A Discriminative Model to Predict Comorbidities in ASD

Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my deepest appreciation to Professor Finale Doshi-Velez for serving as my thesis adviser, offering her guidance, and introducing me to this topic area. I thank Angela Fan, Andre Nguyen, and Vincent Nguyen, fellow members of Senior Thesis Club, for deepening my understanding and generously sharing your insights. Thank you to Melih Elibol for allowing me to build off of your code and answering all and any of my questions. Thank you to my thesis readers for your time and consideration.

Chapter 1

Introduction

Autism spectrum disorders (ASD) form a heterogeneous, neurodevelopmental syndrome diagnosed via clinical assessment and defined by atypical social behavior, disrupted verbal and non-verbal communication, and by unusual patterns of restricted interests and repetitive behaviors [5]. ASD typically begin in infancy, with the onset of the three core disturbances before three years of age. The incidence of ASD is most recently estimated at 1 in 68 children [6].

Across the core characteristics, there are significant differences in the extent and quality of symptoms [13]. For example, despite similar presentations at the time of diagnosis, approximately 30 percent of children with ASD remain nonverbal into adulthood, whereas 30 percent demonstrate a reasonably normal verbal IQ, with primary deficits in language use and context [16].

Furthermore, children who have ASD have a higher comorbidity burden than the general pediatric population. A *comorbidity* designates the presence of more than one distinct condition in an individual and is most often defined in relation to a specific index condition. Children with ASD have higher than expected rates of eczema, allergies, asthma, ear and respiratory infections, gastrointestinal problems, severe headaches, migraines, and seizures, among other conditions [19].

1.1 Overview

Different diseases may be found in the same individual due to chance, selection bias, or one or more types of causal association. While a comorbidity that occurs by chance or selection bias remains relevant because it leads to incorrect assumptions about causality, we will focus on the etiological associations between conditions, including direct causation, associated risk factors, heterogeneity (disease risk factors that are not correlated but can each cause diseases associated with the other risk factor), and independence [33].

Genetic variation is known to play a large role in risk for ASD, but a large number of genes, estimated to be near 1000, appear to confer risk for ASD [22]. Without a clearly defined genetic understanding of ASD, we must then turn to comorbidities and their co-occurrence patterns in ASD to reveal etiological associations. Understanding co-occurrence patterns not only has clinical implications for disease management but also can stratify the risk for various conditions across individuals with ASD [12].

Over the course of the thesis, I model co-occurrence patterns in single-concept words within social online ASD forum posts. In particular, by applying logistic regression, a machine learning technique, I predict the presence of a single-concept word in posts written by an author on an ASD subject from the presence of single-concept words in posts written by the same author on that subject at an earlier age. By restricting single-concept words to those semantically associated with comorbidities, this approach can potentially model co-occurrence patterns in ASD comorbidities over time.

1.2 Motivation

This thesis harnesses the data within social online forum posts, which feature positive characteristics in terms of volume, verbosity, and context.

When coping with illness, individuals turn to social support in search of information, which can in turn improve physical functioning and psychosocial well-being [28]. According to a survey by the National Cancer Institute, 7.5 million Americans ventured online to acquire peer support about a health issue during 2012 [15]. For health conditions likely to threaten personal relationships or with a greater potential for loss in the form of death, nurturant messages were more common in these computer-mediated contexts, whereas for chronic conditions, like ASD, action-facilitating messages were more common [29].

For families dealing with ASD, they face economic costs and emotional stress to provide care for the diagnosed family member, and social online forums provide an open and easily accessible platform to share, gather, and exchange information. Forums have become an immense source of knowledge for other members of the community dealing with similar challenges [30].

Electronic health records remain the gold standard for understanding co-occurrence patterns among comorbidities because they contain structured, clean, less noisy, and mostly complete data. The data exists in the form of symptom descriptions, documentation of examinations, diagnostic reasoning, and motivations for treatment decisions. Still, the unstructured text of social online forum posts written by a caretaker presents valuable information. Individuals can flexibly post, unburdened by temporal, geographical, and spatial limitations, and can carefully consider their message, developing it at their own pace before posting. In addition, social online forums may bring together a more varied range of individuals offering diverse perspectives, experiences, opinions, and sources of information. Lastly, participation in an social online forum allows a greater degree of anonymity, facilitating self disclosure and discussion of sensitive issues with less fear of embarrassment or judgement [11].

1.3 Related Work

Recently, a variety of machine learning algorithms have been applied to electronic health records and have incorporated sociodemographic and clinical characteristics for the purpose of assessing risk for certain conditions and predicting comorbidities. These include a Bayesian network model composed of age, sex, race, smoking history, and eight comorbidity variables to predict chronic obstructive pulmonary disease in asthma patients; support vector machines to predict cancer survival since date of diagnosis; a logistic regression model to estimate risk for treatment resistance among outpatients with major depressive disorder; a least squares extreme logistic regression to predict prostate cancer mortality; and random forests and elastic net penalized logistic regressions to predict post-traumatic stress disorder from pre-trauma risk factors [14, 15, 26, 25, 18].

While the given examples primarily relied on structured data from electronic health records, computerized text analytics have also been applied to unstructured medical records, as in the case of a linguistics-driven prediction to estimate the risk of suicide from unstructured clinical notes taken from a national sample of U.S. Veterans Administration medical records [27]. In particular, the latter application studied single-word terms and their numerical counts in a patient record for the model. The pervasive use of machine learning techniques applied to structured and unstructured medical data points to their accuracy and effectiveness in large-scale pattern recognition.

1.4 Contribution

The overarching contribution of this thesis is to present a discriminative model baseline for predicting the trajectory of single-concept words in social online ASD forum posts. The main contributions of this thesis are as follows.

- I explore and aggregate posts from social online forums related to ASD, in which the ASD subject has been age-identified and medical concepts have been extracted via regular expression filters.
- I define a discriminative approach for predicting single-concept words for forum posts regarding an ASD subject greater than five years old from forum posts of that subject under five years old.

- I evaluate and interpret the accuracy of the baseline discriminative model against that of a dynamic topic model - a generative, predictive model of a sequential corpus.

The thesis will be structured as follows. Chapter 2 introduces the notation and basic theory used throughout the thesis. Chapter 3 presents the exploration of social online forum data and the experimental setup for the prediction task. Chapter 4 discusses the results of the model and concludes.

Chapter 2

Basic Theory

In this chapter, I present the notation and basic theory that will be used throughout this thesis. Defined as the field of study that gives computers the ability to learn without being explicitly programmed, machine learning can be subdivided into further categories: supervised and unsupervised [31]. Supervised learning narrows in on the classification problem: given a set of data points that each belong to one of any number of classes, how can one generalize a hypothesis that will properly classify unseen data points?

