Using Topic Models to Investigate Depression on Social Media

William Armstrong
University of Maryland
armstrow@umd.edu

Abstract

In this paper we explore the utility of topic-modeling techniques (LDA, sLDA, SNLDA) in computational linguistic analysis of text for the purpose of psychological evaluation. Specifically we hope to be able to identify and provide insight regarding clinical depression.

1 Introduction

Topic modeling is a well-known technique in the field of computational linguistics as a model for reducing dimensionality of a feature space. In addition to being an effective machine-learning tool, reducing a dataset to a relatively small number of "topics" makes topic models highly interpretable by humans. This positions the technique ideally for problems in which technology and domain experts might work together to achieve superior results. One such application is in identifying and monitoring mental health disorders like depression.

Clinical psychology in practice tends to be hampered by insufficient data gathered from patients in only a few hours of conversation per week in a controlled environment that may bias the information gains. Fortunately, the modern age of social media presents an abundance of individuals sharing their inner thoughts naturally and constantly on sites such as Twitter for anyone with the patience to analyze it. Potentially, topic modeling can both automatically monitor and identify social media users at risk of depression while simultaneously summarizing the findings for the clinicians who may be treating those patients or studying the disease.

For example, a patient could give their clinician access to their Twitter feed, which would be monitored during the time between sessions to identify periods of higher depression and subjects that may be associated with such periods. For example, if schoolwork is identified as a troublesome subject for the patient the clinician can investigate whether homework or grades may be triggering depressive episodes and propose appropriate actions for such. Furthermore, these tools can scale to larger populations to both identify individuals in need of treatment and discover general trends in depressive behaviors.

This paper will explore the utility of three topic modeling techniques in these objectives: latent-Dirichlet allocation (LDA) in its base form, with supervision (sLDA), and with supervision and a nested hierarchy (SNLDA). We will examine the results of each topic modeling technique qualitatively by the potential usefulness of their posterior topics to a clinician as well as quantitatively in their ability to classify text as indicative of depression or not.

2 Related Work

Several recent papers have similarly attempted to identify depression and other mental health disorders through natural language processing of social media. Although some have included features from topic modeling, none have done so as extensively.

Among the groundbreaking papers was research from De Choudhury et al. (2013) at Microsoft Research, who used crowdsourcing to build a comprehensive dataset of posts on Twitter ("tweets") along with psychological evaluation questionnaires and a survey regarding each user's history of clinical depression. She built supervised learning models to detect depression with a basic set of features; how-

ever the experiments were done over a relatively small population (476 users) that will not necessarily scale well to a larger group of subjects.

Schwartz et al. (2014) attempted a similar task using data from Facebook obtained in the MyPersonality project, which includes user responses to clinical psychological evaluation surveys.[1] They were able to identify depression on a larger scale than De Choudhury et al. (2013), and track its temporal and geographic trends in a larger population.

Coppersmith et al. (2014) collected an alternative dataset using a system that searched English language tweets for individuals posting that they had been diagnosed with a mental health disorder, including depression, and collected all that user's tweets for about two years surrounding this self-reporting post. Though this dataset has a large scale and does not rely on psychological questionnaires, some users in the control set may also have depression but did not post about it on Twitter during the time frame explored. Furthermore, it does not track the rise or fall of depression symptoms in a user over time. Nonetheless, Coppersmith et al. (2014) were able to classify depression and several other common mental health disorders. Our objective is to improve on their predictive results on the same dataset by utilizing more refined language analysis tools.

The experiments in this paper build directly on the work in Resnik et al. (2013), which produced topic models useful in analyzing neuroticism. These models were built from stream-of-consciousness essays correlated with a Big-5 personality score for the author's degree of neuroticism (John et al. (2008)), collected by Pennebaker and King (Pennebaker and King (1999)). Though that work suggested the utility of LDA in analysis of that particular dataset, our work deals with the additional considerations from the nature of text in social media–specifically its scale, topical variety, and brevity.

Additionally, the work presented here is a direct expansion of the work we presented in Resnik et al. (2015b) and Resnik et al. (2015a). In particular, we will present the LDA and sLDA models of those papers more comprehensively while expanding and continuing the work with SNLDA in Resnik et al. (2015b).

