#autism versus 299.0: Topic Model Exploration of Multimodal Autism Data

A thesis presented
by

Joy Carol Ming

To
Computer Science
in partial fulfillment of the honors requirement
for the degree of
Bachelor of Arts
Harvard College
Cambridge, Massachusetts

April 1, 2015

Abstract

Though prevalence and awareness for Autism Spectrum Disorder (ASD) has steadily increased, a true understanding is hard to reach because of the behavior-based nature of the diagnosis and the heterogeneity of its manifestations. Parents and caregivers often informally discuss symptoms and behaviors they observe from their children with autism through online medical forums, contrasting the more traditional and structured text of electronic medical records collected by doctors. We modify an anchor word driven topic model algorithm originally proposed by Arora et al. (2012a) to elicit and compare the medical concept topics, or "themes" from both modes of data: the novel data set of posts from autism-specific online medical forums and electronic medical records. We present methods to extract relevant medical concepts from colloquially written forum posts through the use of choice sections of the consumer health vocabulary and other filtering techniques. In order to account for the sparsity of concept data, we propose and evaluate a more robust approach to selecting anchor words that takes into account variance and inclusivity. This approach that combines concept and anchor words selection seeds the discussion about how unstructured text can influence and expand understanding of the enigmatic disorder, autism, and how these methods can be applied to similar sources of texts to solve other problems.

The code and its instructions can be found at `https://github.com/jming/thesis`. Any questions or comments can be directed to `joy.c.ming@gmail.com`.

Acknowledgements

Thank you to Professor Finale Doshi-Velez for her attention and guidance throughout this research project and my research education. I would also like to thank Professors Ryan Adams and Peter Szolovits for generously offering to be my thesis readers. Moreover, I would like to thank Sam Wiseman and Hongyao Ma from Harvard SEAS for their theoretical and code contributions to the project infrastructure and Pei Chen and Guergana Savova on the Boston Children's Hospital Informatics team for their initial processing of the health records. And a special thanks to my family and friends for being there for me through the process.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Motivation

In the information age, there is a plethora of social data available on the Internet (Chou et al., 2009). However, this data is not being used to its fullest potential, especially by professionals. One example of such data is medical forums, communities where patients and their caregivers discuss symptoms, treatments, and other topics related to a given disorder (Denecke and Nejdl, 2009; Hawn, 2009; Sarasohn-Kahn, 2008). These forums present a wealth of information in looking at pharmacovigilance and adverse drug effects (Almenoff et al., 2005; Chee et al., 2011). Yet this information could be used for more.

Autism spectrum disorders (ASDs) include a variable presentation of neurodevelopmental disorders characterized by impairments in communication, reciprocal social interaction, and restricted behaviors or interests (Faras et al., 2010). Recent prevalence rates for ASDs are now estimated at about 1 in 68 children in the United States (Baio, 2012). Symptoms for autism typically are apparent before age three and timely diagnosis is important in beginning early intervention (Baio, 2012). However, because ASD remains a diagnosis that is defined on the basis of behavior, diagnostic assessment is complex (Lord et al., 2013). Furthermore, autism is fairly heterogeneous and has many etiologies, with a wide variety of manifestations of symptoms throughout the spectrum (Geschwind and Levitt, 2007). This means that the same diagnosis of ASDs could include patients with diverse comorbid disorders such as epilepsy or attention deficit hyperactivity disorder as well as different levels of social interaction, from nonverbal to high-functioning (Matson and Nebel-Schwalm, 2007).

Social media provides a unique outlet for characterizing the heterogeneity of autism. Unlike electronic health records, which is the status quo for understanding symptoms of disease and disorders, medical forums can record more detailed descriptions given by caregivers who spend time with autism patients, information that might not be present in clinical records

(Frost and Massagli, 2008). In this project, we attempt to make use of this unstructured online information and compare it to electronic health records to determine more truths about the diversity of symptoms of autism.

1.2 Approach

It is difficult to understand the content and meaning of hundreds of thousands of forum posts because of the difficulty in textual processing. This task is especially difficult because the online forum posts are unstructured, as is most text found in web spaces, with a myriad of errors in spelling and grammar.

One method of trying to find order in the unstructured text is topic modeling. *Probabilistic topic modeling* is a statistical method used to analyze, organize, and understand large sets of texts by uncovering themes, or topics, that run through them as well as how those themes are connected to each other and how they change over time (Blei, 2012). These algorithms have been successfully applied to understanding text in many areas, including, but not limited to email, scientific abstracts, and newspaper archives (Griffiths et al., 2004; Blei et al., 2003).

This project uses topic modeling to explore the online medical forum information related to autism in order to better understand how conditions or characteristics cluster and the overall heterogeneity of autism. By uncovering the underlying themes in online medical forum data and electronic medical records, we can better understand the discussion of symptoms and potentially apply this to future diagnostics or treatments.

First, more than 200,000 posts are scraped from main autism forums, including Autism Forum, ASD Friendly, Autism Web, and Talk About Autism. Then, the posts are analyzed using a modified version of the anchor words driven probabilistic topic modeling algorithm proposed by Arora et al. (2012a). Using this same pipeline, electronic medical records that have been processed using natural language processing are examined to see whether clustering in the more structured and verified data of electronic medical records matches with patterns found in the unstructured lay text of online medical forums.

1.3 Contributions

This paper makes the following contributions:

- Exploration of an aggregation of novel data from many of the top online medical forums related to the enigmatic and increasingly prevalent disorder of autism.

- Description of the approach of the translation from colloquial, unstructured sources to medical concepts using a trie created based on the consumer health vocabulary, selectively pruned to include only concepts related to the behavior and diagnosis of autism.

- Modification and evaluation of the anchor-word-driven topic modeling algorithm propose by Arora et al. (2012a) to find more robust anchors for datasets with less words to account for variance and inclusivity.

- Comparison of topics generated from two distinct data sources, online medical forums and electronic medical records, based on the application of the pipeline of concept-mapping and the more robust anchor-word driven topic modeling algorithm.

The rest of the report is organized in the following manner. First, background related to the application area, autism, is explored as well as past work in the methodology, or topic models in Chapter 2. Chapter 3 explains the methods, namely the basic exploration of the data and a description of the pipeline. Then Chapter 4 highlights results of the pipeline applied to the datasets and the subsequent adjustments to the methods. Finally, the results and their implications are explored in Chapter 5.

Chapter 2

Background

2.1 Why Autism?

Autism spectrum disorders (ASDs) are a range of neurodevelopmental disorders characterized by social, communicative, and behavioral impairments and affect 1 in every 68 children in the United States (Baio, 2012). Though prevalence and awareness has steadily increased, diagnosis is currently behavior-based and the heterogeneity of the manifestations of the disease make it different to fully understand the disorder. We select autism as the focus of this study due to its increasing awareness (Section 2.1.1), pervasive uncertainty (Section 2.1.2), and behavioral diagnosis (Section 2.1.3).

2.1.1 Increasing Awareness

The enigmatic disorder of autism has taken its place in the spotlight in the past few years because of increasing prevalence. The graph below shows autism diagnosis rates from 2000-2010 as measured by the Autism and Developmental Disabilities Monitoring (ADDM) Network.

Surveillance Year	Birth Year	Prevalence (per 1,000)	Approximation
2000	1992	6.7 (4.5-9.9)	1 in 150
2002	1994	6.6 (3.3-10.6)	1 in 150
2004	1996	8.0 (4.6-9.8)	1 in 125
2006	1998	9.0 (4.2-12.1)	1 in 110
2008	2000	11.3 (4.8-21.2)	1 in 88
2010	2002	14.7 (5.7-21.9)	1 in 68

Table 2.1: Identified prevalence of Autism Spectrum Disorder (ADDM).

These increasing diagnoses have been played a part in the increase in awareness. However, there has also been more press around autism, creating an "epidemic of panic," where its

broad diagnosis has fuelled fears about vaccines despite no evidence for a link (Konner, 2011).

This translates to more activity on social media, from the talk in the general public about autism with regards to the vaccines to the increased discussion among caregivers and patients of autism.

2.1.2 Pervasive Uncertainty

While there are some leads to hormonal, genetic, neurological, and environmental underpinnings of autism, the exact causes and manifestations of autism are not clear. Some of the hypotheses are listed below:

- **Hormonal.** According to statistics on the prevalence, ASD is almost 5 times more common among boys (1 in 42) than among girls (1 in 189) (Baio, 2012). Foetal testosterone (FT) influence later social and communication development and can be linked to the number of autistic traits a child has (Baron-Cohen, 2006).

- **Genetic.** The existence of autism in twins and siblings emphasizes the possibilities of genetics. Gene mutations, gene deletions, copy number variants, and other genetic anomalies are persuasively linked to autism (Sutcliffe, 2008). Most research has looked at multiple interacting genetic factors as the main causative determinants of autism (Muhle et al., 2004). Others look at the association of autism with comorbid or related disorders like epilepsy or Fragile X (Muhle et al., 2004). However, the search for an "autism gene" has been less successful.

- **Neurological.** Pyschologists have examined the possibilities of the theory of mind and mirror neurons in their influence on the inability of individuals diagnosed with autism to reciprocate or understand emotions (Baron-Cohen, 2004). Researchers in neuroscience often look towards different regions of the brain such as the cerbellum or thalamus, in their functionality and their size (Carper and Courchesne, 2000; Tsatsanis et al., 2003). Others have investigated functional and structural connectivity of the different regions of the brain, through white/grey matter or long/short distance connections (Alexander et al., 2007; Courchesne and Pierce, 2005; Belmonte et al., 2004).

- **Environmental.** Epidemiologic studies have investigated environmental factors, finding that external exposures to lead, ethyl alcohol, and methyl mercury or teratogens, perinatal insults, and prenatal infections may be plausible but synthetic chemicals contained in immunizations with the measles-mumps-rubella are not (Landrigan, 2010; Muhle et al., 2004).

In the end, some have declared that it is "time to give up on a single explanation of autism"—there will be no single genetic or cognitive cause for the diverse symptoms defining autism (Happé et al., 2006).

2.1.3 Behavioral Diagnosis

Since there is no clear picture of its cause, autism has been relegated to a purely behavioral diagnosis. This diagnosis is guided by the DSM-5, the latest version of the Diagnostic and Statistical Manual of Mental Disorders (DSM), placing autism under code 299.0 with language similar to the following excerpt (Association et al., 2013):

> Persistent deficits in social communication and social interaction across multiple contexts, as manifested by the following, currently or by history (examples are illustrative, not exhaustive; see text):
>
> 1. Deficits in social-emotional reciprocity, ranging, for example, from abnormal social approach and failure of normal back-and-forth conversation; to reduced sharing of interests, emotions, or affect; to failure to initiate or respond to social interactions.
> 2. Deficits in nonverbal communicative behaviors used for social interaction, ranging, for example, from poorly integrated verbal and nonverbal communication; to abnormalities in eye contact and body language or deficits in understanding and use of gestures; to a total lack of facial expressions and nonverbal communication.
> 3. Deficits in developing, maintaining, and understand relationships, ranging, for example, from difficulties adjusting behavior to suit various social contexts; to difficulties in sharing imaginative play or in making friends; to absence of interest in peers.

The DSM-5 often supplemented with associated behavioral test checklists such as Autism Diagnostic Observation Schedule (ADOS), Autism Diagnostic Inteview (ADI), Social Responsiveness Scale, and Childhood Autism Rating Scale (CARS) (Lord et al., 2012; Rutter et al., 2005; Constantino et al., 2003; Schopler et al., 1988). These generally consist of observation sessions of behaviors by the diagnostician and reports from caregivers of ratings of the child's behavior.

However, purely behavioral diagnoses are quite volatile. Changes in practices for diagnosis autism have a had a substantial effect on autism caseloads in the past, accounting for a quarter of the observed increase in prevalence between 1992 and 2005 (King and Bearman, 2009). And the recent changes in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) diagnosis of autism renegated up to 19% of the classification of males

who were originally diagnosed as autistic under the DSM-IV (Baio, 2012). Furthermore, there are variations in the interpretation of the diagnosis across variables such as culture and socioeconomic class (Ennis-Cole et al., 2013).

Another issue is that autism is not picked up until very late, the average age of diagnosis is around 3 years old, as most of the more obvious symptoms such as nonverbal behaviors or aversion to eye contact are not exhibited until then (Mandell et al., 2005). But this places physicians at a disadvantage, because it is imperative that the treatments be given as early as possible, early intervention playing a large role in combatting the symptoms of autism (Valicenti-McDermott et al., 2012).

2.1.4 Contribution

This project aims to better understand this enigmatic disorder, harnessing the increased awareness and discussion, specifically in relation to the parents, the caregivers that are interacting with children with autism. Taking this dialogue around autism to better understand the behaviors associated with it can contribute to another facet to the behavior-based diagnosis process or explore insights into the understanding of treatments. Not only will this approach leverage the existing trends in the increasing prevalence and discussion, it also explores a new dimension into an age-old problem.

2.2 What Are Topic Models?

As the amount of text builds up online, through mediums including web pages, social media, articles, images, genetic sequences, user ratings, and more, it becomes difficult to sort and search manually, motivating the need for automatic methods to create order and understanding. Topic models help structure this growing and increasingly disorganized corpus of accessible text by using themes, or *topics*. Topic models uncover the underlying semantic structure of a document collection based on hierarchical Bayesian analysis of the original texts, discovering patterns of word use and connecting documents that exhibit similar patterns (Blei and Lafferty, 2009).

2.2.1 Example

Topic modeling builds upon the assumption that documents are a mixture of topics, where a topic is a probability distribution over words (Blei, 2012). Consider the example of online medical forum posts related to autism. Each of the posts contains words that are associated

with larger topics. In this example, we will use the hypothesized topics of "comorbid disorders", "dietary supplements", and "treatment options" as well as the fabricated related words seen in Table 2.2.

comorbid disorders (0.3)	dietary supplements (0.2)	treatment options (0.5)
epilepsy (0.14)	melatonin (0.20)	applied behavior analys (0.65)
ADHD (0.11)	Vitamin B12 (0.12)	pivotal response (0.12)
GI problems (0.08)	Casein-free (0.09)	music therapy (0.02)

Table 2.2: Sample topics and associated words (and their probabilities).