2.1 Discriminative Classifiers

A *discriminant function* takes an input vector $\mathbf{x}$ and assigns it to one of k classes, denoted C_k. For a linear discriminant function, the decision surface is a hyperplane. In the case of two classes, the simplest representation of a linear discriminant function is as follows,

$$y(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_0$$

where $\mathbf{w}$ is a weight vector, and w_0 is a bias. An input vector $\mathbf{x}$ is assigned to class C_1 if $y(\mathbf{x}) \geq 0$ and to class C_2 otherwise.

Through Bayes' theorem, we can compute the posterior probability for class C_1 as,

$$p(C_1|\mathbf{x}) = \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_1)p(C_1) + p(\mathbf{x}|C_2)p(C_2)}$$

If we define

$$a = \ln \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)},$$

then

$$p(C_1|\mathbf{x}) = \frac{1}{1 + e^{-a}}.$$

The logistic sigmoid function is defined by,

$$\sigma(a) = \frac{1}{1 + e^{-a}}$$

so,

$$p(C_1|\mathbf{x}) = \sigma(a).$$

Logistic Regression

For the two-class classification problem, we can use the functional form of the generalized linear model explicitly and determine its parameters directly by using maximum likelihood. By maximizing a likelihood function defined through the conditional distribution $p(C_k|\mathbf{x})$, there will typically be fewer adaptive parameters to be determined and may also lead to improved predictive performance, particularly when the class-conditional density assumptions give a poor approximation to the true distributions [7].

For a data set $\{\mathbf{x}_n, t_n\}$ where correct label $t_n \in \{0, 1\}$, with $n = 1, ..., N$, the likelihood function can be written,

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n} (1 - y_n)^{1-t_n}$$

where target vector $\mathbf{t} = (t_1, ..., t_N)^\top$ and $y_n = p(C_1|\mathbf{x}_n)$. To achieve the *cross-entropy* error function, we can take the negative logarithm of the likelihood,

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N}(t_n \ln y_n + (1-t_n)\ln(1-y_n))$$

where $y_n = \sigma(a_n)$ and $a_n = \mathbf{w}^\top \mathbf{x}_n$.

The derivative of the logistic sigmoid function can be expressed in terms of the sigmoid function itself,

$$\frac{d\sigma}{da} = \sigma(1-\sigma).$$

Taking the gradient of the error function with respect to $\mathbf{w}$, we obtain

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\phi_n$$

where $\phi_n = \phi(\mathbf{x}_n)$, a fixed nonlinear transformation of the inputs using a basis function.

The error function can be minimized by the *Newton-Raphson* iterative optimization scheme, which uses a local quadratic approximation to the log likelihood function. The Newton-Raphson update for minimizing a function $E(\mathbf{w})$ takes the form,

$$\mathbf{w}^{(\text{new})} = \mathbf{w}^{(\text{old})} - \mathbf{H}^{-1}\nabla E(\mathbf{w})$$

where $\mathbf{H}$ is the Hessian matrix whose elements comprise the second derivatives of $E(\mathbf{w})$ with respect to the components of $\mathbf{w}$ [7].

If we apply the Newton-Raphson update to the cross-entropy error function for the logistic regression model, the gradient and Hessian of this function are given by

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\phi_n = \mathbf{\Phi}^\top(\mathbf{y} - \mathbf{t})$$

$$H = \nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1-y_n)\phi_n\phi_n^\top = \mathbf{\Phi}^\top\mathbf{R}\mathbf{\Phi}$$

where $N \times N$ diagonal matrix $\mathbf{R}$ has elements $R_{nn} = y_n(1 - y_n)$.

The Newton-Raphson update formula for the logistic regression model then becomes

$$\mathbf{w}^{(\text{new})} = \mathbf{w}^{(\text{old})} - (\mathbf{\Phi}^\top \mathbf{R} \mathbf{\Phi})^{-1} \mathbf{\Phi}^\top (\mathbf{y} - \mathbf{t})$$

$$= (\mathbf{\Phi}^\top \mathbf{R} \mathbf{\Phi})^{-1} (\mathbf{\Phi}^\top \mathbf{R} \mathbf{\Phi} \mathbf{w}^{(\text{old})} - \mathbf{\Phi}^\top (\mathbf{y} - \mathbf{t}))$$

$$= (\mathbf{\Phi}^\top \mathbf{R} \mathbf{\Phi})^{-1} \mathbf{\Phi}^\top \mathbf{R} \mathbf{z}$$

where $\mathbf{z}$ is an N-dimension vector with elements $\mathbf{z} = \mathbf{\Phi} \mathbf{w}^{(\text{old})} - \mathbf{R}^{-1}(\mathbf{y} - \mathbf{t})$.

Discriminative vs. Generative Classifiers

Whereas discriminative classifiers model the posterior probability $p(y|x)$ of the inputs x and the label y directly, or learn a direct map from inputs x to the class labels, generative classifiers learn a model of the joint probability $p(x, y)$ and make their predictions by using Bayes rules to calculate $p(y|x)$ and then picking the most likely label y [20].

Ng and Jordan (2002) found that while a generative model does have a higher asymptotic error than the discriminative model, as the size of the training set increases, the generative model may also approach its asymptotic error faster than the discriminative model. Indeed, in the generative case, the number of training examples needed to approach the asymptotic error might be on the order of $\log n$, rather than on the order of n for logistic regression. All together, this suggests that there can be two courses of performance, one in which the generative model has already approached its asymptotic error and consequently performs better, and the other in which the discriminative model approaches its lower asymptotic error and performs better [24].

I will return to the discussion on the distinction between discriminative and generative classifiers when evaluating the comparative performance of logistic regression, Latent Dirichlet Allocation, and a dynamic topic model in Chapter 4.

2.2 Topic Modeling

When presented with a corpus, or a group of documents, *topic modeling* uncovers the hidden thematic structure in the documents, empowering the searching, browsing, and summarizing of texts [8]. Topic models are based on the notion that each document can be represented by a mixture of topics. In text analysis, topic models typically adopt the *bag-of-words* assumption that ignores the information from the ordering of words. Each document in a given corpus is thus represented by a histogram containing the occurrence of words. The histogram is modeled by a distribution over a certain number of topics, each of which is a distribution over words in the vocabulary.