[1] See http://www.mypersonality.org

3 Data

3.1 Pennebaker Essays

The Pennebaker and King (1999) dataset consists of 6,459 stream-of-consciousness essays (˜780 words each) collected over the course of a decade from college students in Texas. These students were then evaluated by a questionnaire to obtain a set of "Big-5" personality scores for each document, of which we will focus on the *neuroticism* personality trait, characterized by emotional instability, anxiety, and *depression* (Matthews et al. (2003)). Since the authors of the essays were college students, they can reasonably be assumed to be in roughly the same demographic as our Twitter dataset (most Twitter users are under 50 and college-educated according to Pew Research Center (2014)) and both datasets appear to share a similar informal vocabulary.

This dataset was chosen because it is relatively "clean", includes scores from a validated psychological instrument, and was successfully used in a similar computational psycho-linguistic experiment (Resnik et al. (2013)). It was therefore useful to tune the topic-modeling techniques in an environment with less noise before applying them to the larger Twitter dataset.

3.2 Twitter Posts

Derived from the collection of Coppersmith et al. (2014), the second dataset we used was a set of anonymized tweets used in Coppersmith et al. (2015). It consists of approximately two million tweets (140 characters or less each) from 869 users (3,000 or fewer tweets per user), of whom 314 self-identified as having been diagnosed with depression. As discussed in Coppersmith et al. (2014), self-identification means a user publicly tweeted something along the lines of "I was diagnosed with depression today", with some manual validation by the individuals preparing the data. As mentioned in Section 2, this means some users in the depression set may not have been truthful about their diagnosis and some individuals in the control set ("control users") may have been diagnosed with depression without mentioning it in their public Twitter account. For simplicity in the rest of this paper we refer to individuals from the depression set as "depressed users" but emphasize that we know nothing absolute about

the current mental health state of these users, merely that they claimed to have been diagnosed as having depression at some point. Even assuming the claim is truthful, it could be a minor case, or one that was quickly resolved, or even one that was misdiagnosed by a clinician. However, for our classification purposes we assume such instances to be "noise" in the dataset, which does not prevent us from observing general trends.

3.3 Data Preparation[2]

Full details of the pre-processing of the data can be found in Resnik et al. (2015a) and Resnik et al. (2015b). In summary, we performed basic sanitizing of the data (removing stop-words and words with special characters), then lemmatized the words and filtered them according to the number of documents they appear in.[3] Since initial experimentation with a similar Twitter dataset showed topic modeling to be less effective on either very short individual tweets or very long aggregations of tweets by author, we chose to define a document as a concatenated set of tweets for a particular author during a particular week, as was done in Resnik et al. (2015a). We built a shared vocabulary for all systems to ensure uniformity across experiments and to allow them to build off of each other (see Section 4.4). The vocabulary for all systems consisted of any lemmas found in more than 100 Twitter documents or more than 5 student essays. About 48% of this combined vocabulary was shared between the two sets and roughly 14% came from the essays, 38% from the tweets.

Data splits The tweets were divided into a training and testing set, using 80% of the users for training and 20% for testing, in order to evaluate the performance of prediction models on this set.[4] All models were built exclusively from the training set and the the true labels of the test users remained hidden except for use in evaluation.

[2]Pre-processing work represented in this section was done in close collaboration with Thang Nguyen in Resnik et al. (2015a).

[3]Though most standard emoticons were removed, some unicode emoticons were later discovered to still be present in the data and are simply identified here with the tag "EMOJI".

[4]The testing set here is identical to the "development" set in Resnik et al. (2015a), so it was not used for training in that experimentation and remains a valid test set.

4 Methods

4.1 Latent Dirichlet Allocation (LDA)

System description LDA, introduced in Blei et al. (2003), is a model used for unsupervised dimensionality reduction of datasets that is typically applied to a corpus of text. It is a generative model that assumes text has been produced from a discrete distribution of words known as a "topic", which together form a document containing a probabilistic distribution of topics. This allows us to represent a document with only K topics instead of V words. Note that since a document is just a collection of words, LDA does not take in to account the ordering of these words, just their frequency.

