus the intruder will not stand out. This would require IRB approval for such human involvement in experimentation, as well as access to a large enough pool of dementia specialists necessary for the score calculations of the task. Accordingly, it is left for future work.

Temporal Analysis

This thesis centers around dementia, a condition that may develop over a person's life with neither a clear onset nor a well-understood cause. Cognitive and mental illnesses, along

with other long-term diseases such as cancers, may require more data and historical context for effective predictive modeling. Analysis of temporal ordering has been implemented in structured medical data [Batal et al., 2013, Wang et al., 2008], but its role in research on unstructured text data has focused on extraction of temporal expressions [Harkema et al., 2005, Zhou et al., 2006], rather than modeling how a patient's EHR texts changes over time in relation to a disease condition. Specifically, there is potential for interesting temporal analysis using topic models. Changes in the prominence of topic or topics over time, or even simply their prominence at a particular time, may provide important information and be useful in classification. For example, in a study by Hall et al. [2008], LDA was performed on 12,000 papers from the ACL Anthology archive, spanning 28 years, to model changes and trends in topics of computational linguistics research over time. Similar work might be possible on a larger and more diverse electronic medical dataset. Although the ADNI dataset does contain time-ordered information, it is simply too small to be effectively analyzed in this way. In contrast, this would be worthwhile to explore on distinct data, such as large scale electronic medical records.

Bibliography

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

C.A.L. Bailer-Jones and K. Smith. Combining probabilities. GAIA-C8-TN-MPIA-CBJ-053, July 2011. URL `http://www.rssd.esa.int/doc_fetch.php?id=2968255`.

Deborah E. Barnes, Irena S. Cenzer, Kristine Yaffe, Christine S. Ritchie, and Sei J. Lee. A point-based tool to predict conversion from mild cognitive impairment to probable Alzheimer's disease. *Alzheimer's & Dementia*, 10(6):646–655, Nov 2014. URL `http://www.sciencedirect.com/science/article/pii/S1552526013029415`.

Iyad Batal, Hamed Valizadegan, Gregory F. Cooper, and Milos Hauskrecht. A temporal pattern mining approach for classifying electronic health record data. *ACM Transactions on Intelligent Systems and Technology*, 4(4):63:1–63:22, October 2013.

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

David M. Blei and Jon D. McAuliffe. 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, 2007.

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

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

Roger B Bradford. 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, 2008.

Joseph Bullard, Rohan Murde, Qi Yu, Cecilia Ovesdotter Alm, and Rubén Proaño. Inference from structured and unstructured electronic medical data for dementia detection. In *Operation Research and Computing: Algorithms and Software for Analytics*, 14th INFORMS Computing Society Conference (ICS2015), pages 236–244, Richmond, Virginia, 2015.

Katherine Redfield Chan, Xinghua Lou, Theofanis Karaletsos, Christopher Crosbie, Stuart Gardos, David Artz, and Gunnar Rätsch. 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 2013.

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

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

Carol Friedman, Stephen B. Johnson, Bruce Forman, and Justin Starren. 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, 1995.

David Hall, Daniel Jurafsky, and Christopher D. Manning. Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 363–371, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

Henk Harkema, Andrea Setzer, Rob Gaizauskas, Mark Hepple, Richard Power, and Jeremy Rogers. Mining and modelling temporal clinical data. In *Proceedings of the 4th UK e-Science All Hands Meeting*, pages 259–266, Nottingham, UK, 2005.

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

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

Liangjie Hong and Brian D. Davison. 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, 2010.

Nilesh L. Jain and Carol Friedman. 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, 1997.

Elizabeth F. O. Kern, Miriam Maney, Donald R. Miller, Chin-Lin Tseng, Anjali Tiwari, Mangala Rajan, David Aron, and Leonard Pogach. Failure of ICD-9-CM codes to identify

patients with comorbid chronic kidney disease in diabetes. *Health Services Research*, 41 (2):564–580, 2006.

Li Li, Herbert S. Chase, Chintan O. Patel, Carol Friedman, and Chunhua Weng. 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, 2008.

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

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

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

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 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, 2011.

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. Automated identification of postoperative complications within an electronic medical record using natural language processing. *The American Journal of Medicine*, 306(8):848–855, August 2011.

Michael J. Paul and Mark Dredze. 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 2011.

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. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

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

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

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 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, 2009. Association for Computational Linguistics.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 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 2013.

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

Richard Sproat, Alan W. Black, Stanley F. Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards. Normalization of non-standard words. *Computer Speech & Language*, 15(3):287–333, 2001.

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

Michael E Tipping. Sparse Bayesian learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

Radim Řehůřek and Petr Sojka. 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 2010. ELRA.

Chong Wang, David Blei, and Fei-Fei Li. Simultaneous image classification and annotation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1903–1910, Miami, Florida, 2009.

Taowei David Wang, Catherine Plaisant, Alexander J. Quinn, Roman Stanchak, Shawn Murphy, and Ben Shneiderman. Aligning temporal data by sentinel events: Discovering patterns in electronic health records. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 457–466, Florence, Italy, 2008.

Hua Xu, Zhenming Fu, Anushi Shah, Yukun Chen, Neeraja B. Peterson, Qingxia Chen, Subramani Mani, Mia A. Levy, Qi Dai, and Josh C. Denny. 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, 2011.

Li Zhou, Genevieve B. Melton, Simon Parsons, and George Hripcsak. A temporal con-

straint structure for extracting temporal information from clinical narrative. *Journal of Biomedical Informatics*, 39(4):424–439, 2006.