Latent Dirichlet Allocation

Latent Dirichlet Allocation is a generative topic model in which each document is composed of multiple topics [10]. The vocabulary will have V words and a topic will be a distribution over this vocabulary. We will use K topics and the kth topic is a vector β_k, where $\beta_{k,v} \geq 0$ and $\sum_v \beta_{k,v} = 1$. Each document can be described by a set of word counts $\mathbf{w}_d$ where $w_{d,v}$ is a nonnegative integer. Document d has N_d words in total, such that $\sum_v w_{d,v} = N_d$. The unknown overall mixing proportion of topics is θ, where $\theta_k \geq 0$ and $\sum_k \theta_k = 1$. Each of the D documents has a distribution over the topics. Thus,

$$\alpha : \text{Dirichlet prior on the per document topic distributions}$$

$$\beta : \text{Dirichlet prior on the per topic word distribution}$$

$$\theta_i : \text{topic distribution for document } i$$

$$\psi_k : \text{word distribution for topic } k$$

$$z_{i,j} : \text{topic for the } j\text{th word in document } i \text{ and}$$

$$w_{i,j} : \text{specific word.}$$

The Dirichlet distribution is the multivariate generalization of the beta distribution and the conjugate prior of the categorical distribution and multinomial distribution. The algorithmic process is formalized below as,

Algorithm 1 LDA

```
procedure
    Choose a topic distribution $\theta_i \sim \text{Dir}(\alpha)$ for each document in the corpus.
    Choose a word distribution $\psi_k \sim \text{Dir}(\beta)$ for each topic.
    for every word $i$ in document $j$ do,
        Choose a topic $z_{i,j} \sim \text{Multinomial}(\theta_i)$.
        Choose a word $w_{i,j} \sim \text{Multinomial}(\psi_{z_{i,j}})$.
```

Dynamic Topic Model

The dynamic topic model is an extension of Latent Dirichlet Allocation that no longer treats words exchangeably but instead captures the evolution of topics in a sequentially organized corpus of documents [9]. In a dynamic topic model, the data is divided by time slice. The documents of each slice are modeled with K topics, where the topics associated with slice t evolve from the topics associated with slice $t-1$.

Because Dirichlet distributions, typically used to model uncertainty about the distribution over words, are not amenable to sequential modeling, the per-topic word distributions are chained in a state space model that evolves with Gaussian noise. Consequently, $\beta_{t,k}|\beta_{t-1,k} \sim N(\beta_{t-1,k}, \sigma^2 I)$. Similarly, the per-document topic proportions θ are drawn from a Dirichlet distribution in LDA, but in the dynamic topic model, a logistic normal with mean α expresses uncertainty over proportions. The sequential structure between models is captured $\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2 I)$. A collection of topic models are sequentially considered by chaining topics and topic proportion distributions. Thus,

$$\alpha_t : \text{per document topic distribution at time } t$$

$$\beta_{t,k} : \text{word distribution of topic } k \text{ at time } t$$

$\eta_{t,d}$: topic distribution for document d in time t

$z_{t,d,n}$: topic for the nth word in document d in time t, and

$w_{t,d,n}$: specific word.

The algorithmic process at time slice t is formalized below, where $\pi(x)$ is a mapping from the natural parametrization x to the mean parametrization via $\pi(x_i) = \frac{e^{x_i}}{\sum_i e^{x_i}}$.

Algorithm 2 Dynamic Topic Model

```
procedure
    Choose a word distribution $\beta_{t,k}|\beta_{t-1,k} \sim N(\beta_{t-1,k}, \sigma^2 I)$ for each topic.
    Choose a topic distribution $\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2 I)$ over the corpus.
    for each document do,
        Choose a topic distribution $\eta_{t,d} \sim N(\alpha_t, a^2 I)$.
        for each word do,
            Choose topic $z_{t,n,d} \sim$ Multinomial$(\pi(\eta_t, d))$.
            Choose a word $w_{t,n,d} \sim$ Multinomial$(\pi(\beta_t, z_{t,d,n}))$.
```

Chapter 3

Experimental Design and Methods

This section describes the methods that were applied to analyze the sources of text, social online forums related to ASD. This unstructured text has generally been previously unexplored in terms of predicting single-concept word trajectory with a focus in paralleling comorbidity co-occurrences.

3.1 Data Collection

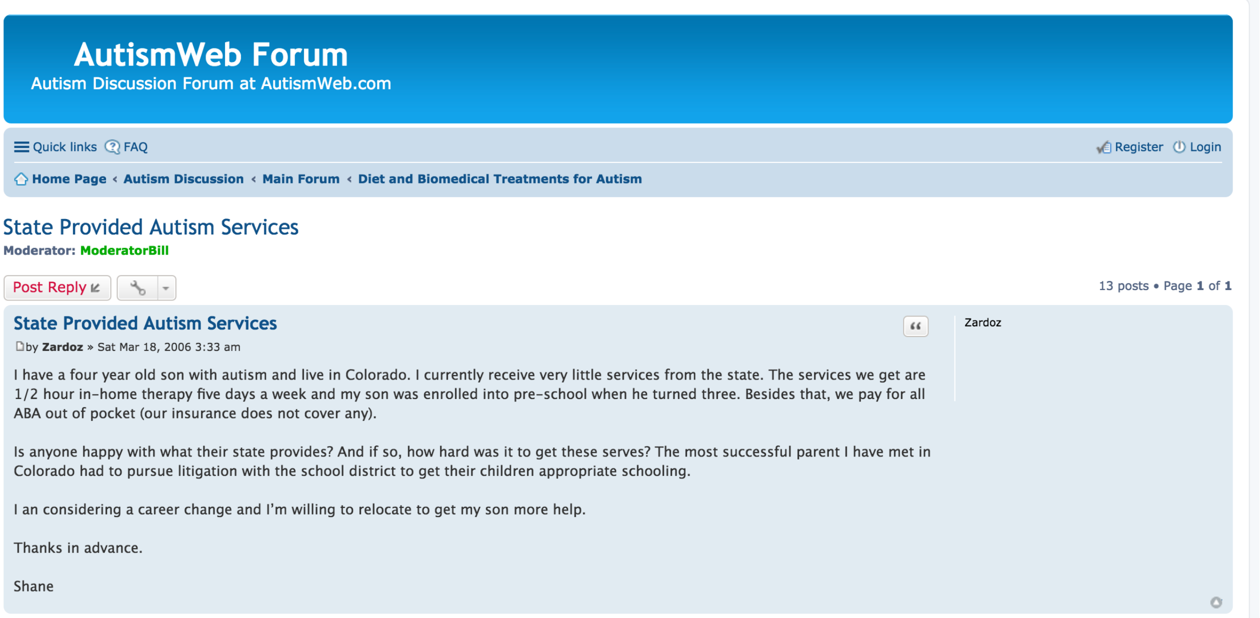
The data set consists of text postings from social online forums focused on ASD. These forums are AutismWeb, "a community of parents interested in autism, Pervasive Developmental Disorder (PDD), and Asperger Syndrome [1]," ASD Friendly, a "close-knit community of parents and carers of people with Autism and Asperger's Syndrome [2]," and Asperger's and ASD UK Online Forum [3]. In February 2016, all text postings were extracted by scraping all subforums with the BeautifulSoup package. This entailed 80,927 threads and 664,954 posts, with an average of 700 characters per post. *Regular expression* filters were created to extract suspected instances in a text post where age may be mentioned in passing. We applied this filter on all scraped text postings from these forums. Additionally, we incorporated additional information about the text posts, such as authorship and time of posting.