Therefore, the assumption of topic models is that in order to create each forum post, a new word is added by drawing a topic at random according to the distribution and then selecting a word from that topic (Steyvers and Griffiths, 2007). In this example, one of topics is chosen, such as "treatment", and then a word is selected from that topic at some probability, such as pivotal response. Then another topic is selected and a word within that topic, etc. This continues until each document will be created. This is a *generative model* that is used to build each of the individual elements in topic modeling, in this case, each individual forum post.

In this case, a possible forum post would look something like this: "pivotal response epilepsy casein-free ADHD." Of course, this seems nonsensical because it is missing filler words that might not be related to a medical topic but are necessary to create a flow in the English language. It is also nonsensical because it is unlikely that someone would choose to talk about so many distinct concepts in a single forum post, which is typically shorter and less concept-heavy. Furthermore, given the nature of the Internet, forum posts are typically spelled incorrectly or have the words in different orders.

Therefore it is difficult to determine the original topics and distribution that was used to generate each of the individual elements simply by looking at the text which may contain additional elements or be less structured than assumed. Using the observed documents to infer the most likely hidden topic structure is the central computational problem for topic modeling, effectively "reversing" the generative process described above (Blei, 2012). Therefore, in order to uncover this underlying structure, it is necessary to guess a hidden structure in the observed data and learn that structure using posterior probabilistic inference, matching a *hidden variable model* of the documents (Blei and Lafferty, 2009).

2.2.2 Latent Dirichlet Allocation

One of the basic approaches is *latent Dirichlet allocation (LDA)*, whichi s based on the Dirichlet distribution, Dir(α), the multivariate generalization of the beta distribution and

the conjugate prior of the categorical distribution and multinomal distribution (Johnson et al., 2002).

LDA is a generative probabilistic model of a corpus, a statistical model of a collection of texts *hidden random variables* that encode its thematic structure, which is articulated by its *probabilistic generative process* (Chaney and Blei, 2012). This process that was described in the example is formalized below with algorithm descriptions from Blei et al. (2003) and Chaney and Blei (2012):

1. For K topics, K Poisson(ξ), choose each topic distribution β_k, or the distribution over the vocabulary.
2. For each document in the collection:
   (a) Choose a distribution over topics θ Dir(α), or the distribution over K elements.
   (b) For each word in the document:
      i. Choose topic assignment z_n Multinomial(θ), a number from 1 to K, from θ_d
      ii. Choose a word w_n from the topic distribution $p(w_n|z_n, \beta)$, selecting the z_nth topic from step 1.

Beyond the basic LDA approach are approaches that challenge some of the basic assumptions innate to the topic modeling strategy. This includes building a Markov chain or Hidden Markov Models (HMM) to break the bag of words assumption (Wallach, 2006; Griffiths et al., 2004); using dynamic topic models to respect the ordering of the documents (Blei and Lafferty, 2006); or Bayesian nonparametric topic models to confront the assumption that the number of topics is assumed to be known and fixed. Other methods will incorporate metadata such as the author-topic model (Rosen-Zvi et al., 2004) or include a mixed-membership model of grouped data (Pritchard et al., 2000). The main alternative to LDA is probabilistic latent semantic analysis (PLSA), which is based on a mixture decomposition derived from a latent class model as opposed to using singular value decomposition (Hofmann, 1999).

2.2.3 Anchor Words

Arora et al. (2012a) pioneered an approach to topic modeling that no longer relies on singular value decomposition (SVD), rather, utilizes nonnegative matrix factorization (NMF), creating a polynomial-time algorithm for learning topic models that allows for the uncovering of anchor words (Arora et al., 2012b). This is based on the *separability* assumption, which is that every topic contains at least one *anchor word* that has a non-zero probability only in that topic (Donoho and Stodden, 2003). Therefore, if a document contains a given anchor

word, that document is guaranteed that the corresponding topic is among the set of topics used to generate the document (Arora et al., 2012a).

Some models, such as the probablistic relational model, use both words on the page and anchor words on the links to predict the category, better understanding the linkages between the texts (Getoor et al., 2001). Anchor text, in addition to links and user clickthrough data, has also been used to cluster web pages (Ramage et al., 2009).

In order to find the anchor words, it is important to ensure that the words in the document can be interpreted as points in an plane (Arora et al., 2012a). This paper will elaborate on a new method of uncovering anchor words based on the convex hull assumption. Instead of focusing only on heuristics that try to find anchor sets that span all of points that represent the words in the documents, this algorithm will actually test the anchor sets on the actual points given. This is explored later on in the paper.

2.2.4 Contribution

Topic models have been applied to many kinds of documents, including email, scientific abstracts, and newspaper archives (Griffiths et al., 2004; Blei et al., 2003). Additionally, topic models have also been applied to various more abstract concepts, including mapping relations between webpages or understanding the geneome (Getoor et al., 2001; Flaherty et al., 2005).

Now that there is more text on the web, topic models are being used for the text that is now being coded on the web. This research project looks at online medical forums, text that is generally previously unexplored with the use of topic models. This particular source of text does lend some properties that make it a great candidate for topic modeling, including having a nested nature, with subforms and responses regarding a specific topic, having a rich variety of topics and discussions, and simplying have a large volume to explore. Furthermore, this research project focuses on autism because of its increasing prevalence but sustained uncertainty, as described in the previous section.

Chapter 3

Methods

This section describes the methods that were applied to analyze the two sources of text, online medical forums and electronic medical records. These methods are used on both of the sources separately and the resulting topics compared for further analysis.

Section 3.1 describes the process by which the mentions and references to medical concepts were extracted from the text using the consumer health vocabulary. Section 3.2 describes the pipeline, which starts with the creation of a matrix that counts the number of times the extracted concepts appear in the same posts (Section 3.2.1), application of an anchor-driven algorithm for topic modeling through anchor selection (Section 3.2.2) and recovery (Section 3.2.3), and finally translation and exploration of the results (Section 3.2.4).

3.1 Concept extraction

This section describes the methods used to organize the unstructured text written by lay people in the online medical forums: first mapping the words to medical concepts (Section 3.1.1) and then creating a count matrix of how often these medical concepts appear together in a given post (Section 3.1.2).

3.1.1 Concept Mapping

In order to find truth in the unstructured text of the posts scraped from various online medical forums, a common vocabulary is needed in order to analyze the clustering of autism characteristics and behaviors.

There are many vocabularies used by professionals with regards to medical concepts. The base corpus is the unified medical language system (UMLS), which links health and biomedical vocabularies and standards together to enable interoperability between computer

systems and research goals (Lindberg et al., 1993). The most specific subset of the UMLS used in this research study is the international classification of diseases (ICD-9), which includes only words associated with diseases and diagnoses (of Health, 1980).

Looking at the unstructured text of the posts scraped from the various online medical forums, it was decided to find health-related concepts through the use of *consumer health vocabulary* (CHV). This helps to address the mismatch between terms used by laypersons on the internet and the concepts highlighted by healthcare professionals and help highlight the important ideas needed for further analysis (Smith and Stavri, 2005).

CHV database description

The Consumer Health Vocabulary (CHV) Initiative provides a comprehensive database with a mapping between every day words and phrases about health ("heart attack") to technical terms or jargon used by health care professionals ("myocardial infarction") associated with the same given concept (Zheng, 2014). Each of the concepts, which many lay words can map to, has a unique UMLS concept ID (CUI) and a preferred name. There are 158,519 terms mapped to 57,819 unique CUIs in the CHV database.

In the following example, all of the listed terms and a few unlisted terms are all associated with CUI C0027051 and the preferred name "myocardial infarction."

CUI	**Term**	**Preferred Name**
C0027051	AMI	myocardial infarction
C0027051	attack heart	myocardial infarction
C0027051	attack hearts	myocardial infarction
C0027051	attacking heart	myocardial infarction
C0027051	attacks coronary	myocardial infarction
C0027051	cardiac infarction	myocardial infarction
C0027051	coronary attack	myocardial infarction
C0027051	disorder infarction myocardial	myocardial infarction
C0027051	heart attack	myocardial infarction
C0027051	...	myocardial infarction

Table 3.1: Example of set of entries in consumer health vocabulary.

CHV trie

Each post is processed so that its content is matched up with phrases in the CHV. However, the CHV has 158,519 terms, so it is difficult to match up each individual phrase of the 245,717 posts scraped from the online medical forums with each of the CHV terms.

In order to process the posts, a trie was created to store the various terms and the CUIs they are associated with. The trie is created using nested dictionaries and loaded at the beginning of the process. Figure 3.1 is the trie associated with the terms in the example in Table 3.2.

CUI	Term	Preferred Name
C1261512	attack	attack
C0235462	attack angina	anginal attack
C0700031	attack anxiety	anxiety attack (finding)
C0027051	attack heart	myocardial infarction
C0027051	attack hearts	myocardial infarction
C0086769	attack panic	panic attacks
C0683920	attack rate	attack rate
C0683920	attack rates	attack rate
C0855247	attack sleep	sleep attacks

Table 3.2: Examples of CHV entries associated with the trie in Figure 3.1.

A trie was selected because of its optimal computational complexity for search and recovery of concepts. A standard binary search or binary search tree over the concepts would take $O(M \log N)$ time to find a key, where M is the maximum string length and N is the number of keys, which in this case is , A trie will take simply $O(M)$ time to recover a key. Furthermore, the trie data structure builds upon the unique feature of this database, that often times a word with a similar prefix will be mapped to the same CUI. This is helpful in the next step, with the concept lookup for each of the posts.

The posts that are scraped are processed individually, feeding in each of the posts to look for the longest instances of a matching concepts in each post. For example, matching the sentence "The bear attacks me" to Table 3.2 would result in highligthing with C1261512, whereas the sentence "The bear attacks me at a high attack rate" results in highlighting C1261512 and C0683920.

This process results in a count where the unique CUIs are associated with the post numbers of the posts in which it is included.

3.1.2 Concept Co-occurrences

Once the posts were processed so that there is a list of concepts associated with each post, this was transformed into a data structure that would reflect the relationship between the different concepts in a given post. This is seen through conditional probability and the co-occurrence matrix.

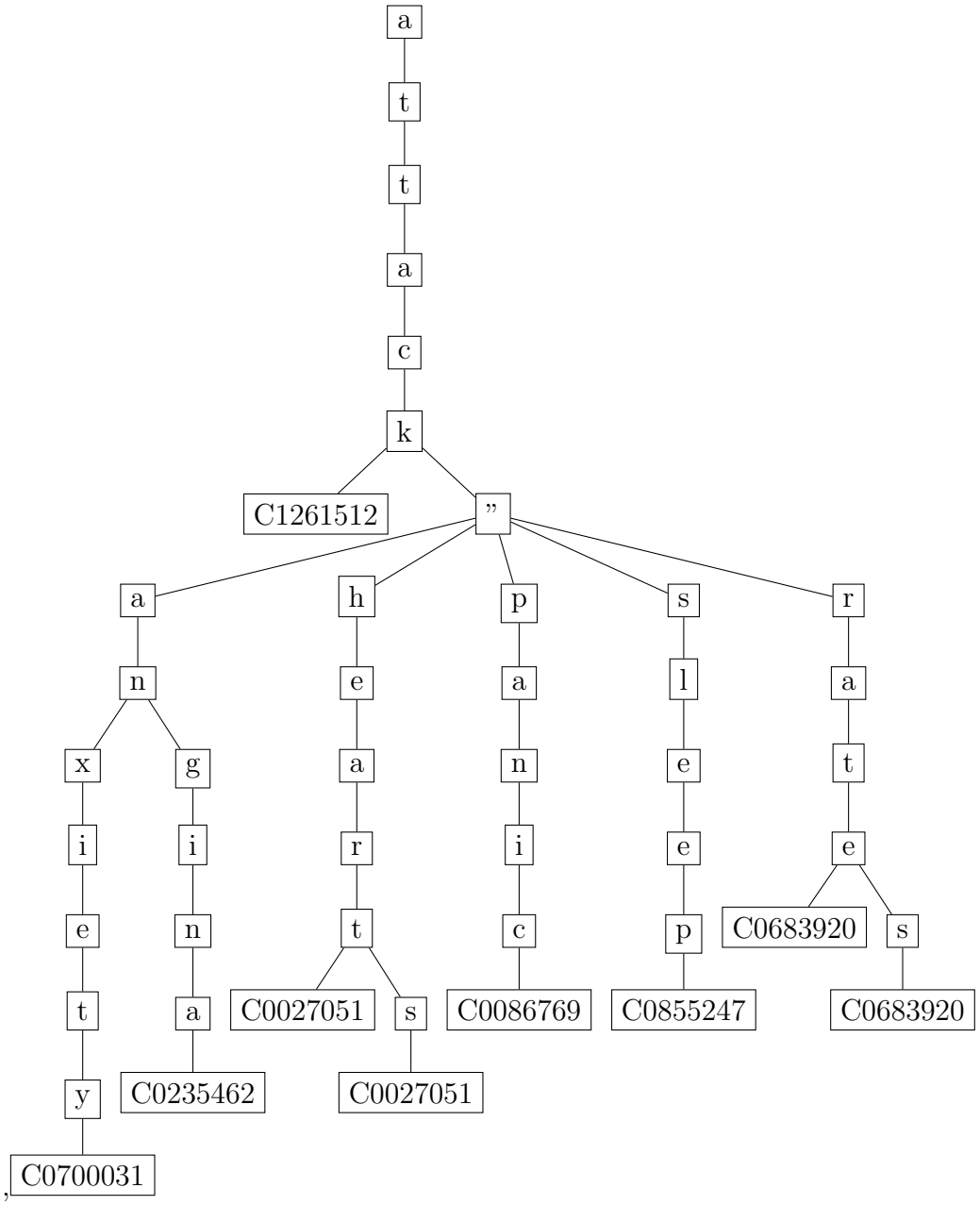


Figure 3.1: Example section of CHV trie based on terms in Table 3.2.

Conditional probability

The idea behind the clustering of different concepts within a given post is first explored by looking at the conditional probabilities between each of the concepts, or the probability that concept B will appear in a post given that concept A already appears in the post.