Blei et al. (2003) defines the generative model for each document of N words, $\mathbf{w} =< w_1, w_2, ..., w_N >$, in a corpus D with K topics as follows:

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$, where θ is a K-vector Dirichlet random variable defining the topic mixture, and $\text{Dir}(\alpha)$ is a Dirichlet distribution parameterized by the K-vector α.
3. For each of the N words w_n:
   (a) Choose topic z_n from $\text{Multinomial}(\theta)$.
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n and the $K \times V$ matrix β where $\beta_{ij} = p(w^j = 1|z^i = 1)$.

Figure 1: Generative process of LDA. Reproduced from Blei et al. (2003).

The probability of generating a particular document is therefore:

$$p(\mathbf{w}|\alpha, \beta) = \int_\theta p(\theta|\alpha) \prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \; d\theta \tag{1}$$

The corpus probability is the product of all document probabilities.

Since the observed variable here is the corpus of text and the "latent" topics that generated it are the

desired output, these must be approximated through an inference algorithm. This is commonly done with Gibbs Sampling, a Markov-chain Monte Carlo technique where each dimension of a distribution is sampled alternately while all others are fixed (Heinrich (2005)).

The objective of Gibb's sampling is to infer the topics for a document, represented by the "latent" parameter $\mathbf{z}$. We obtain these with the equation below (leaving out the hyperparameters θ, α, and β):

$$p(\mathbf{z}|\mathbf{w}) = \frac{p(\mathbf{z}, \mathbf{w})}{p(\mathbf{w}} = \frac{\prod_{i=1}^{W} p(z_i, w_i)}{\prod_{i=1}^{W} \sum_{k=1}^{K} p(z_i = k, w_i)}$$

W is the the sequence of all words in the corpus, which makes the denominator too large to compute directly. Since Gibb's sampling uses a Markov assumption, it instead approximates $p(\mathbf{z}|\mathbf{w})$ with $p(z_i|\mathbf{z}_{\neg \mathbf{i}}, \mathbf{w})$, which is drawn from the result of the update step:

$$p(z_i = k|\mathbf{z}_{\neg \mathbf{i}}, \mathbf{w}) \propto \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_t} \cdot \frac{n_{m,\neg i}^{(k)} + \alpha_k}{[\sum_{k=1}^{K} n_{m,\neg i}^{(k)} + \alpha_k] - 1} \quad (2)$$

Here m is the index of the current document, n is the index of the current word in that document, t is the index of that word in the vocabulary, and $\neg i$ indicates excluding item i from a set. The sampling process iterates over each document and then each word of the document, updating the assignment of a topic to that word by re-sampling from $p(z_i|\mathbf{z}_{\neg \mathbf{i}}, \mathbf{w})$ according to Equation 2, then using that assignment to change the appropriate counts of words assigned to topics and topics assigned to the document. This iteration is done until convergence is reached and the final distribution of topics for each document and words for each topic is output (Heinrich (2005)).

4.2 Supervised LDA (sLDA)

For datasets that contain additional information in the form of a response value (e.g. rating level from online product reviews, grades on student essays), Blei and McAuliffe (2007) introduces sLDA to guide the topic inference of LDA towards more effective topics with respect to a prediction goal.

This generative model is very similar to that of LDA except it includes an additional random draw to determine the response value $\mathbf{y}$, which is drawn from a normal distribution based on the response hyperparameter η and the document's topic distribution $\bar{z}$. This response value is based on the topic frequencies in the generated document, so it is considered completely separate from the unknown topic distribution generating those topics, which means the rest of the document generation proceeds the same as in LDA (see Figure 1). The additional step is:

3. Draw response variable $y \sim \mathrm{N}(\eta^\top \bar{z}, \sigma^2)$ where $\bar{z} = (1/N) \sum_{n=1}^{N} z_n$

Figure 2: Generative step for response value in sLDA. Reproduced from Blei and McAuliffe (2007), continued from Figure 1.

As in LDA, the exact posterior cannot be computed and must be approximated. The only difference from the LDA algorithm is an additional step to update the current response values based on the current topic distributions and the inferred values of η and σ^2.