Age Extraction

Of the total number of posts, more than 27,000 posts were associated with the age of the ASD subject concerned as determined by regular expression filters. *Regular expression*, the standard algebraic notation for characterizing text sequences, is used for specifying text strings in situations like web searching and information retrieval, in word processing, in computing frequencies from corpora, and in other such tasks. A regular expression search function will search through the corpus returning all texts that contain the pattern [17].

For this application, the regular expression greedily matched on all posts indicative of age, pattern matching on phrases like ”six-year-old” and ”is six years.” The patterns were designed to incorporate specific uses of language present in these online forums. For example, authors often abbreviated pronouns using colloquialisms such as `dd` for *dearest daughter*, `dgs` for *dearest godson*, and `yo` for *years old*. These abbreviations were taken into account when creating the search patterns. In addition to understanding the context leading up to a mention of age, words after a possible age match were captured to verify that the match did not refer to an entity other than age, such as weight, height, or time. To avoid multiple mentions of age in a single text post, namely the problem of coreference, only filtered posts that contain a single mention of age were retained. For more information on the age filter, please see Vincent Nguyen’s thesis.

As an illustrative example, user Zardoz posted the following on AutismWeb in March 2006. The phrase ”four year old son” matches the regular expression pattern that establishes the age of the ASD subject. Consequently, the post was associated with the age of 4, and all other posts of user Zardoz will now be associated with the appropriate age depending on the time interval between it and the March 2006 post. Zardoz only published one additional post in February 2006, and because that post was within less than one year from the March 2006 post, the February 2006 post also was tagged with the age of 4.


AutismWeb Forum
Autism Discussion Forum at AutismWeb.com
Quick links
FAQ
Register
Login
Home Page ‹ Autism Discussion ‹ Main Forum ‹ Diet and Biomedical Treatments for Autism
Post Reply
13 posts • Page 1 of 1
Zardoz

State Provided Autism Services

Moderator: ModeratorBill

State Provided Autism Services

by Zardoz » Sat Mar 18, 2006 3:33 am

I have a four year old son with autism and live in Colorado. I currently receive very little services from the state. The services we get are 1/2 hour in-home therapy five days a week and my son was enrolled into pre-school when he turned three. Besides that, we pay for all ABA out of pocket (our insurance does not cover any).

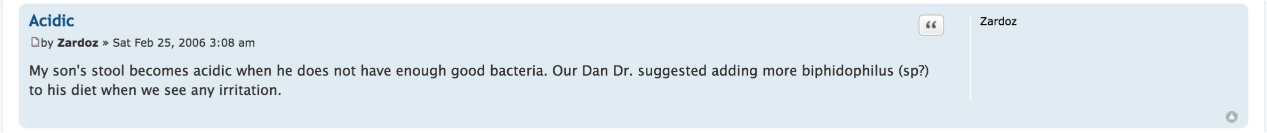
Is anyone happy with what their state provides? And if so, how hard was it to get these serves? The most successful parent I have met in Colorado had to pursue litigation with the school district to get their children appropriate schooling.

I an considering a career change and I'm willing to relocate to get my son more help.

Thanks in advance.

Shane

Figure 3.1: User Zardoz's post on AutismWeb in March 2006.



Acidic

by Zardoz » Sat Feb 25, 2006 3:08 am

My son's stool becomes acidic when he does not have enough good bacteria. Our Dan Dr. suggested adding more biphidophilus (sp?) to his diet when we see any irritation.

Figure 3.2: User Zardoz's only other post on AutismWeb, published in February 2006.

Concept Extraction

Now with age-identified posts, a common vocabulary across the posts was required, specifically a vocabulary of medically relevant words. Developed by the U.S. National Library of Medicine, the Unified Medical Language System [4] is a repository of biomedical vocabularies that integrates more than 2 million names for some 900,000 concepts from more than 60 families of biomedical vocabularies, as well as 12 million relations among these concepts [21].

Within the UMLS Metathesaurus, synonyms, or the words from all of the source vocabularies that have the same intended meaning, are all mapped to one concept unique identifier (CUI), which contains the letter C followed by seven numbers.

To process the posts such that unstructured text content can be mapped to concepts, we must first recognize that healthcare terminology is not consumer English; for example, in a study on communities in PatientsLikeMe, a social networking site for patients, about 43 % of patient-submitted terms are present either as exact (24 %) or as synonymous matches (19 %) to the UMLS Metathesaurus [32]. To address inconsistencies between everyday words likely to be present in forum posts and the technical jargon of the UMLS Metathesaurus used by health care professionals, the *Consumer Health Vocabulary* was leveraged.

The Consumer Health Vocabulary (CHV) Initiative provides a comprehensive database with a mapping between colloquial phrases about health (heart attack) to technical terms (myocardial infarction) associated with the same given concept [34]. Each of the concepts, to which many colloquial words can map, has a unique CUI and a preferred name. There are 158,519 terms mapped to 57,819 unique CUIs in the CHV database.

To extract CUIs from forum posts, a trie was created using nesting dictionaries that stored word sequences and their associated CUIs. Each scraped post was processed individually and returned the longest instances of matching CUIs within the text. Once again, for more information on CUI extraction, please see Vincent Nguyen's thesis.

To continue with the earlier example, Zardoz's post on Autism Web in February 2006 contained the following unstructured text:

> My son's stool becomes acidic when he does not have enough good bacteria. Our Dan Dr. suggested adding more biphidophilus (sp?) to his diet when we see any irritation.

This text mapped to the CUIs of C0424522 (Asleep), C0442696 (Waking), C0234451 (Sleep, Slow-Wave), C1299582 (Unable), C0580846 (Does pull), C0218063 (Ensure - product), and C0557351 (Employed). As demonstrated, while extracted concept C0218063 can be intuitively interpreted from ”biphidophilus” in the original text, because bifidophilus refers to a probiotic that aids digestion and Ensure refers to a liquid nutrition shake, the other extracted CUIs do not necessarily exhibit obvious semantic relationships with the post. The relative bluntness of CUI extraction from the posts via the CHV trie preprocessing remains a matter of future investigation.

3.2 Exploratory Analysis

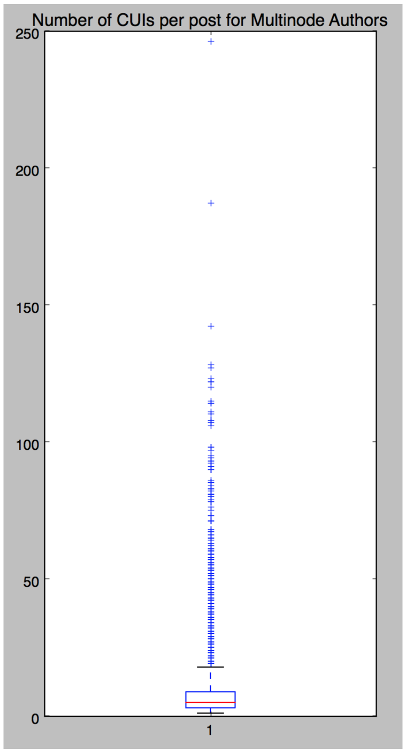


Figure 3.3: Number of CUIs extracted per post.