For example, we can look at the probability the "seizure" concept will appear in a post given that the concept "epilepsy" is already present in the text. By Bayes' rule, this probability is equivalent to the probability both words appear in a post divided by probability just epilepsy appears in the post.

$$P(\text{seizure} \mid \text{epilepsy}) = \frac{P(\text{seizure} \cap \text{epilepsy})}{P(\text{epilepsy})} \tag{3.1}$$

The simple probability of a concept appearing in a post is equivalent to the number of times that concept appears in all the posts divided by the total number of posts. And the probability two concepts are in the same post is the number of times both concepts appear in the same post in all the posts divided by the number of posts.

$$P(\text{seizure} \cap \text{epilepsy}) = \frac{\#\text{ posts with seizure and epilepsy}}{\#\text{ total posts}} \tag{3.2}$$

$$P(\text{epilepsy}) = \frac{\#\text{ posts with epilepsy}}{\#\text{ total posts}} \tag{3.3}$$

Therefore, looking at the original equation, the count of the total number of posts in both the denominators cancels out, resulting in the following equation that can be applied to any number of words appearing in a post given that another word appears in a post.

$$P(\text{seizure} \mid \text{epilepsy}) = \frac{\#\text{ posts with seizure and epilepsy}}{\#\text{ total posts}} \tag{3.4}$$

For a sanity check, this method was applied to some example concepts that may or may not appear in the same post. This is based on some basic intuitions about the concepts. For example, it is intuitive that a post that includes the word "epilepsy" should have a high probability of also including the word "seizures."

Concept co-occurrence matrix

Since the calculation of conditional probability is based on the number of times two concepts appear in a given post, it makes sense to count the co-occurrence of terms in a post. The desired end result is a representation of the number of times a concept will co-occur with another concept.

This was calculated using the concept mapping list as described in Section 3.1.1, which has a list of the CUIs and the posts they are included in. The original brute-force approach looked at pairs of CUIs, counting the length of the set intersection of posts to see the number of posts they share. However, this $O(n^2)$ algorithm took hours to process the 158,519 CUIs.

In order to make the calculation time reasonable, the concept mapping list was then converted into a matrix $C \times P$ matrix, where C is the number of concepts or number of CUIs and P is the number of posts. Each entry a_{ij} is a 0 or 1 to indicate whether concept i appears in post j.

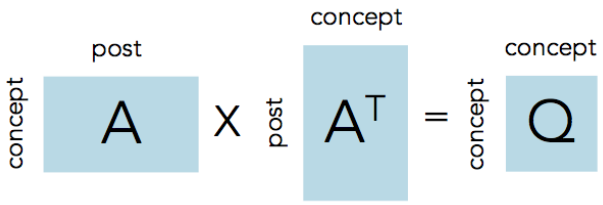


Figure 3.2: Visual representation of the process of deriving the concept-concept co-occurrence matrix Q from a concept-post count matrix A

Therefore, finding the $C \times C$ matrix where each entry a_{ij} is a sum of the number times the ith CUI appears in the same post as the jth CUI was simply reduced to a matrix multiplication. The original $C \times P$ matrix, A multiplied by its transpose, or the $P \times C$ matrix A^T results in the necessary $C \times C$ matrix. This ended up being a more tractable and reasonable approach to finding the number of co-occurrences of a given concept.

Finally, this matrix is normalized so that each column sums up to 1. Since the matrix is stored as a sparse matrix as a dictionary of keys, the initial approach looked at the nonzero CUI, CUI key pairs and dividing each entry by the sum of its column. However, summing the 158,519 columns on the spot took too much time. Therefore, the final approach keeps a separate matrix with sum of each column from the original matrix creation step and divides the entries based on these sums.

3.2 Data Pipeline

The posts are first processed to produce a concept-concept co-occurrence matrix Q that tracks the number of times a CHV concept appears with another concept in the same post (Section 3.2.1). Then, the Q matrix is processed with the anchor-words driven topic modeling algorithm which undergoes two steps: *anchor selection* (Section 3.2.2), which identifies anchor words, and *anchor recovery* (Section 3.2.3), which recovers the concepts associated

with each anchor word (Arora et al., 2012a; Halpern, 2014). The high-level algorithm is presented in Algorithm 1.

Algorithm 1: High-level algorithm, drawn from Arora et al. (2012a).

Data: Posts P, Concepts C, Number of anchors K, Tolerance parameters $\epsilon_\alpha, \epsilon_\beta > 0$.
Result: Concept-topic matrix A, topic-topic matrix R.
begin
  $P^C \longleftarrow$ Post-concept mapping(P, C);
  $Q \longleftarrow$ Concept co-occurrences(P^C);
  $\bar{Q} \longleftarrow = \{\bar{Q}_1, \bar{Q}_2, ..., \bar{Q}_V\}$ Normalize rows(Q);
  $S \longleftarrow$ FastAnchorWords($\bar{Q}, K, \epsilon_\alpha$) (Algorithm 2);
  $A, R \longleftarrow$ RecoverKL(Q, S, ϵ_β) (Algorithm 3);
  return A, R
end

The code and instructions for running this pipeline can be found at `https://github.com/jming/thesis` written with Python using the numpy, pandas, matplotlib, and scikit-learn libraries and is built upon by code from Sam Wiseman, Hongyao Ma, and Yoni Halpern.

3.2.1 Matrix Creation

The co-occurrence matrix detailed in Section 3.1.2 can be modified to serve as a concept-concept co-occurrence matrix Q through the process described in Figure 3.2. First, the posts are mapped to the concepts using the CUI trie, resulting in a dictionary of CUIs and the posts it appears in. Then, this is changed into a $C \times P$ matrix A where C is the number of concepts and P is the number of posts and each entry a_{ij} indicates whether concept i appears in post j. Finally, in order to find a $C \times C$ concept-concept matrix, the original matrix A is multiplied by its transpose A^T, resulting in a matrix that tracks the number of times a CHV concept appears with another concept in the same post.

This matrix then needs to then be normalized so that the sum of all entries is 1. This is similar to the normalization of the matrix based on the sum of all occurrences of each CUI. Therefore, the number of occurrences of each CUI in all of the posts is tracked and then all entries of the matrix are divided by this value based on their row.

3.2.2 Anchor Selection

This step will uncover a set of anchor words. Anchor words can be used as representations each of the topics because they hold the nonzero probability of being in the topic. In order

to find these anchor words, the assumption is made that each of the entries in the co-occurrence matrix can be treated as a point in a space with as many dimensions as there are concepts. Then, the points in the Q-matrix can be treated as a convex hull. These anchor words essentially will form a basis, or simplex, which can be modified to include a linear combination of the points.

The original approach from Arora et al. (2012a) uses a heuristic to find the set of anchor points that best covers the convex hull given. This heuristic is based on finding the points that are the furthest away from each other and therefore most likely to include the most number of points in the convex hull. This is calculated in the algorithm to recalibrate the distance based on the span of the current basis with multiple rounds of replacement.

This algorithm is described in Algorithm 2. Given V points $\{d_1, d_2, ..., d_V\}$ in V dimensions, with K vertices, start with random point, such as the farthest point from the origin, to initialize the basis set, or simplex, S. Then, for $K-1$ times, find the point with the most error, or the largest orthogonal component to the span, and add it to the basis set S. Given this S with K elements, repeat the previous step, removing one point at a time and replacing it with the point that has the largest distance to the span of the new set. Then, return this set S as K points $\{v'_1, v'_2, ..., v'_K\}$ that are close to the vertices of the simplex, where v'_i is the row of the ith anchor word.

Algorithm 2: FastAnchorWords, drawn from Arora et al. (2012a).

Data: V points $\{d_1, d_2, ..., d_V\}$ in V dimensions, almost simplex, K vertices and $\epsilon > 0$.
Result: K points that are close to the vertices of the simplex.
begin
  Project the points d_i to a randomly chosen $4 \log V \epsilon^2$ dimensional subspace;
  $S \leftarrow \{d_i\}$ s.t. d_i is the farthest point from the origin;
  for $i = 1$ *to* $K-1$ **do**
    Let d_j be the point in $\{d_1, d_2, ..., d_v\}$ that has the largest distance to span(S);
    $S \leftarrow S \cup \{d_j\}$;
  end
  $S = \{v'_1, v'_2, ..., v'_K\}$;
  for $i = 1$ *to* K **do**
    Let d_j be the point that has the largest distance to span($\{v'_1, v'_2, ..., v'_K\}\{v'_i\}$);
    Update v'_i to d_j;
  end
  return $\{v'_1, v'_2, ..., v'_K\}$
end

3.2.3 Anchor Recovery

Then, after the anchor words are discovered, the words associated with them in each topic are recovered using a quadratic program. This then returns the concept-topic matrix A and the topic-topic matrix R. This is described in Algorithm 3. Given a $C \times K$ or concept by topic matrix, where each row sums to 1 and each entry is nonnegative, for each row, find C_{ik}, the variable in the quadratic program, with the objective $||\bar{Q}_i - \sum_k C_{ik}\bar{Q}_{s_k}||$ and constraints $C_{ik} \geq 0$ and $\sum_k C_{ik} = 1$.

Algorithm 3: RecoverKL, drawn from Arora et al. (2012a).

Data: Matrix Q, Set of anchor words S, tolerance parameter ϵ.
Result: Matrices A, R.
begin
  Normalize rows of Q to form $\bar{Q}$;
  Store normalization constances $\vec{p}_w = Q\vec{1}$;
  $\bar{Q}_{sk}$ is the row of Q for the k^{th} anchor word;
  for $i = 1, ..., V$ **do**
    Solve $C_i = \text{argmin}_{\vec{C_i}} D_{KL}(\bar{Q}_i || \sum_{k \in S} C_i, k\bar{Q}_{s_k})$;
    Subject to $\sum_k C_{i,k} = 1$ and $C_{i,k} \geq 0$;
    With tolerance ϵ;
  end
  $A' = \text{diag}\ (\vec{p}_w)C$;
  Normalize the columns of A' to form A;
  $R = A^{''}QA^{''T}$;
  return *A, R*
end

3.2.4 Result Translation

In order to treat each of the concepts as words as required by the algorithm, the matrices are indexed by the CUI or concept ID. The final step in ensuring the readability of the result is the translation into plain text English. This use the CHV the dictionary's mapping of the concept ID to the concept name.

Chapter 4

Results

This chapter will tell the story of the application of the methods on the data sets as described in Chapter 3. The chapter starts with a description of the data sources in Section 4.1. When the methods are first applied, as described in Section 4.2, the results are scattered and less than promising. After adjusting different aspects of the pipeline, including the matrix creation, as described in Section 4.3, and anchor words selection, as described in Section 4.4, the results seemed much more reasonable. These adjustments are evaluated using cross-validation in Section 4.5

4.1 Data Sources

4.1.1 Online Medical Forums

Data was collected from main autism forums, including Autism Forum asd (2014a), ASD Friendly asd (2014b), Autism Web aut (2014), and Talk about Autism tal (2014). Data was scraped from the subpages of these sites related to general discussion and medical forums using Beautiful Soup on 14 December 2014. The dataset presented includes 245,717 posts from 4 forums for a total of 8 subforums and 33,818 threads. The breakdown of these sources is shown in Table 4.1.

The posts ranged from 1 to 11,317 words in length, with the average length of a post at 108 words. The distribution of post lengths is seen in the Figure 4.1.

4.1.2 Electronic Medical Records

The data from electronic medical records was collected from Boston Children's Hospital (BCH) for all patients with records with at least one code for ASD (299.*). Because this

Forum	Subforum	Threads
ASD Friendly	Medical & Health	1902
ASD Forum	General Discussion	270
	Education	120
	Help and Advice	121
	ASD Related Conditions	29
	PDA	18
	ADD & ADHD	11
Autism Web	Treatments for Autism	16535
	Popular Posts	14000
Talk About Autism	Medical and Health	435
	Diagnosis	377

Table 4.1: Forum, subforum, and thread breakdown for discussions on autism-related online medical forums.

method tracks only mentions of autism as opposed to confirmed diagnoses, this could have included people being tested but did not have ASD in the end. Notes were processed by the BCH informatics and natural language processing using CTAKES, extracting UMLS concepts from the notes and whether these concepts were positive, negative (not concept), or history. This resulted in two sets of data: one that processes only ICD9 codes and the other which includes all UMLS concept matches that physician notes down.

For this pipeline, the record data was filtered on using only concepts that were referred to in the positive concept (ie, excluding negative phrases such as "not autism"), records for patients that were at least 15 years old and with records of at least 4 visits.

4.2 Initial Results

The pipeline, as described in Section 3.2, with no embellishments, was first run on the online medical forum data set, using L2 loss and 20, 50, and 100 potential anchors. This process was run to only include anchor words that appear greater than 100 times in the documents and the result displays each of the anchor words and the top 10 words associated with each anchor word and topic.

It was found that these topics did not at all cohesive, with some of the topics resulting in a strange conglomeration in the supposedly related words. For example, Table 4.2 presents the top topic in the 20 anchor word set trial with anchor word "acute myocardial infarction," we see that there are words that are not related to symptoms and behaviors and words that are not related to each other.

Looking at these results qualitatively, there are some problems that arise, including:

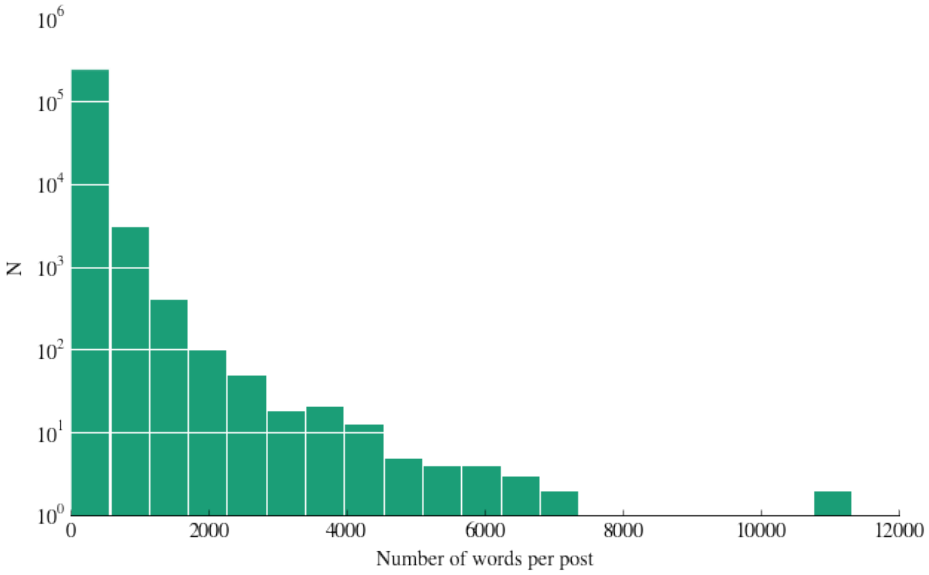


Figure 4.1: Distribution of posts lengths scraped from autism-related online medical forums.