This model becomes particularly useful as it can now predict a response value for unseen documents without having to rely on a classifier learning the topic-posterior features. Instead, we use the mean of the expected response value distribution for the document as follows:

$$E[\mathbf{Y}|\mathbf{w}, \alpha, \beta, \eta, \sigma^2] \approx \eta^\top E[\bar{z}]$$

4.3 Supervised Nested LDA (SNLDA)

We hypothesize that the results from sLDA could be further improved by introducing a layered hierarchy to the topics as proposed in Nguyen (2015) and explored briefly in Resnik et al. (2015b). For instance, if sLDA inferred a general topic about *sports* associated with a single response value, SNLDA could infer that same topic with additional subtopics like *soccer*, *hockey*, or *basketball*, each with a response value of its own that may vary greatly from the parent's. This tree is referred to as τ and has fixed dimensions.

The generative process for SNLDA extends that of sLDA by requiring the tree τ to be generated first.

This is done by drawing a topic ϕ_k and regression parameter η_k for each node k in τ as follows:

1. If k is the root, $\phi_k \sim \text{Dir}(\beta_0 \boldsymbol{u})$ and set $\eta_k = 0$
2. If k is first-level, $\phi_k \sim \text{Dir}(\beta_1 \phi_k^\star)$ and $\eta_k \sim \mathcal{N}(0, \sigma_1)$, where $\phi_k^\star$ specifies either an informed or a symmetric uninformed prior.
3. Otherwise, $\phi_k \sim \text{Dir}(\beta_{l_k} \phi_{p_k})$ and $\eta_k \sim \mathcal{N}(0, \sigma_{l_k})$, where l_k and p_k are the level and the parent of node k respectively.

Figure 3: Generative model for τ in SNLDA. Provided by Viet-An Nguyen

To generate a document we then consider, instead of its distribution across all topics, its distribution across all child topics for a given node in the tree. The generation process for a document d is:

1. For each non-terminal node k in the tree
   (a) Draw a distribution over k's children $\theta_{d,k} \sim \text{Dir}(\alpha_{l_k})$
   (b) Draw a stochastic switching variable $\omega_{d,k} \sim \text{Beta}(\pi\gamma_{l_k})$
2. For each token $n \in [1, N_d]$
   (a) Draw a node $z_{d,k} \sim \mathcal{B}(\boldsymbol{\theta}_d, \boldsymbol{\omega}_d)$
   (b) Draw $w_{d,k} \sim \text{Mult}(\phi_{z_{d,k}})$
3. The probability of the binary response to be 1, is
$$p(y_d = 1) = \Phi\left(\sum_{k \in \mathcal{T}} \frac{N_{d,k}}{N_{d,\cdot}} \eta_k\right)$$
where $\Phi(x) = \exp(x)/(1 + \exp(x))$ is the logistic function, $N_{d,k}$ is the number of tokens in document d assigned to node k, and marginal count is denoted by '$\cdot$'.

Figure 4: Generative model for each document in SNLDA. Provided by Viet-An Nguyen

Note that the tree is traversed by means of a "stochastic switching variable", $\omega_{d,k} \sim \text{Beta}(\pi\gamma_{l_k})$, at each level, which determines whether to use that node or select again from its children.

To infer an SNLDA model based on a given corpus we again need to use an approximation algorithm. This time, at each iteration we (1) assign each word in a document to a node in τ, (2) sample topics for nodes along the path to the node selected in (1), and (3) optimize the regression parameters, η. The selection of a node for a word in (1) is done by selecting a node at the current level through sampling from a conditional probability distribution, and then choosing to either stay at that node or recursively sample from its children based on the number of child nodes and number of tokens in the document assigned to that node and/or its subtree.

4.4 Informative priors

Wallach et al. (2009) propose that an improvement can be made over traditional topic modeling by drawing topic distributions for documents from an asymmetric prior instead of the typical symmetric Dirichlet priors with a "heuristically set" α hyperparameter. We take this a step further by incorporating an "informed prior" (also later referred to as simply "prior"), which is a known good topic distribution from which the corpus may have been sampled.