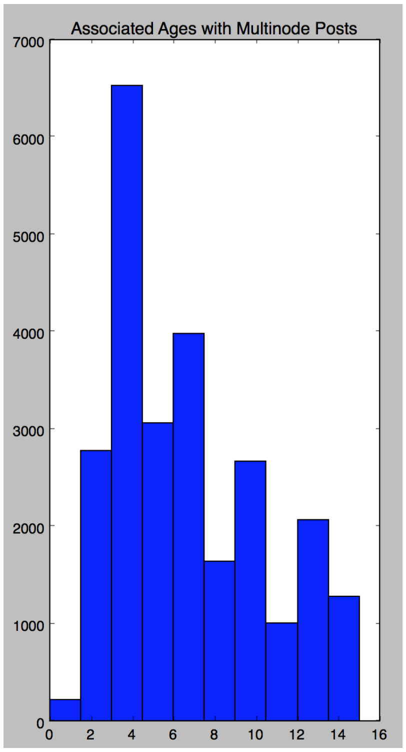


Figure 3.4: Distribution of associated ages extracted from posts.

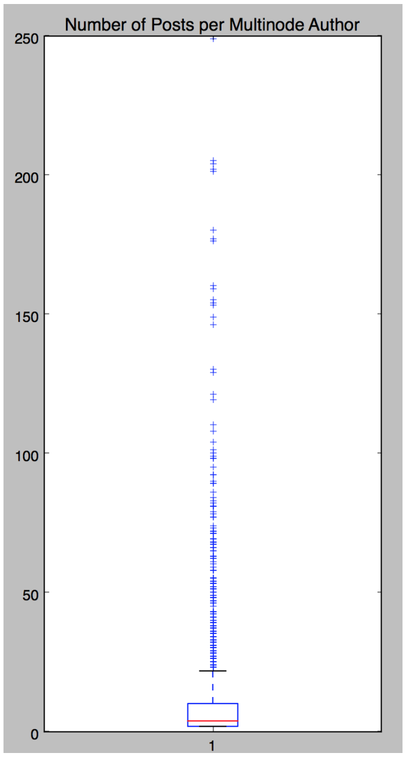


Figure 3.5: Number of posts written per author

For the 27,022 age-identified posts across the three ASD forums, the content was attributed to 4085 authors, and of the 4085 authors, 2304 had posted more than once and are referred to as *multinode* authors. The average number of CUIs extracted per post for a multinode author was 7.4885, but a few outlying authors with upwards of 200 CUIs extracted from their posts inflated the average. Thus, the median number of CUIs extracted per post for a multinode author was 5 CUIs. With regards to the number of posts written by a multinode author, the average was 10.9553 posts, but once again, the average was inflated by a few outliers, including users miami girl and Jack'sMum, who had both posted more than 7000 times on ASD Friendly. As such, the median number of posts written by an author was 4 posts.

To now consider the CUIs extracted from the posts, a total of 3050 different CUIs and 180,851 instances of CUIs were extracted from the posts by multinode authors. As expected, the most frequently extracted CUI, C0004352, referenced ASD, and the other top-appearing CUIs pertained to family designations (parent, daughter) and general activities children undergo in development but may face delays if autistic (speaking, reading).

CUI	Name	Frequency
C0004352	Autistic Disorder	5446
C0011900	Diagnosis	4548
C0234856	Speaking (activity)	4442
C0034754	Reading (activity)	4090
C1299581	Able (finding)	3712
C0011011	Daughter	3134
C1273517	Used by	3075
C0030551	parent	3071
C0032214	Play	3024

Table 3.1: Most Frequent CUIs Extracted from Multinode Authors' Posts

3.3 Experimental Setup

With the goal of discriminatively predicting the presence of a single-concept word in an author's posts regarding an older-aged ASD subject from the presence of single-concept words in her posts regarding the same subject at a younger age, we must define an experimental setup such that we can train a discriminative classifier. One option would be to train several binary logistic regression classifiers, one for each single-concept word in the vocabulary. In this scenario, $X_{ij} = 1$ if single-concept word j is present in author i's posts regarding the younger-aged ASD subject, and 0 otherwise, and $Y_{ij} = 1$ if single-concept word j is present in author i's later posts regarding the subject, and 0 otherwise. Then, the classifiers can be evaluated in aggregate, as detailed in Chapter 4.

Boundary Definition

In order to distinguish an author's ”later” posts from her ”earlier” posts, we must choose a boundary for the inferred age of the ASD subject for which single-concept words present in the author's posts pertaining to an ASD subject older than that boundary age will be accounted in the Y matrix, and those pertaining to an ASD subject that boundary age or younger will be accounted in the X matrix. Considering that only about 2300 authors in the data set had written more than one post, the ideal boundary age would allow the greatest number of authors to still be incorporated, entailing that they had written at least one post on an ASD subject who had been that age or younger and had later written at least one post on the subject who had been older than that age. Given these parameters, the following number of authors would be valid in the data set for the proposed age boundary. Because the boundary age of 5 maximized the number of valid authors, it was determined that 5 would serve as the boundary age.

Concept Semantic Types

Although extracting CUIs from the text of posts provides structure to unstructured data, and features medically relevant vocabulary, not all CUIs reference comorbidities, our focus for this investigation. To elaborate, C0011011, referencing ”Daughter,” is among the top-occurring CUIs in the data set, but an author having the single-concept word of ”Daughter” in her

Proposed Age Boundary	Number of Authors
2	1061
3	1568
4	1759
5	1842
6	1806
7	1741
8	1631
9	1472
10	1343
11	1196
12	942
13	695
14	391

Table 3.2: Number of relevant authors with posts referring to an ASD subject younger and older than the proposed age boundary

post does not indicate a comorbidity of the ASD subject. Thus, we needed to further narrow the CUIs within our vocabulary of single-concept words to better reference comorbidities.

To address the issue that not all CUIs referenced ASD comorbidities, we leveraged the semantic type of CUIs. The semantic type is the basic semantic category to which a term may be assigned. The types are assigned based on the inherent properties of a concept, and occasionally based on its functional properties. For example, the semantic type ”Mental Dysfunction” is assigned to *Dementia*. A network of these semantic types accompanies the Metathesaurus, which provides a consistent categorization of all concepts represented and elucidates the permissible relationships between and among these concepts [23].