Anchor word	Top words
acute myocardial infarction	internets, United States, vision, all, work, happiness, ns (qualifier value), sound - physical agent, will not try (finding), magnesium

Table 4.2: Example of anchor word and associated top words in initial findings applying basic pipeline of medical concept extraction and anchor word driven topic modeling to posts scraped from online medical forums.

- **Relevance.** The concepts that appear in the result, both with regards to the anchor words and the top words, do not strongly relate to behaviors and characteristics associated with autism, or the topic at hand. In this example, this includes words like "suggestion" or "mail". While one of the benefits of the CHV is its comprehensiveness, we hard to reduce the search space to words that are symptom- or behavior-related, as explained in Section 4.3.1.

  Additionally, it is interesting that "internet" appears as one of the top related words, with 10 of the topics in the 20 anchor word run have the word internet as one of their top words. It turns out that the letter combination of "http" matches with the "internet" concept, meaning every time a user posts a link, the internet concept will be highlighted, artificially inflating the actual importance of the concept among the topics. Removal of stop words like "http" is explored in Section 4.3.3.

- **Coherence.** Not only were the topics not relevant to the question at hand, but they were also not discernibly related to each other within a given anchor word or the top

words associated with it in the topic. For example, there is a tenuous relationship between "acute myocardial infarction" and "vision", but that relationship is even weaker when including "happiness" and "magnesium", as proposed by the sample results. Two hypotheses for this include the method of anchor word selection, as explored in Section 4.4, as well as the top word probabilities, seen in Section 4.3.2.

However, while the first example of the topics and topwords do not necessarily seem promising, a quick qualitative assessment found that there were some relatively interesting and somewhat cohesive topics, some of which are included in Table 4.3.

Anchor word	Top words
whey protein	whey protein; proteins; caseins; out (qualifier value); food; poaceae; free of (attribute); acids; milk; glutathione
antimicrobials	tic disorder; probiotics; antimicrobials; testing; saccharomyces cerevisiae; bacteria; ring, device (physical object); enzymes; antibiotics; hydroxymethylglutaryl-coa reductase inhibitors
ascorbic acid	ascorbic acid ; vitamin A ; internet ; minerals ; vitamin D ; vitamin B complex ; vitamin E ; oils ; all ; out (qualifier value)

Table 4.3: Examples of more promising initial findings from the application of the basic pipeline on online medical forum data, the procedure used to generate Table 4.2.

These are examples of three topics that seemed more viable. While there are still some stop words included, such as "all" or "out" as the final words of the topic with anchor word "ascorbic acid", these results demonstrate both **relevance** to autism and **cohesiveness** amongst themselves as a given theme or topic, which are necessary for informative results. The existence of these topics encourages the continuation of this research project and presents the promise for these methods as a great starting point.

For example, a casein-free diet is a trend for combating some of the negative side effects of autism, and this is embodied in the "whey protein" topic. Moreover, "antimicrobials" and "probiotics" are a new trend in autism treatment, especially with regards to the microbiome and the potential for regulation of gut bacteria in the reduction of some symptoms. In addition to these examples of relevance, the final example topic with anchor word "ascorbic acid," which is strongly linked colloquially to Vitamin C, also demonstrates stronger cohesion, as suggested by the grouping of various vitamins and minerals within the same topic.

4.3 Matrix Creation

One of the potential areas for improvement in the basic approach is the creation of the concept-concept co-occurrence matrix. A basic approach to the creation of this matrix will take the naive count of all letter combinations that match with any concept in the CHV. However, the preliminary results proves this wide search space is much too large, necessitating some reasonable paring down of the possibilities. The following section explores pruning of the CHV categories (Section 4.3.1), probabilistic display of the top words (Section 4.3.2), and removal of stop words (Section 4.3.3) as various means of reducing the search space.

4.3.1 CHV Trie Pruning

The full CHV includes a wide range of concepts. The metathesaurus of the UMLS lists various categories under which each term can be found, ranging from "Anatomical Abnormality (A1.2.2)" to "Daily or Recreational Activity (B1.2)." For a full list of topics covered by the UMLS as well as a highlighting of the topics included in each of these individual explorations with pruning, see Appendix 6.1.

In order to narrow down the search space for the concept mapping algorithm and to reduce the space complexity of the trie, the concepts from the CHV were pruned to include only metathesaurus topics related to the analysis at hand. Below are a few of the higher level categories that were tested. These include looking only at words associated with behaviors and symptoms as well as diagnoses.

Behaviors, symptoms, diagnoses, and diet

The first filter was applied to include categories and the associated subcategories that were related to behaviors, symptoms, diagnoses, and diet. The reasoning behind selecting behaviors, symptoms, and diagnoses was to limit the discussions to concepts that were more directly related to the the main question at hand. And the reasoning behind including the words associated with diet is because of the strong relevance and cohesiveness exhibited by those associated words in the preliminary round of results, as seen in Table 4.3.

Table 4.4 includes a few of the anchor words and top words found in this exploration. An interesting finding is that with the reduction of the search space to include only more tightly related words, the topics that emerge are much more relevant. For example, the topics with anchor words "inappropriate behavior" and "speech and language" are much more relevant to and even expected to rise out of discussions of the manifestations of autism. Some of the words that appear in the top words list are also quite related to each other, including the

Anchor word	Top words
inappropriate behavior	vision ; butting (finding) ; aortic valve stenosis ; thinking ; chronic fatigue syndrome ; obsessive-compulsive disorder ; tic disorder ; deterioration of status ; autistic disorder ; liking
speech and language	vision ; autistic disorder ; therapeutic procedure ; liking ; butting (finding) ; speaking, function (observable entity) ; comprehension ; thinking ; seizures ; speech delay
seizure activity	aortic valve stenosis ; chronic fatigue syndrome ; seizure activity ; electroencephalography ; autistic disorder ; epilepsy ; speech ; reading ; epilepsy, absence ; sleep
hay fever	food hypersensitivity ; aortic valve stenosis ; thinking ; hay fever ; chronic fatigue syndrome ; behavior ; allergy testing ; asthma ; intolerance, function (observable entity) ; deterioration of status

Table 4.4: Example of anchor words and top words findings with a filter to include only CHV concepts associated with behaviors, symptoms, diagnoses, and diet.

link between the anchor word "speech and language" and the top words associated with it of "therapeutic procedure" and "comprehension."

While "autistic disorder" does show up in the top words related to "speech and language," the concept of "autism" or any directly related concepts do not show up many times throughout the anchor words and top words. This is perhaps due to the underlying nature of the discussion with autism being an implicit topic of discussion as opposed to explicitly state in the text. But more likely, this is a sign that more processing needs to be done on the topics.

Furthermore, the anchor word "seizure activity" seems to not only have a strong cohesive presence, including words like "epilepsy" and "electroencephalography," but also an interesting contribution in the ability to the potential of investigating more comorbid disorders. The entry with the anchor word "hay fever," which is interestingly fairly coherent, could also be a result of comorbid disorders or the frequency and recognition of cold-like symptoms in general discussions about diseases and diagnoses by lay persons.

Behaviors, symptoms, diagnoses

After realizing that many of the topics included were related to diet and nutrition, the topics were focused to include only the behaviors and characteristics associated with autism that could potentially inform diagnosis of autism and associated or potential co-morbid disorders.

Diagnoses or ICD-9

The electronic medical record data includes filtering to only include ICD-9 codes, or only the diagnoses. In order to examine the differences between the online medical forum data and the electronic health record data, a filter for only diagnoses was placed on the online medical forum data and the results of the pipeline run on both sources of data is compared.

Anchor word	Top words
diaper rash	eczema ; diaper rash ; tinea ; miliaria ; viral exanthem (disorder) ; urticaria ; scratches (disorder) ; feeling upset (finding) ; dumping ; rash mouth
cold symptoms	chronic fatigue syndrome ; cold symptoms ; aortic valve insufficiency ; pulmonary embolism ; influenza ; pharyngitis ; abstract thought (finding) ; asthma ; polyendocrinopathies, autoimmune ; migraine
streptococcal infections	infection of ear (disorder) ; streptococcal infections ; chronic fatigue syndrome ; bacterial infections ; yeast infection ; virus diseases ; urinary tract infection ; chronic infectious disease (disorder) ; disease ; tic disorder

Table 4.5: Example of anchor words and top words findings with a filter to include only CHV concepts associated with only diagnoses.

Table 4.5 includes examples of the anchor words and top words found using a diagnosis filter on the online medical forum data. There are some interesting results, with strong coherence within the idea of itches and rashes such as "eczema" and "scratches" embodied in the topic with anchor word "diaper rash." Also it is interesting that the topics "cold symptoms" and "streptococcal infections" both appear, having similar symptoms in reality but having some different associated top words in the findings. Both some strongly cold-related symptoms, respiratory concepts, or general disease-related concepts in the associated top words.

Table 4.8 lists some examples of the findings of the topic model algorithm run on the data from the electronic medical records specifically on only diagnoses in the ICD-9. In this example, the translation of the concepts is slightly different from the concepts in the online medical forums due to the translation from ICD-9 to UMLS compared to CHV, resulting in more granular and specific concept names.

This data shows that the results from the electronic medical records seem to be much more cohesive, resulting in stronger conglomeration of the topics into diagnoses. For example, looking at the diabetes related topic, all of the top words are concepts that are related to diabetes, mainly organized around different types of diabetes (including type I vs II or

juvenile diabetes). This also shows that the data is much more granular, distinguishing between the different types of diabetes, which would be difficult to discern in an online setting. The other topics also seem to be fairly cohesive, including more respiratory terms for the asthma topic and convulsion related terms in the epilepsy topic. These are also fairly relevant to the autism concept.

The fact that the topics that emerge from disease and disorder related entries in the electronic medical records are much more cohesive than the topics from online health forums makes sense. First and foremost, the data for disease and disorder should be much cleaner in hospitals where input could take the form of billing codes and familiar terms to that environment. This difference in language might not necessarily be bridge by the consumer health vocabulary in that individuals will discuss disease more formally in formal settings compared to using slang, colloquialisms, or simply not knowing the proper terms in online settings. Also, clinical settings are more driven by discussions by disease and disorder, rather than the wider span of topics that could be discussed on online medical forums.

4.3.2 Probability

The original implementation of the Arora et al. (2012a) algorithm includes the top n words that are associated with a given anchor word and topic. In the previous cases, as seen in the examples in other sections, the $n = 10$, meaning that each topic will display the top 10 words associated with that given topic. However, with a glance at some of the top words returned by the previous methods, it seems that while there are some words that are well associated with a given anchor word, others may seem less related to the anchor word.

Another approach would be to include only top words that are associated with the anchor words and topics at a given probability threshold. The two probability thresholds that were tested include 0.01 and 0.03. After running the algorithm using both thresholds on different types of filters and data sources, the threshold 0.01 seems the more reasonable because it maintains a relatively consistent number of top words associated with each topic.

Table 4.6 presents some examples of the anchor words and top words that emerge using a more discerning display of the top words. Comparing these results to those seen in Table 4.5, there are fewer top words associated for each. However, they seem to be, qualitatively, more related to the anchor words. For example, in the topic with the anchor word "diaper rash," the term "dumping," which is much less relevant is no longer included. The term "rash mouth," which is related to the concept of rash but potentially not diaper rash, is also removed.

Anchor word	Top words
diaper rash	eczema ; diaper rash ; tinea ; miliaria ; viral exanthem (disorder) ; urticaria ; scratches (disorder)
streptococcal infections	infection of ear (disorder) ; streptococcal infections ; chronic fatigue syndrome ; bacterial infections ; yeast infection ; virus diseases ; chronic infectious disease (disorder) ; disease
cold symptoms	chronic fatigue syndrome ; cold symptoms ; aortic valve insufficiency ; influenza ; pulmonary embolism ; pharyngitis ; abstract thought (finding)

Table 4.6: Examples of anchor words and top words with probability > 0.01, with filtering for only diagnoses.

4.3.3 Stop Words

Since the consumer health vocabulary is very extensive, some combinations of letters matched with concepts that were not intended due to common misspellings. For example, "te" matches with the concept associated with the element Telluride, causing it to be counted as a concept match more commonly than it actually appears in the text. Another example is the concept "butting", which is matched in the CHV trie with the combination of letters "but", a common stop word but not necessarily common as the concept of "butting".

Therefore, *stop words*, or some of the most common words including short function and lexical words should be filtered out before processing to avoid unnecessary matching. The list of stop words used in this report is drawn from Arora et al. (2012a) and includes 525 words as seen in Appendix 6.2.

Anchor word	Top words
autistic disorder	autistic disorder ; parents ; work ; recommending
gastrointestinal tract structure	gut ; gastrointestinal tract structure ; liver ; intestines ; inflammation ; blood

Table 4.7: Examples of anchor words and top words with probability > 0.01, including diagnoses, behaviors, and symptoms, and stop words filtering.

Table 4.7 shows examples of anchor words and the associated topwords with probability > 0.01 for diagnoses, behaviors, and symptoms with stop words filtering. In this case, the topics definitely seem much more cohesive.

4.3.4 Summary

Since the initial results were less than promising, one approach to getting more relevant and coherent results is to adjust the way in which the matrix is created.

The first attempts include pruning the CHV trie (Section 4.3.1), which reduces the search space greatly to only include relevant words. The reductions to include only concepts related to behaviors, symptoms, diagnoses, and diet was promising in terms of relevance for the topics related to diagnoses and in terms of coherence for topics related to diet. Another filter applied to the online medical forum data only include diagnoses was compared against the electronic health records that included only ICD-9 codes and it was found that both show increases in relevance but the results from the electronic health records was much more coherent and granular in the discussion of diagnoses and diseases.

Another approach includes reducing the number of top words that are displayed to include only those with higher probability (Section 4.3.2). This demonstrated a tightening in the top words that are displayed. And the final approach includes removing stop words from the search space (Section 4.3.3).