In the generative models above, a topic distribution, θ, for a document is drawn (see step 1 of Figure 1) from a Dirichlet distribution with the concentration parameter α built on an implied symmetric distribution, the "uninformative prior". Such a prior gives all topics an equal weight without any biases for or against particular topics, which is good if nothing is initially known about the underlying topic structure. However, if something is known about the underlying structure we can instead define an informed prior from which to build the Dirichlet distribution. This prior is built from the "known" topics (defined by their words in a V-vector), whose influence in the resulting topics can now be tuned with the α hyperparameter.

This means that the topic distribution selected for a document is more likely to represent the topics from the prior, although the iterative nature of the algorithm still allows for the topics to converge on other strong topical signals from the data. Note that for SNLDA the prior only affects the level 1 nodes, since the root node's topic distribution is drawn from a Dirichlet distribution over a given prior and the other levels are drawn from a distribution with a

prior based on their parent node (see Figure 3, noting that the hyperparameters there are β_l where l is the level).

Initial experiments with the Twitter data revealed a major difficulty in working with social media, which is that users do not generally post explicitly about mental health topics, such as feelings, but instead write about surface activities like school, work, hobbies, and so forth. Thus topic models derived directly from the data would be more likely to identify trends in conversation topics (like sports teams, music groups, or current events), which merely reveals which of those topics depressed users tended to discuss. Therefore, to get the kind of clinically significant topics we desired, like the ones we found in the Pennebaker essays, we used the posterior topic distributions from those models to construct an informed prior that would drive the Twitter topics in the desired direction.

Though using the prior from the Pennebaker essays helped mitigate some of the difficulties of working with the social media data, forcing the two to share a vocabulary may have lost some of the value added by unconventional forms of communication used in social media (ie emoticons, abbreviations, etc.). Some information "lost in translation" is also possible due to the different document sizes and population characteristics. On the whole, however, the two datasets are similar enough to be be useful, and the results reinforce this.

5 Qualitative Results

For the "human-intuition" component of topic modeling, we ran each of the models on the datasets and for each topic in the resulting topic posterior identified the 20 words with the strongest weight, which are typically considered to be the words that best identify that topic. These posteriors were the topic-word distributions averaged over samples from 500 iterations of sampling (excluding the first 100 as burn-in).

5.1 LDA

We ran LDA with parameters: number of topics (k) = 50, document-topic Dirichlet hyperparameter (α) = 1, topic-word Dirichlet hyperparameter (β) = 0.01, in order to best match the experiments from Resnik et al. (2013).

Pennebaker essays The application of LDA to the Pennebaker dataset in order to produce insight about neuroticism and depression was explored in Resnik et al. (2013). Though their use of LDA was not particularly complicated, its application to this dataset provided coherent and relevant results, the utility of which was confirmed by a clinical psychologist.

We replicated this work in our own experiments using a different implementation of LDA and the vocabulary from the combined, preprocessed datasets. A rough comparison of the top 20 words in the resulting topics showed strong correlation with Resnik et al. (2013) in that 35% of the top words were found in both listings (for comparison, the Twitter topics showed only a 18% overlap) and 45 of the topics had over five or more of their top words in the top words of a topic from Resnik et al. (2013). Seven of the topics were particularly similar, with over half of the top 20 words shared between both models. These are shown in Table 1. This is especially notable because the vocabulary used in our work was substantially different from that of Resnik et al. (2013), being much smaller and using lemmatization.

Twitter Though the LDA topic modeling with a traditional symmetric prior (which we will refer to as "uninformed LDA") appeared to work well in its role of summarizing the Twitter dataset, the topics do not appear to be useful in providing intuition to a clinician. Instead, as expected, we see discussion of things like *politics*, *celebrities*, *pets*, *kids*, *relationships*, and *sports*. Some of the topics that appeared to be more relevant to mental health are reproduced in Table 2.

Twitter with informed prior The influence of informative priors on LDA appears through another rough comparison of top words to the top words of the prior, which yields a 42% correlation–a 20% increase from the Twitter topics alone. Though there are a few of the same topics as in uninformed LDA (ie *sleep*, and *food*), some other useful topics about *relationships* and *emotions* also appear.These topics might be more useful to a clinician. The topics most similar to those of the prior are shown in Table 3.