For this investigation, of the 135 semantic types, interested semantic types included ”Disease or Syndrome”, ”Mental or Behavioral Dysfunction”, ”Neoplastic Process”, ”Acquired Abnormality”, ”Age Group”, ”Behavior”, ”Congenital Abnormality”, ”Clinical Drug”’, ”Cell or Molecular Dysfunction”, ”Diagnostic Procedure”, ”Individual Behavior”, ”Mental Process”, ”Social Behavior”, and ”Sign or Symptom.” Of note, these semantic types more likely directly contained concepts associated with comorbidities and that we expected would be potent in prediction. After processing each single-concept word in the vocabulary and excluding those that did not fall under the interested semantic types, the top-occurring CUIs were as follows. This selection better captured the behaviors and conditions associated with ASD comorbidities, as evidenced by how "Abstract thought disorder" and "Temper tantrum," behaviors and syndromes within the featured heterogeneity of ASD, rose to be among the top-occurring CUIs.

CUI	Name	Frequency
C0004352	Autistic Disorder	5446
C0012634	Disease	1162
C0683607	allowing	1111
C0237876	Sharing (Social Behavior)	1082
C0424324	Fighting	1055
C0009452	Communication	901
C0233642	Abstract thought disorder	770
C0233558	Temper tantrum	716
C0001807	Aggressive behavior	714

Table 3.3: Most Frequent Semantically-Reduced CUIs

At this point, we have reduced the vocabulary of single-concept words from 3050 to 800, and we can finalize the X and Y matrices for logistic regression. With a number of authors, $|N| = 1842$, and a size of the vocabulary, $|V| = 800$, we set up the application as follows.

For X, $X_{ij} = 1$ if author i had CUI j present in her posts on an ASD subject younger than or at five years old, and $X_{ij} = 0$ otherwise. Similarly, for Y, $Y_{ij} = 1$ if author i had CUI j present in her posts on an ASD subject older than five years old, and $Y_{ij} = 0$ otherwise. Following a split of the data set, 80 % for training and 20 % for testing, we trained a logistic regression classifier for each CUI, with that CUI's column in Y being the target vector. We utilized the logistic regression classifier from the scikit-learn library with an L2 penalty, a lbfgs solver, and an inverse of regularization strength (C) of 0.1.

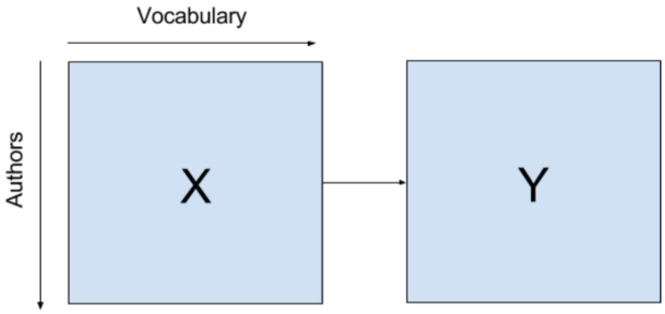


Figure 3.6: Experimental Setup

Chapter 4

Results and Conclusion

In this chapter I provide an evaluation of the logistic regression classifiers trained for each CUI and compare their effectiveness against generative classifiers like LDA and DTM. I also discuss limitations of predicting comorbidities associated with ASD from social online forums, suggest open areas for further research, and conclude.

4.1 Evaluation

Of the 800 single-concept words in the vocabulary, 512 single-concept words had more than one class in the training Y set, meaning that 512 logistic regressions could be trained because there existed at least one author who had the single-concept word in her later posts. For each of those 512 single-concept words, we then computed the log probabilities of that single-concept word's presence in each author's later posts from her earlier posts. We then computed the Area Under the Curve (AUC) from the author by word log probabilities using the roc auc score function from the sci-kit learn library.

Area Under the Curve

Area Under the Curve, a performance metric of a logistic regression, is a commonly used evaluation metric for binary classification problems. The interpretation is that given a random positive observation and negative observation, the AUC gives the proportion of the time you guess which is correct. It is less affected by sample balance than accuracy. A perfect model will score an AUC of 1, while random guessing will score an AUC of around 0.5.

Please find below boxplots displaying the distribution of AUCs for the single-concept word logistic regressions. On the right, we have the distribution of AUCs for the continuation of a single-concept topic from an author's earlier posts to her later posts. To elaborate, if our logistic regression model for single-concept words paralleled the trajectory of comorbidities in ASD, then the presence of a single-concept word in a forum post regarding an ASD subject at an earlier age should likely persist in forum posts about that subject at later ages because chronic health conditions would not disappear. As illustrated, although the AUC scores for the single-concept word logistic regression models stand at a relatively high value near 1, the AUC scores for the perpetuation of single-concept words from X to Y reveal poor performance, hovering around 0.5.

Due to the high AUC scores, we investigated further and plotted the proportion of authors in the training set that featured the single-concept word on the x-axis and the AUC score of the logistic regression model predicting that word on the y-axis. We found a distinctive upward-facing curved shape that demonstrated that the high AUC scores for predicting single-concept words stemmed from single-concept words with low frequency in the data set. Suspecting that the logistic regression models for low-frequency, high-AUC single-concept words ”memorized” rather than ”learned” features for prediction, we consequently only included single-concept words that appeared in at least 10 % of the testing documents to mediate the issue. This decreased the average AUC to about 0.75, which better reflects the performance of the logistic regression models.