In the end, the resulting anchor words and top words presented became much tighter, both more relevant and more coherent. Future analysis could include slightly varying the number of words that are included in each filter, as seen in Figure 6.1. For example, closer analysis would deem some of the categories, such as "B2.2.1.2.1.2 Neoplastic Process" under the heading of cisease diagnostics as less relevant and a more detailed trimming of the categories could help improve the relevance of each of the words. It would also be interesting on the flip side to include analysis on only words related to diet because of its predilection for coherence. Or even only behaviors and symptoms and not diagnoses, to see if the symptoms themselves will organize despite not having the explicit diagnosis concepts included in the analysis. Other approaches could include including only results in which "autism" shows up as a top word, making explicit the assumption that all of the documents are related to autism.

While the top words and anchor words seem much cleaner after this preprocessing, the overall results still seem scattered. This is potentially due to the way in which the anchors are selected. The exploration of this is described in the following section.

Anchor word	Top words
Diabetes mellitus without mention of complication, type I [juvenile type], not s	Diabetes mellitus without mention of complication, type I [juvenile type], not s ; Diabetes mellitus without mention of complication, type II or unspecified type, ; Diabetes mellitus without mention of complication, type II or unspecified type, ; Diabetes with ketoacidosis, type I [juvenile type], uncontrolled ; Diabetes mellitus without mention of complication, type I [juvenile type], uncon ; Diabetes with ketoacidosis, type I [juvenile type], not stated as uncontrolled ; Diabetes with renal manifestations, type I [juvenile type], uncontrolled ; Diabetes with ketoacidosis, type II or unspecified type, not stated as uncontrol ; Diabetes with ophthalmic manifestations, type I [juvenile type], not stated as u ; Other chronic infective otitis externa
Asthma, unspecified type, with (acute) exacerbation	Asthma, unspecified type, with (acute) exacerbation ; Chronic respiratory disease arising in the perinatal period ; Other and unspecified hyperlipidemia ; Other diseases of trachea and bronchus ; Acute bronchiolitis due to respiratory syncytial virus (RSV) ; Other deficiencies of circulating enzymes ; Acute tracheitis without mention of obstruction ; Other specified disorders of liver ; Chronic kidney disease, Stage I ; Influenza with other respiratory manifestations
Generalized nonconvulsive epilepsy, without mention of intractable epilepsy	Generalized nonconvulsive epilepsy, without mention of intractable epilepsy ; Generalized nonconvulsive epilepsy, with intractable epilepsy ; Hypersomnia due to medical condition classified elsewhere ; Cerebral artery occlusion, unspecified with cerebral infarction ; Developmental dislocation of joint, other specified sites ; Mineral deficiency, not elsewhere classified ; Other specified disorders of esophagus ; Other and unspecified conditions of umbilical cord affecting fetus or newborn ; Sleep related movement disorder, unspecified ; Pyogenic arthritis, pelvic region and thigh

Table 4.8: Example of anchor words and top words findings using electronic medical record data that includes only diagnoses.

4.4 Anchor Word Selection

However, regardless of how clean the top words results were based on concept and top word filtering and probability, the anchor words were still sometimes off-kilter. This section describes the methods applied to create more robust anchor word selection in order to improve the results presented by the topic modeling algorithm.

The original approach from Arora et al. (2012a) is a combinatorial algorithm for finding anchor words by iteratively finding the furthest point from the subspace spanned by the anchor words found so far. As explained earlier, this does provide a heuristic for finding the most all-encompassing basis but does not necessarily do so in a robust manner. One option to extending this algorithm include using more repetitions of the greedy selection algorithm given the new bases, which would be able to improve anchor selection but has a theoretical limit to improvement.

We propose an updated procedure that will account for potential variations due to sparse data as well as testing the span of the anchor set, iteratively finding the most inclusive span of anchor words based on a slight variance to each of the points. The main differences between the updated approach and the original approach are:

- **Variance.** In order to account for the potential variance not captured by the sparsely populated data, this approach induces reasonable variations among each of the points.
- **Inclusiveness.** The updated approach uses number of points included as a measure of variance as opposed to simply using the furthest point.

In the following section, we describe the proposed implementation and evaluation of a more robust means of anchor word selection that will account for the sparsity of the matrix. Section 4.4.1 lays out a higher-level description of the resulting algorithm, including the proposed updated algorithm (Algorithm 4) as well as modifications to a more tractable hill-climbing algorithm (Algorithm 5). The establishment of the exact parameters for each of the novel steps for the algorithm are explored, including the point selection (Section 4.4.2), checking of inclusiveness (Section 4.4.3), and finally the calculation and implementation of variance (Section 4.4.4). Section 4.4.5 explores the sensitivity of the number of anchor words on the error tolerance and running time of this approach.

4.4.1 Overview

The first proposed implementation of the new algorithm starts with a full search of the space of possible anchor combinations on a set of test configurations generated with slight variance

and selection of the set that includes the most number of points. As described in Algorithm 4, the algorithm first starts with a greedily selected basis, similar to Arora et al. (2012a). Then, for R repetitions, each anchor is attempted to be replaced by another point that performs, on average, better than the greedily selected anchor set on the N configurations generated using a calculated variance. Then, with all of the proposed anchor sets, including the greedily proposed anchor set and each anchor set that performs better than the greedy algorithm for each anchor removal and replacement, these are tested on M generated configurations and take the anchor set that has the best average performance.

Algorithm 4: FastAnchorWords, Updated

Data: V points $\{d_1, d_2, ..., d_V\}$ in V dimensions, almost simplex, K vertices, $\epsilon > 0$, R repetitions, N and M test configurations.
Result: K points that are close to the vertices of the simplex.
begin
  $S \leftarrow \{d_i\}$ s.t. d_i is the farthest point from the origin;
  for $i = 1$ *to* $K - 1$ **do**
    Let d_j be the point in $\{d_1, d_2, ..., d_v\}$ that has the largest distance to span(S);
    $S \leftarrow S \cup \{d_j\}$;
  end
  $S = \{v'_1, v'_2, ..., v'_K\}$;
  for $r = 1$ *to* R **do**
    for $j = 1$ *to* K **do**
      for $n = 1$ *to* N **do**
        Generate a new configuration based on the variations;
        for $i = 1$ *to* C **do**
          Count the number of points included by the candidate anchor set;
          Store this count;
        end
      end
      Select anchor set including most points average across configs;
      Store this anchor set;
      Replace S with this anchor set;
    end
  end
  for $m = 1$ *to* M **do**
    Generate a configuration;
    for $i = 1$ **do**
      Count the number of points included by the candidate anchor set;
    end
  end
  Select anchor set including most points average across configs;
  return $\{v'_1, v'_2, ..., v'_K\}$
end

However, Algorithm 4 proved to be intractable, the bottleneck originating from each

inclusion count for the upwards of 2,000 points taking approximately 1 minute each, meaning a single iteration of attempting to replace an anchor, with $N = 5$ took upwards of 24 hours, and the whole process would take days. This does not even account for the possibility of having even more robust approaches that increase the N and M test configurations for each potential switch and for each reconfirmation test.

Therefore, a hill-climbing style algorithm was proposed. Instead of attempting to exhaustively search the possibility of replacing each anchor with every point, the algorithm looks to attempt swapping each point out based on a certain selection criteria. As seen in Algorithm 5, the algorithm starts by heuristically selecting an anchor to be replaced and then heuristically selecting an anchor that will replace that given anchor, keeping the proposed swap only if the new set performs better on the N test configurations than the original. The N test configurations are refreshed every $k*$ iterations and the entire process runs for R repetitions.

The main differences of steps in this algorithm are selecting points to swap in and out, described in Section 4.4.2, checking the inclusiveness of the points, described in Section 4.4.3 and inducing variance, described in Section 4.4.4.

Algorithm 5: FastAnchorWords, Updated with Hill Climbing

Data: V points $\{d_1, d_2, ..., d_V\}$ in V dimensions, almost simplex, K vertices, $\epsilon > 0$, R repetitions, N and M test configurations.
Result: K points that are close to the vertices of the simplex.
begin
  $S \leftarrow \{d_i\}$ s.t. d_i is the farthest point from the origin;
  for $i = 1$ *to* $K - 1$ **do**
    Let d_j be the point in $\{d_1, d_2, ..., d_v\}$ that has the largest distance to span(S);
    $S \leftarrow S \cup \{d_j\}$;
  end
  $S = \{v'_1, v'_2, ..., v'_K\}$;
  for $k = 1$ *to* K **do**
    if $k \mod k* == 0$ **then**
      create a set of C new configurations based on variance;
    end
    Select an anchor to swap out proportional to variance;
    Select an anchor to swap in proportional to distance and variance;
    Replace swapped out anchor with swapped in;
    Evaluate the swap by counting how many points are included in this configuration on average;
    Replace S with the swap if it performed better than the previous ;
  end
  return $\{v'_1, v'_2, ..., v'_K\}$
end

4.4.2 Point selection

Heuristics were used in order to select the points to be swapped out and the potential point to be swapped in to the new anchor set. The selection of the points from the original anchor set to be swapped out is proportional to the calculated variance, as described in Section 4.4.4. The reasoning behind this is that points with higher variance would be less stable with the new configurations and therefore more stable points should be preferred over those that are less stable in establishing a strong anchor set.

And the selection of the replacement point is proportional to variance, similar to the selection of points to be swapped out, as well as the average of the distance of the point to the rest of the anchors, a similar calculation to the greedy algorithm used in the Arora et al. (2012a) algorithm. This is to leverage both the element of variance as well as the established heuristic of distance. This algorithm also ensures that there are no repeat sets by selecting from the set difference of the possible points and the tried points to consistently test novel anchor sets.

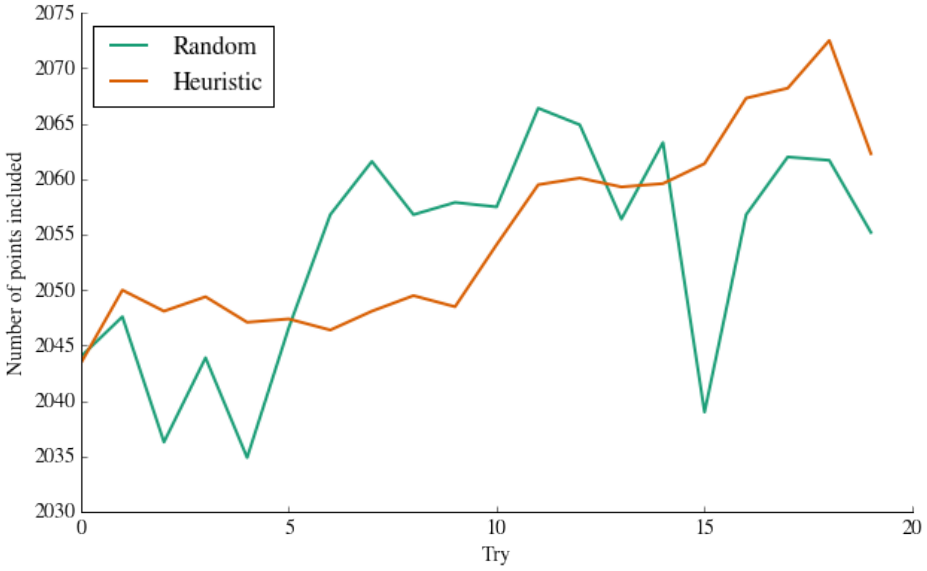


Figure 4.2: Effect of number of attempts on the number of points included on proposed anchor sets by a random and heuristic point selection algorithm.

Based on the initial runs of the algorithm, this point selection algorithm showed some promise, as seen in Figure 4.2. This graph depicts the number of points included on the proposed anchor sets, before deciding whether to retain the proposed anchor set in order to get a sense of whether the point selection algorithm is proposing reasonable anchor sets.

The green line indicates the performance of a proposed anchor set by a semi-random selection of points to be swapped in and out whereas the orange line indicates the use of

this heuristic on selecting the points to be swapped in and out. A qualitative analysis of the curve indicates that there is a positive correlation between the number of tries and the number of points included in proposed anchor set the case of the use of the heuristic.

Since the line for the heuristic tapers off slightly towards the end, this graph also indicates that there is benefit to the increased number of iterations. It does not yet seem like the algorithm has reached a plateau of sorts, therefore, investigation of the number of iterations before the plateau should be established. Reasonable stopping cases for the hill climbing algorithm include stopping at a point where the changes in the algorithm are less than some ϵ or a reasonable number of tries has been exhausted, whichever comes first.

Improvements on the point selection heuristic can be implemented in the form of different proportions, such as changing the amount of weight placed on the proportional variance and distance in selecting points to be swapped in. Other heuristics such as number of neighbors or probability of being located outside a certain radius after variation could also be tested to establish their effects on the anchor set.

4.4.3 Checking Inclusiveness

This solution uses the heuristic of the number of points included in the created convex hull in order to determine which anchor sets are better than others. This is in contrast to the use of greedily selected anchors from the original algorithm in that this approach will actually test the changing of the scope of the anchors.

In order to check which of the potential sets of anchors included the most number of points, it was necessary to check the number of points from a given configuration that are included within the convex hull created by that set of anchors. This can be done by finding the distance of the closest point within the convex hull to the given point. If the point is within the convex hull, the distance should be close to 0. Points within the convex hull can be formed as a linear combination of each one of the anchor vectors, meaning this problem can be formulated as a linear program.

The linear program has the objective of minimizing the distance between the point in the convex hull to the given point, or $\min ||Bx - P||$, where B is the set of anchor vectors, P is the point, and x is the set of coefficients for the linear combination. The constraints are that the coefficients are nonnegative and they sum up to 1, or $\sum x = 1$ and $x \geq 0$.

The linear program returns the error the attempted fit of the anchor set as a basis to the given data point, as seen in the equation $||Bx - P||$. The two proposed methods of handling this error include setting a threshold to some small epsilon and only counting points that have an error below that threshold or looking at the average error among the points tested.

In the end the heuristic that is used is average error.

Threshold

Originally, the approach was to look at simply the number of errors that were less than a given small epsilon value. Based on tests of the Scipy Optimize linear program solver, an epsilon of $\epsilon = 10^{-7}$ was set. This is the method seen in calculating the number of points included in Figure 4.2. However, upon closer examination of test cases, an artificially inflated number of points were included in each of the anchor sets, up to the point of including all the points in the set, which was possible but unlikely. Since the means of implementing variance had a significant effect on the range of errors produced, there would also be cases of the other extreme, where only the anchor points themselves were included in the span of the anchor set.

Average error

However, visualizing the spread of the errors for some of the attempts and the differences in the different approaches for inducing variance as described in the following section and seen in Figure 4.5 prompted to exploration of alternative methods to determining the error than a simple threshold.