Clinician Label	Top 20 words
FUTURE ACTIVITIES	weekend week day home time friday haven boyfriend san homework austin lot drive saturday hour sunday plan month antonio tomorrow
VEGETATIVE ENERGY LEVEL	sleep night morning wake class bed hour late tomorrow fall start asleep nap bus shower yesterday sleepy awake monday friday
FOOD	eat food hungry dinner lunch weight cook pizza ice gain cream buy stomach chicken meal taste gym fat watch lose
MUSIC	music song listen play band sing sound remind guitar hear favorite concert rock cool awesome radio lyric voice beat ipod
E-MAIL AND PAPERS	computer time email lab internet sit type check guess line assignment library suppose day screen wait send write paper online
IMMATURE	yeah wow haha minute guess funny cool suck weird stuff write hmm gosh type ugh lol fun freak crazy lot
ANGER/FRUSTRATION	damn hell suck shit stupid fuck crap piss time stop real screw care lose bad stick girl hot hey blah

Table 1: LDA topics most similar to Resnik et al. (2013), along with their clinician's label for the topic

Top 20 words
love text miss phone boyfriend call hair talk sister feel sleep bore wake wanna house tomorrow stop walk die picture
dont people hate talk feel ill stop youre didnt ive care yeah call bad whats wont girl school doesnt laugh
eat food drink lose chicken weight skinny cheese share fat start water pizza month workout fry cook bacon dinner body
game play team win time football tonight goal season hit player ball sport score wow leave field baseball pick half
god life jesus love lord bless pray word heart day church live change woman family friend faith prayer christ peace
school class college study camp hike teacher day test homework math fall senior student grade life friday kid exam home
hope time lovely forward excite enjoy lot luck london film nice amaze weekend train race glad book idea sort meet
feel sleep tire bed hard wake bad doe stay hate start happen sick hurt pain mine morning mind lot forever

Table 2: Topics from uninformed LDA on Twitter that were most similar to the Pennebaker topics

5.2 SLDA

Our sLDA experimentation used parameters: $k = 50$, $\alpha = 1$, $\beta = 0.01$, Gaussian variance for document responses $(\rho) = 1$, Gaussian variance for topic's regression parameters $(\sigma) = 1$, and Gaussian mean for topic's regression parameters $(\mu) = 0.0$. The response value for both sets was treated as continuous, despite the binary label of the tweets, for uniformity between them and ease of interpreting the results based on confidence. Note the number of topics remains fixed to allow for the use of the LDA topics as the prior and for better comparison between models.

Pennebaker essays Table 4 shows that the work in Resnik et al. (2013) is greatly enhanced by the introduction of supervision by the essay's neuroticism score. The topics appear to be as coherent and meaningful as those from LDA, but now they are also easily arranged in order of their regression parameter. At the top of Table 4, indicating a high neuroticism response, there are emotional topics dealing with *stress*, and *relationship issues*. For a low neuroticism response we see activities such as *sports*, *academics*, and *music*. It was shown in Resnik et al. (2015b) that the results obtained by the algorithm in this case correlated well with the intuition of our clinical psychologist collaborator as well.

The quality of the topics and response values indicated by sLDA on this data suggest it would also be a good candidate to use as an informed prior for the tweets, since the prior could then also be leveraged on learning the response values associated with each topic. This exercise has been left for future work.

Twitter Documents in the Twitter dataset were labeled with a binary value to indicate which set (depression or control) it belongs to, which was treated as a numeric response in the system. Then, for the continuous response values of y in the final sLDA model (see Figure 2), larger numbers indicate higher confidence the document belongs to the depression category. This is reasonable since depression can occur in various degrees.

As shown in Table 5, running sLDA on the Twitter set with a symmetric, uninformed prior ("uninformed sLDA") demonstrated improvement over both LDA models and the response values are helpful in identifying the most useful topics. The topics themselves appear to be stronger than the ones identified by LDA and the model also identified some new highly relevant topics. For instance, the topic in Table 5 at response value 3.536 did not exist in the LDA models and appears to be highly intuitively associated with depression. Another interesting example is the appearance of a *baby* and *pregnancy* related topic at regression value 2.877, which could indicate signals of postpartum depression in the dataset.