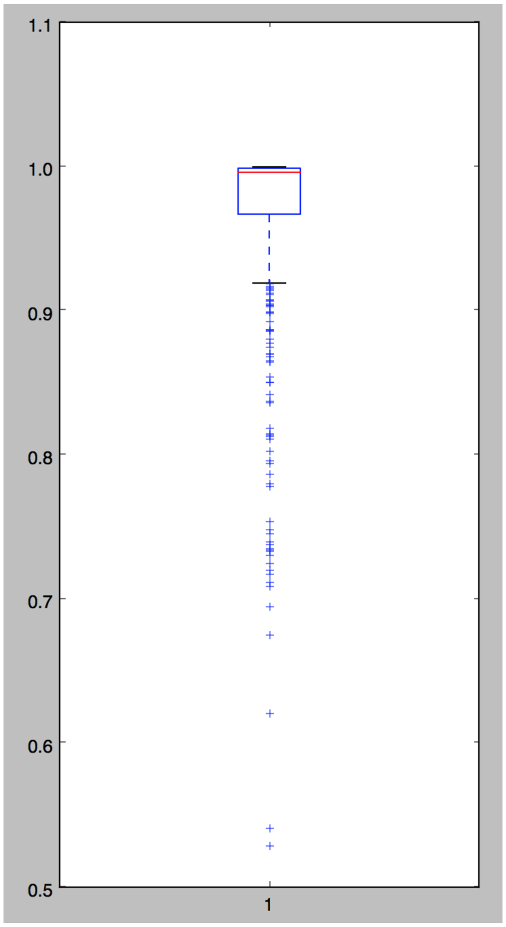


Figure 4.1: Distribution of AUCs for single-concept word predictions

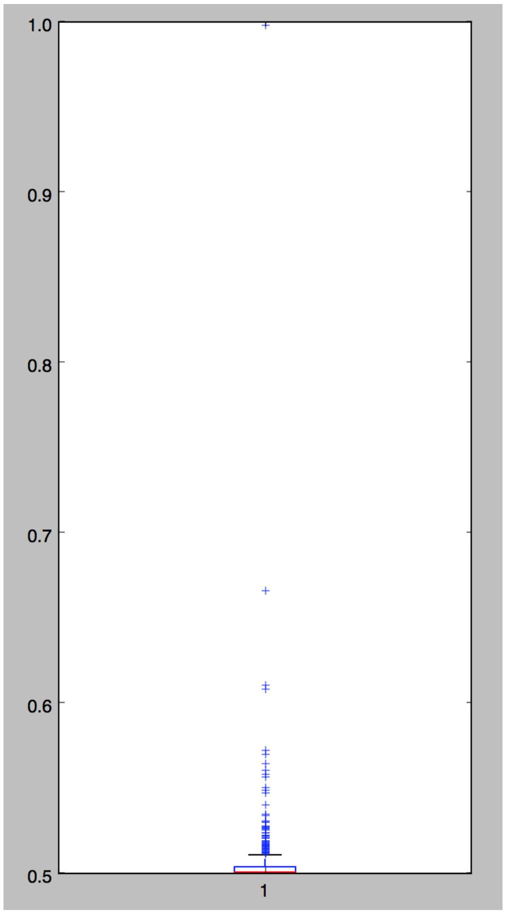


Figure 4.2: Distribution of AUCs for single-concept word persistence

Comparison to DTM

Whereas the discriminative model of LR learns the boundary between classes, generative models like DTM model the distribution of individual classes. Generative models can outperform discriminative models on smaller data sets because their generative assumptions place some structure on your model that prevent overfitting. For example, Naive Bayes assumes conditional independence of the features, while logistic regression (the discriminative ”counterpart” of Naive Bayes) does not [24].

To compare the performance of the LR and DTM on the social online forum data set and

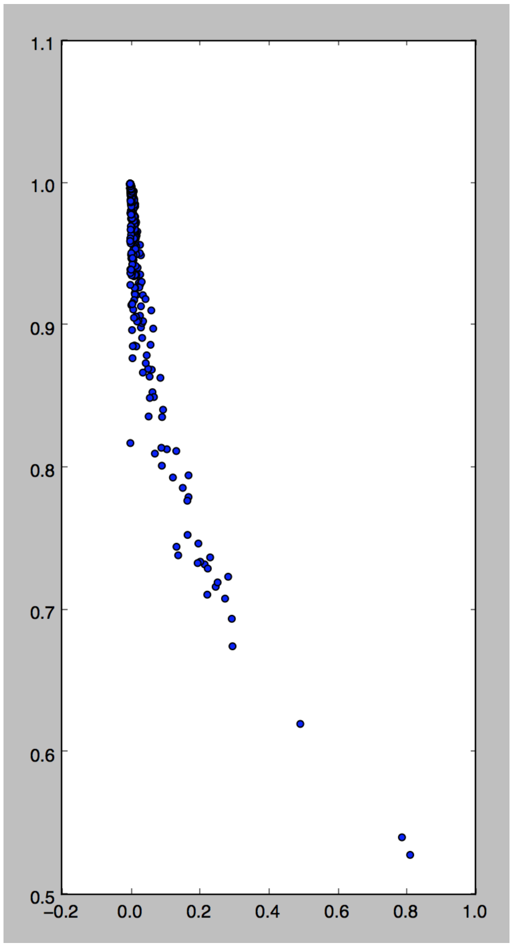


Figure 4.3: AUC of single-concept word vs. Frequency of single-concept word

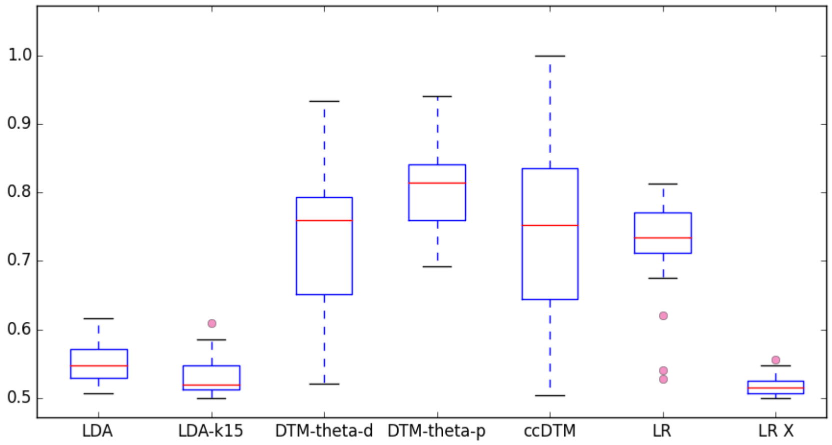


Figure 4.4: AUC of single-concept word vs. Frequency of single-concept word

then evaluate whether a discriminative or generative approach is better suited to this application of single-concept word modeling, we must first ensure that the testing set includes the same authors and same words. Whereas the log probabilities from DTM are document by word, meaning that we computed the log probability that a certain document contains a word, the log probabilities from LR are author by word, indicating the probability of whether an author includes a certain word in later forum posts. We thus converted the log probabilities of LR from author by word to document by word by replicating the LR probabilities for each author in the testing X matrix and Y matrix by the number of documents the author had written, ensuring that it's a time insensitive model. Now with document by word predictions for the document by word binary outputs for the DTM and LR, we can compare the computed AUCs for each model. Although the absolute numbers in DTM log probabilities and LR log probabilities are not directly directly comparable because they are used to compute different likelihoods, once used for an AUC, they become comparable again.

As illustrated in Figure 4.4, the LR model outperforms the LDA models, which would expected given that LDA treats documents within a corpus exchangeably and would not take into account the sequential nature of documents by an author. The DTM models significantly outperform LDA model, which would also be expected considering that topic and word distributions depend upon earlier topic and word distributions. Interestingly, the LR model outperforms the LDA models but slightly under performs against the DTM models, suggesting that the gains of DTM when applied to this corpus are significant despite that discriminative models tend better perform when labelled training data is plentiful.

4.2 Limitations

Although our approach was novel in terms of applying a discriminative model to sequential topic modeling by dividing an author's posts into earlier and later subcategories based on the inferred age of the ASD subject involved, it also faced several limitations by utilizing social online forum data in this manner. Beyond the inconsistencies in extracting age and medical concepts from unstructured text, we did not verify that the concepts extracted from the forum post indisputably pertained to the ASD subject. Furthermore, for an author with at least one post age-identified, we inferred the age of that subject for the author's other posts by comparing the time of posting. However, the concepts extracted from these additional posts may not have pertained to the ASD subject. Overall, we made several assumptions in how this process could potentially parallel the trajectories of ASD comorbidities, and those assumptions do not prove to be true, as evidenced by how few single-concept topics persisted across time for authors, which diverges from the actual trajectories of ASD comorbidities.