Looking at the errors, most form reasonable distributions. Some of the parameters that may differ for each run include the mean and the variance of the curves. In the context of the problem at hand, it makes sense to select anchor sets that perform well in terms of the mean and take less into consideration the spread of the distribution of the errors. Therefore, the heuristic used is the mean errors.

4.4.4 Inducing Variance

Since there are few instances of each concept in each document, it is difficult to ascertain whether there is slight variation than the points captured in the data. Therefore, we introduce slight shifts among each of the points to simulate the unaccounted variance. First, the method of calculating variance is described and then the different methods that were explored of implementing this variance are described.

Calculating variance

The amount of variance the anchors would have is calculated using the following equations where $var_{i,i}^{(1)}$ or $var_{i,j}^{(1)}$ the variance of concept i with itself or with concept j, respectively, is

calculated with the variables p_i and p_j, the probabilities of word i or word j appearing in a given post, and n, the average length of the post.

$$var_{i,i}^{(1)} = n(-2p_i^2 + 8 + p_i^3 - 6p_i^4) + n^3(4p_i^3 - 4p_i^4) + n^2(2p_i^2 - 12p_i^3 + 10p_i^4) \tag{4.1}$$

$$\begin{aligned} var_{i,j}^{(1)} = n^3(p_i^2 p_j + p_i p_j^2 - 4p_i^2 p_j^2) + n^2(p_i p_j + 10p_i^2 p_j^2 - 3(p_i^2 p_j + p_i p_j^2)) \\ +n(-p_i p_j - 6p_i^2 p_j^2 + 2(p_i^2 p_j + p_i p_j^2)) \end{aligned} \tag{4.2}$$

These are derived from the underlying distribution assumption for topic modeling, that for a single document with length n, word count x which is drawn from some distribution $p = Aw$ for some w $Dir(\alpha)$:

$$\mathbb{E}[x^T x - diag(x)|w] = n(n-1) \cdot pp^T \tag{4.3}$$

Implementing variance

In order to implement this level of variance, a few heuristics were used. The original heuristic was to induce variance for each element in the concept i-concept jth-entry, $p_{i,j}$ in the Q matrix based on variance, $var_{i,j}$, as described, by using the equation:

$$p'_{i,j} = 2(var_{i,j} \times rand(0,1)) - var_{i,j} + p_{i,j} \tag{4.4}$$

which effectively selects adds a random number between $[-var_{i,j}, var_{i,j}]$.

However, implementing only this naive equation proved to be an issue for values that were on a boundary, such as those whose probabilities are initially 0 and if varied would result in a negative number. Various strategies to help to induce some variance but within viable boundaries, including the following:

- **Border.** Set all probabilities that were shifted negative to 0.
- **Scale.** Scale all probabilities so that the lowest value would be set to 0.
- **Log.** Use the log of the variance and exponentiate the entire equation but first setting the values with 0 probability to a small number, some epsilon, ϵ.
- **Dirichlet sampling.** Sample a set of points based on the the Dirichlet distribution, the basis for the Latent Dirichlet Allocation in topic modeling.

Each of these strategies are tested on data created using the Q-matrix filtered to exclude stopwords and to include only concepts related to behaviors, symptoms, and diagnoses. Initial testing was done on the number of anchors $K = 20$ in order to be more tractable and used 10 configurations with the configurations refreshed every $k* = 5$ on 30 tries. The merits and the downfalls of each of these strategies is described in the following sections. While scaled seemed like the most viable originally, ultimately, the Dirchlet sampling approach initialized to probabilities proved to be the most reasonable.

Border

By setting all probabilities that became negative to zero effectively creates a "border", bouncing all those that fell below the threshold to the threshold. However, since on average approximately 30% of the points in the new confirmation were below 0, a significant number of elements would be on the border and this approach would be less successful than necessary.

Scale

This method takes the configuration generated by the naive method described before and effectively shifts the points by the difference between the smallest point in the new configuration and 0.

Upon first glance, this method seems to be the most viable option because of its low errors, as seen in Figure 4.5a. This figure shows the error distributions of 30 anchor sets tested on 6 different configurations.

However, one of the concerns is the possibility of reducing the differences between each of the values as all were scaled equally. This would mostly take place if there is a gap between the minimum point and 0 that is not proportional to the maximum point, completely shifting the scale.

trial	1	2	3	4	5	6	7	8	9	10
min	-12.31	-0.22	-0.32	-14.35	-6.24	-0.33	-11.70	-0.65	-0.32	-0.74

Table 4.9: Minimum values or amount scaled for 10 sample configurations using the scaled approach.

Upon closer examination, the original data of the Q-matrix had a range of points from 0 to approximately 0.55. However, looking at the minimum points in the 10 new configurations created using variance, as seen in Table 4.9, while some of the minimums are reasonable, closer to -0.22, some configurations result in minimum points that are up to -14.35. This

large value would result in shifts that move the points a significant distance from their original starting point.

These could explain the extreme variances in the differences in the errors seen in Figure 4.5a, which had previously been attributed to whether the anchors were successful or not. In examining the errors of the other approaches to inducing variance, the shapes of the errors seem to be somewhat consistent, making these large variations an anomaly.

Log

Taking the log of the variance and exponentiating the equation takes the form of the following equation:

$$\exp(2(\log(var_{i,j}) \times rand(0,1)) - \log(var_{i,j}) + p_{i,j}) \tag{4.5}$$

where $var_{i,j}$ is the variance, as calculated with the equation described previously and probability $p_{i,j}$ is the probability the two concepts appear in the same post, once again randomly selecting a point within a range but ensuring th at the resulting values would be nonnegative.

It was found that simply taking the log of the probabilities was not feasible because of the nature of the log of 0, so first it was necessary to set values with 0 probability to a small epsilon. For this project, the epsilon selected is $\epsilon = 10^{-7}$.

After applying this change to the calculations, the log approach seems to be tractable, but to have fairly high errors with a wide range, as seen in Figure 4.5b. For that reason, this is not the preferred approach.

Dirichlet sampling.

A configuration of points is sampled using a Dirichlet distribution, which is the underlying distribution used in topic modeling. Dirichlet distributions $Dir(\alpha)$ often take in a single parameter, α, which can be influenced by some concentration parameter, or σ. In this case, the α used is the original probability from the Q-matrix, or the calculated $p_{i,j}$ probability but the concentration, or σ can start with a simpler and more generic equalized amount, or the inverse of the size of the row, or $\frac{1}{|V|}$ or either the column probability, $\sum_j p_{i,j}$.

- $Dir(p_{i,j} + \frac{1}{|V|})$. This calculation of the Dirichlet distribution *initialized equally* uses a concentration parameter that is equal for each α, essentially the inverse of the length of the row, or $|V|$.

- $Dir(p_{i,j}+\sum_j p_{i,j})$. This calculation of the Dirichlet distribution *initialized by probability* is sensitive to both the probability of each of the elements in the rows but also the probability of each of these in the column form.

Looking at Figure 4.5c and Figure 4.5d for the sampling initialized by equally and by probability, respectively, both approaches seem to have similar range of errors but fairly different means. It seems like the sampling initialized by probability has a much lower mean, within a more reasonable 10^{-3} range compared to the 10^{-1} range for the other. This is probably due to the sensitivity of the approach to the distribution of both the row and column probabilities calculated in the original Q-matrix.

Therefore, it is because of its small range and low mean that the Dirichlet sampling initialized by probability is the most viable option for the induction of variance. Future work can look at potentially varying the initialization of the Dirichlet parameters in a different manner, perhaps consider variance in addition probability.

4.4.5 Number of Anchors

Since all of the previous tests were run on the number of anchors $K = 20$ for tractability, analysis of the sensitivity of the number of anchors on the distribution of averages errors runs on varying number of anchors, or $K = 5, 20, 50, 100$ to see how average error changes. The results of the algorithm run on the test data using the Dirichlet sampling initialized to probability for 10 trial anchor sets each. It was hypothesized that the greater the number of anchors, the lower error there would be for the anchor set on inclusiveness.

Figure 4.3 presents the distribution of average errors for each run using the different numbers of anchors. This confirms the hypothesis, that the average errors decreases with increased number of anchors. It also demonstrates the the distribution of errors becomes less varied with the increased number of anchors.

However, one main concern in terms of the algorithm is the tractability of the algorithm in processing multiple anchors. Figure 4.4 highlights the nature of the amount of time it takes to process a larger number of anchors. In the case of 100 anchors, the time to process increases to 302,781 seconds, which is equivalent to 84 hours. This is not very tractable, as these included only 10 anchor set trials and future trials would need to include more trials, even with the reduced range of errors.

There is a tradeoff to be considered. While 5 anchors takes only 801 seconds to run for 10 trials, or a handful of minutes, this does not mean that it is accurate. Another consideration with the number of anchors is also the ability for the number of anchors to properly include

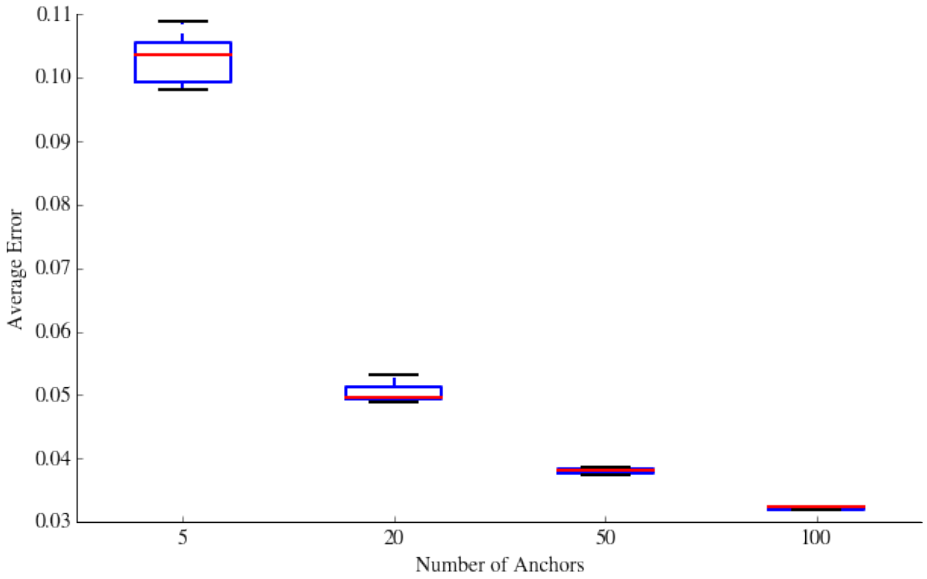


Figure 4.3: Distribution of average errors for each run using different numbers of anchors.

all of the points. Measures other than error should be taken into consideration in future iterations in order to best measure the effectiveness of the number of anchors and their span.

4.4.6 Summary

This section explores a more robust anchor selection approach in order to account for the high variance in the unstructured text of the online medical forum. The new approach differs from the original Arora et al. (2012a) approach in that it considers both variance and explicitly accounts for inclusiveness. An overview of the algorithm is presented in Section 4.4.1 with a description of a basic proposed algorithm (Algorithm 4) and the final implemented hill-climbing variant to increase tractability (Algorithm 5).

The considerations in the heuristic for point selection are explored in Section 4.4.2, including a comparison to a semi-random point selection. Point selection is based on variance and distance, metrics whose proportions and values can be adjusted in future iterations.

Then, Section 4.4.3 presents the methods for checking inclusiveness, which all rely on the linear program that has the goal of minimizing the distance from the point to a point within the span of the basis. After exploring the distribution of errors, the use of average error is selected over the use of a threshold. However, checking inclusiveness of each point is costly. Future approaches could sample and check the inclusiveness of certain points in order to increase tractability and reduce the amount of time it takes for the entire algorithm to run. Other faster heuristics or strategies to checking inclusiveness of larger cohorts of points at a time in a convex hull can also be implemented.