Twitter with informed prior The technique of using sLDA on the Twitter data utilizing an informed prior from LDA of the Pennebaker data was chosen for use in Resnik et al. (2015a) and selected topics from that work are reproduced here in Ta-

Top 20 words
baby birthday cute dad mom sister friend dog miss wait home goodnight cat brother parent family cousin buy sweet drive
day happen week call break start month feel bad ago guess past pain couple nice plan worry fine green hour
happy omg hope justin hot wait tomorrow nice tonight time bae rain cool fun stop tire miss sleep eye glad
people hate person care time bad feel act ignore sleep car wrong true reason nice avi judge stand happen conversation
friend yeah talk meet stay mine summer fun lot school hang short drake bby close learn touch shower EMOJI person
time wanna miss forget lot remember happy hard doe head mind fast haven weird leave kinda bite stuff bad spend
guy girl boy babe date meet single break night mum gorgeous girlfriend kiss katy ugly honestly jealous marry boyfriend chill
night sleep wake tomorrow bed wait morning hour feel till nap weekend asleep late damn excite awake miley haven start
eat food pizza fat chocolate chicken dinner lunch drink cook cream hungry skinny cheese ice weight favorite luv wine breakfast

Table 3: Topics from informed LDA that were most similar to the Pennebaker topics

Regression value	Top 20 words
2.112	time guess doe hate lose feel bad stop bother hope care start run weight don scar wrong haven change fast
1.812	feel worry nervous time relax depress sad lonely comfortable stress feeling afraid reason anxious anymore overwhelm guilty pressure effort frustrate
1.131	hate damn stupid suck hell shit fuck bad blah doe piss crap freak screw bitch care real god lie kick
0.912	relationship time feel happy happen person mind life feeling cry break past girlfriend reason understand trust hurt close depress hold
0.831	friend people meet talk lot roommate hang school close person stay conversation haven friendship huge shy comfortable mine suppose surprise
0.827	time worry hard lot stress start feel school focus trouble harder easier easy figure happen realize frustrate concentrate hop constantly
...	...
0.181	day class time hour start wait sit bus schedule tomorrow homework half campus pick yesterday decide late finish everyday ride
0.167	drink water sick start sit smoke feel taste bite light stop coffee thirsty sound mouth body beer bottle smell teeth
0.122	life live change time person future grow rest realize day decision goal choice plan chance choose moment situation regret everyday
-0.007	eat food hungry dinner cook lunch roommate pizza smell jester ice cream chicken tonight meal stomach homework buy breakfast yesterday
-0.013	yeah wow minute haha funny guess type hmm gosh fun bore lol cool hey yay hmmm yea ugh min suck
-0.018	people understand person care doe act wrong strange change sense realize matter opinion expect reason waste completely accept totally selfish
...	...
-0.917	study test homework week chemistry tomorrow class hard due quiz thursday psychology calculus hour exam finish start biology monday grade
-1.044	dont ill kinda write stuff bad alot leave talk hope love hard start roomate havent fun call gonna didnt min
-1.105	weekend home week austin night drive houston plan stay apartment excite visit leave friday haven saturday sunday hurricane fun town
-1.142	music song listen play band sing hear remind guitar sound change roommate favorite rock concert awesome radio lyric amaze video
-1.796	game play football team win watch ticket run sport texas practice basketball soccer player lose season excite fan tennis coach
-2.039	guess pretty lot time nice bite stuff fun start sort haven figure couple cool alright surprise easier expect awhile bunch

Table 4: Most extreme and neutral sLDA topics from Pennebaker dataset. More positive regression parameters indicate a stronger association with a high Big-5 neuroticism score