4.3 Future Work

This chapter presented a logistic regression model to predict single-concept words within an social online forum data set. There is ample opportunity for research to further improve this application for clinical insight. Because a discriminative approach nicely lends itself to interpretability, we can enable physicians to interpret the co-occurrences of single-concept words in this topic modeling by providing the most significant features, the words with the greatest coefficients, for each word. In addition, we can apply this problem statement to structured electronic health records, where ICD-9 diagnosis codes could serve as the vocabulary, and patient records could be divided into X and Y; X would encapsulate comorbidities present for the patient up to age five and Y would encapsulate those present for the patient after age five.

4.4 Conclusion

In this thesis, I showed that a discriminative logistic regression model to predict single-concepts in social online forum posts over time outperformed latent Dirichlet allocation but under performed against dynamic topic modeling, both generative approaches. I presented exploratory analysis of social online posts within forums related to autism, and I explained the preprocessing step of age identification and concept extraction via regular expression filters. I showed that the performance of the logistic regression for the problem statement can be quite accurate but must be re-evaluated in the context of the frequency of the word it is predicting.

Previous work has theoretically shown machine learning applications to structured and unstructured medical data to predict comorbidities, but this experimentation specifically leveraged unstructured social data to investigate the trajectory of topics for a given author. The excellent experimental performance of logistic regression warrants further research into how discriminative approaches can be applied to binary topic modeling, but overall, this thesis provided an appropriate baseline to evaluate the effectiveness and accuracy gained by the generative DTM.

Bibliography

[1] URL: `http://www.autismweb.com/`.

[2] URL: `http://www.asdfriendly.org/`.

[3] URL: `http://www.asd-forum.org.uk/forum/`.

[4] URL: `http://umlsks.nlm.nih.gov`.

[5] American Psychiatric Association. "Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR (Text Revision)". In: (2000).

[6] Division of Birth Defects. "Autism spectrum disorders: data and statistics". In: *National Center on Birth Defects and Developmental Disabilities 2012* (2014).

[7] Christopher M. Bishop. *Pattern Recognition and Machine Learning.* Cambridge, U.K.: Springer, 2006.

[8] David M. Blei. "Probabilistic Topic Models". In: *Communications of the ACM* 55.4 (2012).

[9] David M. Blei and John D. Lafferty. "Dynamic Topic Models". In: *Proceedings in the 23rd International Conference on Machine Learning* (2006).

[10] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation". In: *The Journal of Machine Learning Research* 3 (2003).

[11] Neil S. Coulson, Heather Buchanan, and Aimee Aubeeluck. "Social support in cyberspace: A content analysis of communication within a Huntington's disease online support group". In: *Patient Education and Counseling* 68.2 (2007).

[12] Finale Doshi-Velez, Yaorong Ge, and Isaac Kohane. “Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis”. In: *Pediatrics* 133.1 (2014).

[13] Daniel H. Geschwind and Pat Levitt. “Autism spectrum disorders: developmental disconnection syndromes”. In: *Current Opinion in Neurobiology* 17 (2007).

[14] Bianca E. Himes et al. “Prediction of Chronic Obstructive Pulmonary Disease (COPD) in Asthma Patients Using Electronic Medical Records”. In: *Journal of the American Medical Informatics Association* 16.3 (2009).

[15] National Cancer Institute. “Health Information National Trends Survey”. In: *http://hints.cancer.gov* (2012).

[16] Shafali S. Jeste and Daniel H. Geschwind. “Disentangling the heterogeneity of autism spectrum disorder through genetic findings”. In: *Nature Reviews Neurology* 10 (2014).

[17] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Englewood Cliffs, New Jersey: Prentice Hall, 1999.

[18] Ronald C. Kessler et al. “How well can post-traumatic stress disorder be predicted from pre-trauma risk factors? An exploratory study in the WHO World Mental Health Surveys”. In: *World Psychiatry* 13 (3 2014).

[19] Isaac S. Kohane et al. “The Co-Morbidity Burden of Children and Young Adults with Autism Spectrum Disorders”. In: *PLoS ONE* 7.4 (2012).

[20] J. A. Lasserre, C. M. Bishop, and T. P. Minka. “Principled Hybrids of Generative and Discriminative Models”. In: *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference* (2006).

[21] DA Lindberg, BL Humphreys, and AT McCray. “The Unified Medical Language System”. In: *Methods of information in medicine* 32.4 (1993).

[22] Li Liu, Jing Lei, and Kathryn Roeder. “Network Assisted Analysis To Reveal The Genetic Basis of Autism”. In: *The Annals of Applied Statistics* 9.3 (2015).

[23] Alexa T. McCray. “The UMLS Semantic Network”. In: *Proc Annu Symp Comput Appl Med Care* (1989).

[24] Andrew Y. Ng and Michael I. Jordan. “On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes”. In: *NIPS-14* (2002).

[25] Che Ngufor et al. “Extreme Logistic Regression: A Large Scale Learning Algorithm with Application to Prostate Cancer Mortality Predictions”. In: *Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference* (2014).

[26] Roy H. Perlis. “A Clinical Risk Stratification Tool for Predicting Treatment Resistance in Major Depressive Disorder”. In: *Biological Psychiatry* 74 (1 2013).

[27] Chris Poulin et al. “Predicting the Risk of Suicide by Analyzing the Text of Clinical Notes”. In: *PLoS ONE* 9 (3 2014).

[28] S. A. Rains and D. M. Keating. “The social dimension of blogging about health: Health blogging, social support, and well-being”. In: *Communication Monographs* 78 (2011).

[29] Stephen A. Rains, Emily B. Peterson, and Kevin B. Wright. “Communicating Social Support in Computer-mediated Contexts: A Meta-analytic Review of Content Analyses Examining Support Messages Shared Online among Individuals Coping with Illness”. In: *Communication Monographs* 82.4 (2015).

[30] Amit Saha and Nitin Agarwal. “Modeling social support in autism community on social media”. In: *Network modeling and analysis in health informatics and bioinformatics* 5.8 (2016).

[31] Arthur L. Samuel. “Some Studies in Machine Learning Using the Game of Checkers”. In: *IBM Journal* 3 (1959).

[32] Catherine Arnott Smith and Paul J. Wicks. "PatientsLikeMe: Consumer Health Vocabulary as a Folksonomy". In: *AMIA Annual Symposium Proceedings 2008* (2008).

[33] Jose M. Valderas et al. "Defining Comorbidity: Implications for Understanding Health and Health Services". In: *Annals of Family Medicine* 7.4 (2009).

[34] Q. T. Zheng. *Consumer health vocabulary initiative*. 2014. URL: `http://consumerhealthvocab.org`.