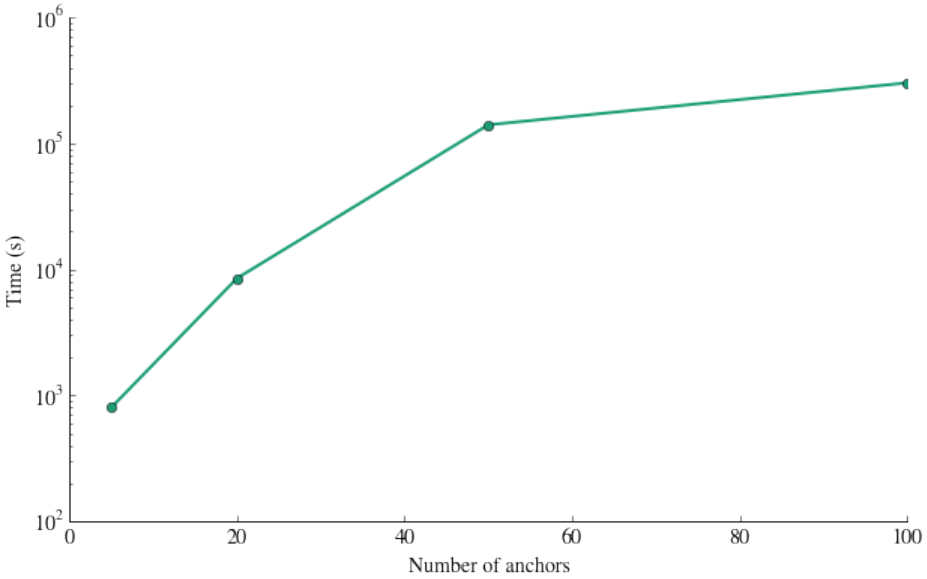


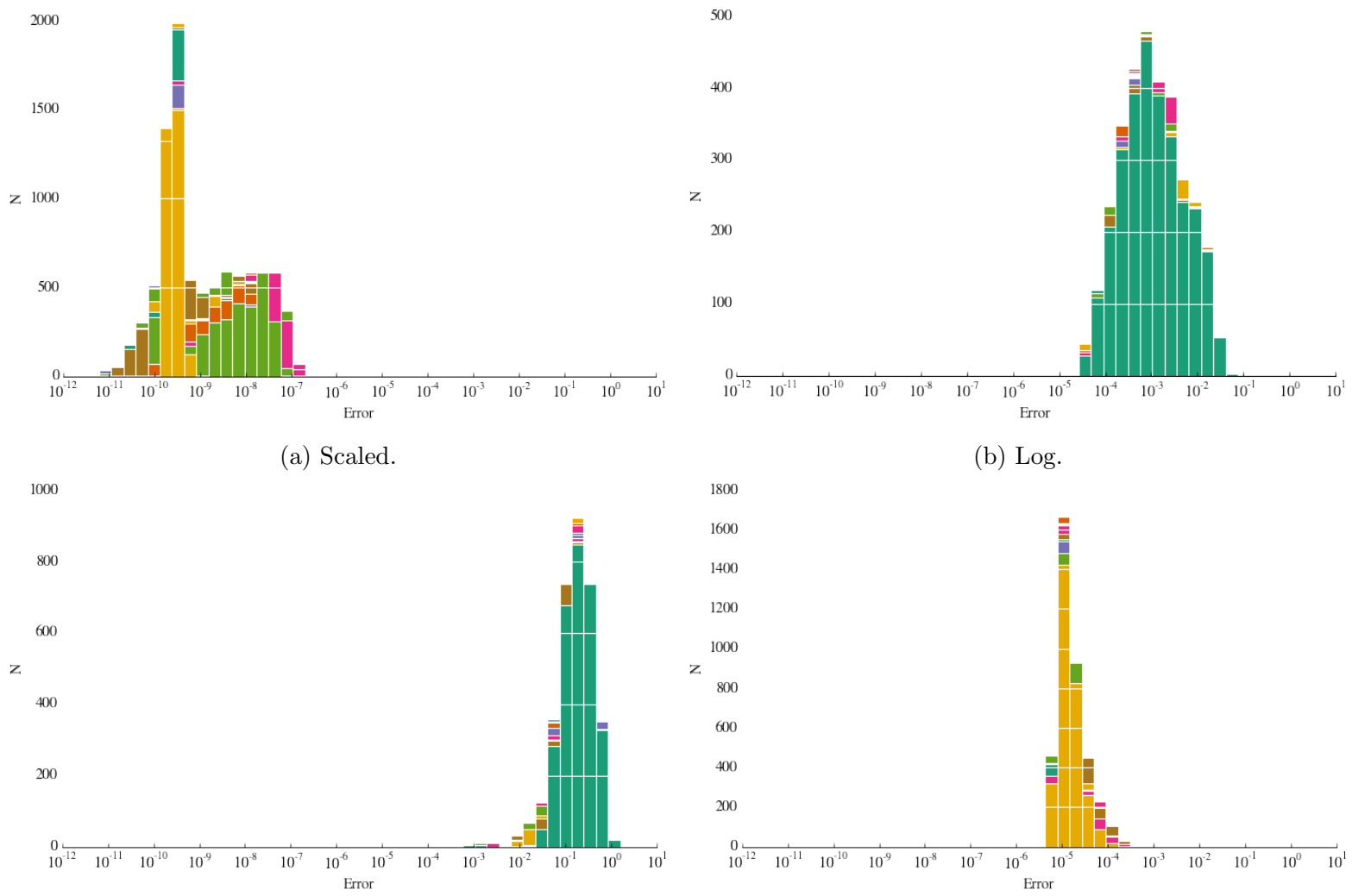
Figure 4.4: Effect of number of anchors on the amount of time it takes to run the algorithm.

Section 4.4.4 looks at the calculation and implementation of variance to create equivalent configurations on which to test the anchor sets. The error distribution of various implementation methods are depicted in Figure 4.5. While naive, scaled, and log implementations of variance seem promising, the "borders" of potential negative errors creates an issue in implementation. The Dirichlet initialized to probabilities as seen in the original Q-matrix the best choice for consistency and lowest mean and smaller range. Future exploration of other parameters for the Dirichlet distribution needs to be done as well as a more exhaustive search of the different types of potential heuristics that improve performance.

Finally, Section 4.4.5 explores the influence of different numbers of anchors, exploring the error (Figure 4.3) and time (Figure 4.4) tradeoff, with much room to explore the effectiveness of differing numbers of anchors in the future.

In addition to the future work mentioned within each section for improvements in performance, there is still much to be done. This section presents the contributions of the exploration of different implementations and parameters on the online medical forum data. Future evaluation includes generated data to establish consistency and accuracy for external validity as well as applications to the other data set, the electronic medical record data.

However, this modified algorithm proves that with more evaluation, it has the promise for demonstrating a more robust means of selecting anchors in an anchor-word driven topic model algorithm. This is especially important for settings in which there is potential for much variation, as seen in the instance here. Future steps could include the application of this modified algorithm on other forums of unstructured text, especially from online sources.



(a) Scaled.

(b) Log.

(c) Dirichlet (initialized equally).

(d) Dirichlet (initialized by probability).

Figure 4.5: Error distributions of 30 anchor sets, each of which is a different color, tested on 6 different configurations using the 4 different approaches as described in Section 4.4.4.

4.5 Evaluation

Given the time constraints of the thesis, the following chapter will be a description of the process that should be taken for evaluation.

Evaluation of the pipeline is a 10-fold cross-validation and comparison of the resulting topics with regards to both the strength of the new anchor words recovery algorithm as well as the differences between the two data sets of online medical forums and electronic medical records.

4.5.1 Evaluation Process

The evaluation is based on a 10-fold cross evaluation separately on the online medical forum posts and the electronic medical records. For each fold, the training data is used to create the concept co-occurrence matrix Q to choose anchors based on the original Arora et al. (2012a) algorithm to produce A and then the updated robust anchor word selection algorithm to produce A'. Then, another concept co-occurrence matrix, Q' is produced. Using held-out probability and L1 distance, the topic similarity between A and A' is calculated based on the test data Q'. This algorithm is described in Algorithm 6.

Based on the sets of all of the $\{A\}$ and $\{A'\}$ between the folds and across the different data sets, e compare the variation between the sets of $\{A\}$'s and $\{A'\}$'s for each. Given that the new algorithm performs as well as and better than the other algorithms, we also qualitatively analyze the differences in topics that are generated by looking at $\{A'_{emr}\}$ and $\{A'_{forum}\}$.

Algorithm 6: 10-fold cross validation evaluation algorithm.

```
Data: Data D
Result: Similarity check (L1 difference, Hungarian bipartite matching, etc.)
begin
    folds = split D into 10-folds for cross validation ;
    for train, test in folds do
        Q_train = build up Q-matrix based on training data ;
        A = choose anchors for Q_train based on Arora et al. (2012a) ;
        A' = choose anchors for Q_train based on updated algorithm ;
        Q_test = build up Q-matrix based on test data ;
        diff = get L1 difference between A and A' ;
    end
end
```

Chapter 5

Conclusion

5.1 Conclusion

This report explores a the discussion of autism in online medical forums compared to electronic medical health records using topic models. We scrape posts from top online medical forums related to autism and use an anchor word driven topic model algorithm to uncover the underlying themes in the structure of the various posts based on words and phrases that match with concepts from the consumer health vocabulary. In the process, we optimize the workflow to narrow the search space by adjusting the creation of the concept-concept matrix and respond accordingly to the sparsity of the data by proposing a more robust means of searching for anchor words. This workflow is then applied to electronic medical records and the elicited topics compared to those found in the online medical forums.

5.2 Future Work

This work makes the first steps towards the application of machine learning methods to understanding the discussions of autism in online medical forums and electronic health records. However, there is still much work to be done before this research realizes its full potential.

As explored in the individual sections, adjustments to some of the facets of the approach such as the exact categories included in the consumer health vocabulary filter or the distribution used to initialized the Dirichlet distribution for inducing variance could find improvements in performance in terms of reduction of time or error. Other examples of this include the number of anchor words used, the number of iterations the algorithm is run, or the various thresholds used to set a lower limit on the number of concepts per document or an upper limit on the amount of error tolerance.

Additionally, after determining the most optimal parameters to be used for the approach, more work needs to be done on the evaluation of the approach. This is the most logical next step, involving the application of this new approach compared to the original Arora et al. (2012a) anchor words driven topic model approach through the cross-validation as described in Section 4.5. This would be useful in understanding the performance of the two algorithms in contrast to each other as well as the differences in the concepts elicited by both algorithms.

Furthermore, additional testing of the algorithm on generated data sets could help solidify the algorithm for external validity. Understanding the nuances of the parameters set for each run of the algorithm could help determine the adjustments necessary for application to different data sets. This could help extend the application of methods to other data sets with high potential variance, including data from other online sources, such as other social media with even higher tendency for misspellings and abbreviations. And these sources can help explore both autism and as well as other puzzles that have low certainty but high popularity.

In an even broader sense, it would be interesting to analyze how this new, proposed approach fares compared to other techniques in terms of running time, information novelty, and error threshold. These additional techniques could be something like probabilistic latent semantic analysis, a more clustering driven algorithm, or deep learning and enhanced neural networks. It could be interesting to apply a variety of algorithms to this same data set or an extension to include more forum posts from different times or more diverse set of data from different sources to establish the strengths and weakness of each approach.

And finally, this research and research like it can help uncover more understanding of the enigma of autism. To fully explore the clinical and policy implications of this research, the due diligence explained above needs to be completed as well as possibilities for extensions of the analysis. It would be interesting to integrate even more modes of data to compare and contrast the results as well as obtain a more comprehensive knowledge of autism. This could simply raise ideas that were not previously explored by scientists and clinicians to investigate as potential causes or associated symptoms. Additionally, it could also be interesting to use these findings to create a multimodal predictive and analytic model to supplement the autism diagnostic process.

Chapter 6

Appendix

6.1 Code

The code and instructions for running this pipeline can be found at `https://github.com/jming/thesis` written with Python using the numpy, pandas, matplotlib, and scikit-learn libraries and is built upon by code from Sam Wiseman, Hongyao Ma, and Yoni Halpern.

Some notable components of this code include the following:

- **Posts.** This includes both the code for scraping the forums as well as compressed versions of the raw sources scraped from forums.
- **Exploration.** This includes the code used in data exploration and the visualizations generated for these explorations and for the paper.
- **Pipeline.** This includes the code that can be used to run the pipeline. Instructions for running the pipeline are included as well as sample `pipeline.sh`.
- **Results.** The results section of the code includes the results of running the algorithm on both the online medical forum and electronic medical records. The textfile `result.txt` explains the parameters used to generate each result. The files are named with the number of anchors and files with `*.topwords.translate` are the files that include the anchor words and top words translated from CUIs.

Any questions or comments can be directed to `joy.c.ming@gmail.com`.

6.2 Additional Figures

A Entity
A1 Physical Object
A1.1 Organism
A1.1.1 Archaeon
A1.1.2 Bacterium
A1.1.3 Eukaryote
A1.1.3.1 Animal
A1.1.3.1.1 Vertebrate
A1.1.3.1.1.1 Amphibian
A1.1.3.1.1.2 Bird
A1.1.3.1.1.3 Fish
A1.1.3.1.1.4 Mammal
A1.1.3.1.1.4.1 Human
A1.1.3.1.1.5 Reptile
A1.1.3.2 Fungus
A1.1.3.3 Plant
A1.1.4 Virus
A1.2 Anatomical Structure
A1.2.1 Embryonic Structure
A1.2.2 Anatomical Abnormality
A1.2.2.1 Congenital Abnormality
A1.2.2.2 Acquired Abnormality
A1.2.3 Fully Formed Anatomical Structure
A1.2.3.1 Body Part
A1.2.3.2 Tissue
A1.2.3.3 Cell
A1.2.3.4 Cell Component
A1.2.3.5 Gene or Genome
A1.3 Manufactured Object
A1.3.1 Medical Device
A1.3.1.1 Drug Delivery Device
A1.3.2 Research Device
A1.3.3 Clinical Drug
A1.4 Substance
A1.4.1 Chemical
A1.4.1.1 Chemical Viewed Functionally
A1.4.1.1.1 Pharmacologic Substance
A1.4.1.1.1.1 Antibiotic
A1.4.1.1.2 Biomedical or Dental Material
A1.4.1.1.3 Biologically Active Substance
A1.4.1.1.3.1 Neuroreactive Substance or Biogenic Amine
A1.4.1.1.3.2 Hormone
A1.4.1.1.3.3 Enzyme
A1.4.1.1.3.4 Vitamin
A1.4.1.1.3.5 Immunologic Factor
A1.4.1.1.3.6 Receptor
A1.4.1.1.4 Indicator
A1.4.1.1.5 Hazardous or Poisonous Substance
A1.4.1.2 Chemical Viewed Structurally
A1.4.1.2.1 Organic Chemical
A1.4.1.2.1.5 Nucleic Acid
A1.4.1.2.1.6 Organophosphorus Compound
A1.4.1.2.1.7 Amino Acid
A1.4.1.2.1.8 Carbohydrate
A1.4.1.2.1.9 Lipid
A1.4.1.2.1.9.1 Steroid
A1.4.1.2.1.9.2 Eicosanoid
A1.4.1.2.2 Inorganic Chemical
A1.4.1.2.3 Element
A1.4.2 Body Substance
A1.4.3 Food
A2 Conceptual Entity
A2.1 Idea or Concept
A2.1.1 Temporal Concept
A2.1.2 Qualitative Concept
A2.1.3 Quantitative Concept
A2.1.4 Functional Concept
A2.1.4.1 Body System
A2.1.5 Spatial Concept
A2.1.5.1 Body Space or Junction
A2.1.5.2 Body Location or Region
A2.1.5.3 Molecular Sequence
A2.1.5.3.1 Nucleotide Sequence
A2.1.5.3.2 Amino Acid Sequence
A2.1.5.4 Geographic Area
A2.2 Finding
A2.2.1 Laboratory or Test Result
A2.2.2 Sign or Symptom
A2.3 Organism Attribute
A2.3.1 Clinical Attribute
A2.4 Intellectual Product
A2.4.1 Classification
A2.4.2 Regulation or Law
A2.5 Language
A2.6 Occupation or Discipline
A2.6.1 Biomedical Occupation or Discipline
A2.7 Organization
A2.7.1 Health Care Related Organization
A2.7.2 Professional Society
A2.7.3 Self-help or Relief Organization
A2.8 Group Attribute
A2.9 Group
A2.9.1 Professional or Occupational Group
A2.9.2 Population Group
A2.9.3 Family Group
A2.9.4 Age Group
A2.9.5 Patient or Disabled Group
B Event
B1 Activity
B1.1 Behavior
B1.1.1 Social Behavior
B1.1.2 Individual Behavior
B1.2 Daily or Recreational Activity
B1.3 Occupational Activity
B1.3.1 Health Care Activity
B1.3.1.1 Laboratory Procedure
B1.3.1.2 Diagnostic Procedure
B1.3.1.3 Therapeutic or Preventive Procedure
B1.3.2 Research Activity
B1.3.2.1 Molecular Biology Research Technique
B1.3.3 Governmental or Regulatory Activity
B1.3.4 Educational Activity
B1.4 Machine Activity
B2 Phenomenon or Process
B2.1 Human-caused Phenomenon or Process
B2.1.1 Environmental Effect of Humans
B2.2 Natural Phenomenon or Process
B2.2.1 Biologic Function
B2.2.1.1 Physiologic Function
B2.2.1.1.1 Organism Function
B2.2.1.1.1.1 Mental Process
B2.2.1.1.2 Organ or Tissue Function
B2.2.1.1.3 Cell Function
B2.2.1.1.4 Molecular Function
B2.2.1.1.4.1 Genetic Function
B2.2.1.2 Pathologic Function
B2.2.1.2.1 Disease or Syndrome
B2.2.1.2.1.1 Mental or Behavioral Dysfunction
B2.2.1.2.1.2 Neoplastic Process
B2.2.1.2.2 Cell or Molecular Dysfunction
B2.2.1.2.3 Experimental Model of Disease
B2.3 Injury or Poisoning

KEY

Behaviors, symptoms, diagnoses, and diet
Behaviors, symptoms, and diagnoses
Diagnoses or ICD-9

Figure 6.1: Categories of the MRSTY and the filtering used in this study, including behaviors, symptoms, diagnoses, and diet (yellow), behaviors, symptoms, and diagnoses (pink), diagnoses only (black box).

a
able
about
above
according
accordingly
across
actually
after
afterwards
again
against
all
allow
allows
almost
alone
along
already
also
although
always
am
among
amongst
an
and
another
any
anybody
anyhow
anyone
anything
anyway
anyways
anywhere
apart
appear
appreciate
appropriate
are
around
as
aside
ask
asking
associated
at
available
away
awfully
b
be
became
because
become
becomes
becoming
been
before
beforehand
behind
being
believe
below
beside
besides
best
better
between
beyond
both
brief
but
by
c
came
can
cannot
cant
cause
causes
certain
certainly
changes
clearly
co
com
come
comes
concerning
consequently
consider
considering
contain
containing
contains
correspondin
g
could
course
currently
d
definitely
described
despite
did
different
do
does
doing
done
down
downwards
during
e
each
edu
eg
eight
either
else
elsewhere
enough
entirely
especially
et
etc
even
ever
every
everybody
everyone
everything
everywhere
ex
exactly
example
except
f
far
few
fifth
first
five
followed
following
follows
for
former
formerly
forth
four
from
further
furthermore
g
get
gets
getting
given
gives
go
goes
going
gone
got
gotten
greetings
h
had
happens
hardly
has
have
having
he
hello
help
hence
her
here
hereafter
hereby
herein
hereupon
hers
herself
hi
him
himself
his
hither
hopefully
how
howbeit
however
i
ie
if
ignored
immediate
in
inasmuch
inc
indeed
indicate
indicated
indicates
inner
insofar
instead
into
inward
is
it
its
itself
j
just
k
keep
keeps
kept
know
knows
known
l
last
lately
later
latter
latterly
least
less
lest
let
like
liked
likely
little
look
looking
looks
ltd
m
mainly
many
may
maybe
me
mean
meanwhile
merely
might
more
moreover
most
mostly
much
must
my
myself
n
name
namely
nd
near
nearly
necessary
need
needs
neither
never
nevertheless
new
next
nine
no
nobody
non
none
noone
nor
normally
not
nothing
novel
now
nowhere
o
obviously
of
off
often
oh
ok
okay
old
on
once
one
ones
only
onto
or
other
others
otherwise
ought
our
ours
ourselves
out
outside
over
overall
own
p
particular
particularly
per
perhaps
placed
please
plus
possible
presumably
probably
provides
q
que
quite
qv
r
rather
rd
re
really
reasonably
regarding
regardless
regards
relatively
respectively
right
s
said
same
saw
say
saying
says
second
secondly
see
seeing
seem
seemed
seeming
seems
seen
self
selves
sensible
sent
serious
seriously
seven
several
shall
she
should
since
six
so
some
somebody
somehow
someone
something
sometime
sometimes
somewhat
somewhere
son
soon
sorry
specified
specify
specifying
still
sub
such
sup
sure
t
take
taken
tell
tends
th
than
thank
thanks
thanx
that
thats
the
their
theirs
them
themselves
then
thence
there
thereafter
thereby
therefore
therein
theres
thereupon
these
they
think
third
this
thorough
thoroughly
those
though
three
through
throughout
thru
thus
to
together
too
took
toward
towards
tried
tries
truly
try
trying
twice
two
u
un
under
unfortunately
unless
unlikely
until
unto
up
upon
us
use
used
useful
uses
using
usually
uucp
v
value
various
very
via
viz
vs
w
want
wants
was
way
we
welcome
well
went
were
what
whatever
when
whence
whenever
where
whereafter
whereas
whereby
wherein
whereupon
wherever
whether
which
while
whither
who
whoever
whole
whom
whose
why
will
willing
wish
with
within
without
wonder
would
would
x
y
yes
yet
you
your
yours
yourself
yourselves
z
zero

Figure 6.2: Stopwords that were used in the filtering from this study, drawn for the most part from Arora et al. (2012a).

Bibliography

Asperger and asd uk online forum. http://www.asd-forum.org.uk/, 2014a.

Asd friendly. http://www.asdfriendly.org, 2014b.

Autism web. http://www.autismweb.com/, 2014.

Talk about autism. http://www.talkaboutautism.org.uk/, 2014.

Andrew L Alexander, Jee Eun Lee, Mariana Lazar, and Aaron S Field. Diffusion tensor imaging of the brain. *Neurotherapeutics*, 4(3):316–329, 2007.

June Almenoff, Joseph M Tonning, A Lawrence Gould, Ana Szarfman, Manfred Hauben, Rita Ouellet-Hellstrom, Robert Ball, Ken Hornbuckle, Louisa Walsh, Chuen Yee, et al. Perspectives on the use of data mining in pharmacovigilance. *Drug safety*, 28(11):981–1007, 2005.

Sanjeev Arora, Rong Ge, Yoni Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. *arXiv preprint arXiv:1212.4777*, 2012a.

Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 1–10. IEEE, 2012b.

American Psychiatric Association et al. *Diagnostic and statistical manual of mental disorders, (DSM-5®)*. American Psychiatric Pub, 2013.

Jon Baio. Prevalence of autism spectrum disorders: Autism and developmental disabilities monitoring network, 14 sites, united states, 2008. morbidity and mortality weekly report. surveillance summaries. volume 61, number 3. *Centers for Disease Control and Prevention*, 2012.

Simon Baron-Cohen. The cognitive neuroscience of autism. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(7):945–948, 2004.

Simon Baron-Cohen. Causes of autism. 2006.

Matthew K Belmonte, Greg Allen, Andrea Beckel-Mitchener, Lisa M Boulanger, Ruth A Carper, and Sara J Webb. Autism and abnormal development of brain connectivity. *The Journal of Neuroscience*, 24(42):9228–9231, 2004.

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

David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

David M Blei and John D Lafferty. Topic models. *Text mining: classification, clustering, and applications*, 10:71, 2009.

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

Ruth A Carper and Eric Courchesne. Inverse correlation between frontal lobe and cerebellum sizes in children with autism. *Brain*, 123(4):836–844, 2000.

Allison J.B. Chaney and David M. Blei. Visualizing topic models, 2012.

Brant W Chee, Richard Berlin, and Bruce Schatz. Predicting adverse drug events from personal health messages. In *AMIA Annual Symposium Proceedings*, volume 2011, page 217. American Medical Informatics Association, 2011.

Wen-Ying Sylvia Chou, Yvonne M Hunt, Ellen Burke Beckjord, Richard P Moser, and Bradford W Hesse. Social media use in the united states: implications for health communication. *Journal of medical Internet research*, 11(4), 2009.

John N Constantino, Sandra A Davis, Richard D Todd, Matthew K Schindler, Maggie M Gross, Susan L Brophy, Lisa M Metzger, Christiana S Shoushtari, Reagan Splinter, and Wendy Reich. Validation of a brief quantitative measure of autistic traits: comparison of the social responsiveness scale with the autism diagnostic interview-revised. *Journal of autism and developmental disorders*, 33(4):427–433, 2003.

Eric Courchesne and Karen Pierce. Why the frontal cortex in autism might be talking only to itself: local over-connectivity but long-distance disconnection. *Current opinion in neurobiology*, 15(2):225–230, 2005.

Kerstin Denecke and Wolfgang Nejdl. How valuable is medical social media data? content analysis of the medical web. *Information Sciences*, 179(12):1870–1880, 2009.

David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in neural information processing systems*, page None, 2003.

Demetria Ennis-Cole, Beth A Durodoye, and Henry L Harris. The impact of culture on autism diagnosis and treatment: considerations for counselors and other professionals. *The Family Journal*, page 1066480713476834, 2013.

Hadeel Faras, Nahed Al Ateeqi, and Lee Tidmarsh. Autism spectrum disorders. *Annals of Saudi medicine*, 30(4):295, 2010.

Patrick Flaherty, Guri Giaever, Jochen Kumm, Michael I Jordan, and Adam P Arkin. A latent variable model for chemogenomic profiling. *Bioinformatics*, 21(15):3286–3293, 2005.

Jeana H Frost and Michael P Massagli. Social uses of personal health information within patientslikeme, an online patient community: what can happen when patients have access to one another's data. *Journal of Medical Internet Research*, 10(3), 2008.

Daniel H Geschwind and Pat Levitt. Autism spectrum disorders: developmental disconnection syndromes. *Current opinion in neurobiology*, 17(1):103–111, 2007.

Lise Getoor, Eran Segal, Ben Taskar, and Daphne Koller. Probabilistic models of text and link structure for hypertext classification. In *IJCAI workshop on text learning: beyond supervision*, pages 24–29, 2001.

Thomas L Griffiths, Mark Steyvers, David M Blei, and Joshua B Tenenbaum. Integrating topics and syntax. In *Advances in neural information processing systems*, pages 537–544, 2004.

Yoni Halpern. Code. `http://www.cs.nyu.edu/~halpern/code.html`, 2014.

Francesca Happé, Angelica Ronald, and Robert Plomin. Time to give up on a single explanation for autism. *Nature neuroscience*, 9(10):1218–1220, 2006.

Carleen Hawn. Take two aspirin and tweet me in the morning: how twitter, facebook, and other social media are reshaping health care. *Health affairs*, 28(2):361–368, 2009.

Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

Norman L Johnson, Samuel Kotz, and N Balakrishnan. *Continuous Multivariate Distributions, volume 1, Models and Applications*, volume 59. New York: John Wiley & Sons, 2002.

Marissa King and Peter Bearman. Diagnostic change and the increased prevalence of autism. *International journal of epidemiology*, 38(5):1224–1234, 2009.

Melvin Konner. Epidemiology: Epidemic of panic. *Nature*, 469(7331):468–469, 2011.

Philip J Landrigan. What causes autism? exploring the environmental contribution. *Current opinion in pediatrics*, 22(2):219–225, 2010.

Donald A Lindberg, Betsy L Humphreys, and Alexa T McCray. The unified medical language system. *Methods of information in Medicine*, 32(4):281–291, 1993.

Catherine Lord, Michael Rutter, Pamela C DiLavore, Susan Risi, Katherine Gotham, and S Bishop. *Autism diagnostic observation schedule: ADOS-2*. Western Psychological Services Los Angeles, CA, 2012.

Catherine Lord, Edwin H Cook, Bennett L Leventhal, and David G Amaral. Autism spectrum disorders. *Autism: The Science of Mental Health*, 28:217, 2013.

David S Mandell, Maytali M Novak, and Cynthia D Zubritsky. Factors associated with age of diagnosis among children with autism spectrum disorders. *Pediatrics*, 116(6):1480–1486, 2005.

Johnny L Matson and Marie S Nebel-Schwalm. Comorbid psychopathology with autism spectrum disorder in children: An overview. *Research in developmental disabilities*, 28(4): 341–352, 2007.

Rebecca Muhle, Stephanie V Trentacoste, and Isabelle Rapin. The genetics of autism. *Pediatrics*, 113(5):e472–e486, 2004.

US Dept of Health. *ICD 9 CM. The International Classification of Diseases. 9. Rev: Clinical Modification.; Vol. 1: Diseases: Tabular List. ; Vol. 2: Diseases: Alphabetic Index. ; Vol. 3: Procedures: Tabular List and Alphabetic Index.* US Government Printing Office, 1980.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

Daniel Ramage, Paul Heymann, Christopher D Manning, and Hector Garcia-Molina. Clustering the tagged web. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 54–63. ACM, 2009.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

Michael Rutter, Ann Le Couteur, Catherine Lord, and Raffaella Faggioli. *ADI-R: Autism diagnostic interview–revised: Manual*. OS, Organizzazioni speciali, 2005.

Jane Sarasohn-Kahn. *The wisdom of patients: Health care meets online social media*. California HealthCare Foundation Oakland, CA, 2008.

Eric Schopler, Robert Jay Reichler, and Barbara Rochen Renner. *The childhood autism rating scale (CARS)*. Western Psychological Services Los Angeles, 1988.

Catherine Arnott Smith and P Zoë Stavri. Consumer health vocabulary. In *Consumer Health Informatics*, pages 122–128. Springer, 2005.

Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.

James S Sutcliffe. Insights into the pathogenesis of autism. *Science*, 321(5886):208–209, 2008.

Katherine D Tsatsanis, Byron P Rourke, Ami Klin, Fred R Volkmar, Domenic Cicchetti, and Robert T Schultz. Reduced thalamic volume in high-functioning individuals with autism. *Biological psychiatry*, 53(2):121–129, 2003.

Maria Valicenti-McDermott, Kathryn Hottinger, Rosa Seijo, and Lisa Shulman. Age at diagnosis of autism spectrum disorders. *The Journal of pediatrics*, 161(3):554–556, 2012.

Hanna M Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984. ACM, 2006.

QT Zheng. Consumer health vocabulary initiative. `http://consumerhealthvocab.org/`, 2014.