Regression value	Top 20 words
4.319	omg cry love gonna demi cute guy feel perfect meet idk tweet omfg pls god wanna song literally bye ily
4.318	people woman doe person human kid word read child understand happen world joke remember real reason write stop change wrong
4.29	fuck shit bitch smoke hate people drink gonna sex damn fuckin dick suck wtf weed life hell feel piss stupid
3.536	feel eat die fat cut hate lose people line cry stop body care cross friend sick hurt life scar start
3.394	home watch week time wait day bed hour cat tomorrow feel call morning friend hope leave buy sleep night ago
3.093	girl guy boy people friend cute mom wear hot hate school life wanna date picture talk boyfriend kiss literally pretty
2.78	week post baby inbox month hey day ago start pregnant feel time pain girl private boy bad doe period child
...	...
0.148	lmao lol talk girl lmfao text love tho baby miss bae phone wanna mad shit fuck call damn bitch oomf
0.062	hair buy nail love dress wear red color blue cute beautiful fall pink black eye flower shoe beauty pretty spring
-0.042	photo post facebook photoset share tumblr skinny picture time update tag pic life timeline day story repost month video challenge
-0.043	happy birthday love day hope guy reason babe miss start nice stop time life night literally bad alive world song
-0.066	EMOJI EMOJI EMOJI EMOJI EMOJI EMOJI EMOJI EMOJI EMOJI girl EMOJI EMOJI EMOJI EMOJI love EMOJI EMOJI people EMOJI wanna
-0.083	game video iphone apple app play add youtube ipad google phone note update internet review free super galaxy playlist pro
...	...
-1.432	nigga shit bitch hoe bout gotta real wanna ima tho aint smh damn lil wit tryna money call cuz female
-1.594	sleep wanna feel gonna hate bed tire wake love day miss baby time people text hungry annoy hair bad nap
-1.837	guy love pic miss hey luv wait die plz hope smile tweet watch true wat soo fan sweet cont day
-1.959	school class tomorrow day college teacher homework study start test hate hour home math sleep people sit friday senior grade
-2.348	lol lmao money yea smh damn dat gotta yal ppl kid time dont dude remember jayz baby lil hell woman
-2.742	night tonight tomorrow time miss wait party weekend summer ready home drink hour week saturday excite friend gonna fun leave

Table 5: Most extreme and neutral sLDA topics from Twitter training data. More positive regression parameters indicate a stronger association with the self-reported depression set

Regression value	Top 20 words
5.362	fuck shit bitch sex smoke dick drink girl damn fuckin suck weed wanna life wtf hell gonna gay hate drug
4.702	omg cute cry gonna god guy demi idk literally feel wow hot pretty dont bye perfect pls ugh omfg laugh
4.204	line feel people cross friend comment doe start time link mental depression life live health submit deal talk lot issue
3.132	watch movie time episode read write season totally book favorite play character awesome scene star stuff cool horror start hug
2.877	week post baby inbox month day hey pain ago pregnant hun girl start doe bad boy feel time ive private
...	...
0.011	girl love kiss boy miss guy talk wanna friend cute baby relationship eye date girlfriend mine text fall call hold
0.01	justin tweet bieber mtvhottest fan meet tour win song time guy omg country picture vote award trend miley read idol
-0.016	video iphone add apple app youtube phone note free super hand camera screen usa update amaze trailer galaxy air shot
-0.025	video music live vote artist song listen album support watch boy tune official reason tonight performance download stop shoot perform
...	...
-1.595	sleep night time bed day wake feel hour tire gonna tomorrow home tonight nap morning miss wait ready bad stay
-1.689	food tonight truck night bring android party dinner tomorrow weekend awesome island game free wine lunch bar complete jack live
-1.87	nigga shit bitch hoe bout real tho gotta ima aint money lil wit bruh tryna mad yall damn ppl smh
-2.584	lol lmao damn smh yea gotta hell dude gon tho watch baby lmfao EMOJI wtf black bro idk boo funny
-2.966	car weekend home house drive summer miss week beach family rain weather run dog ready leave cancer race ride hour
-3.017	haha hahaha yeah hahahaha time night hahah wait watch ill love feel drink dad brother sleep phone sister eat miss

Table 6: Most extreme and neutral sLDA topics from Twitter training data with informed prior. More positive regression parameters indicate a stronger association with the self-reported depression set. Reproduced in part from Resnik et al. (2015a))

ble 6. Although the most polarized topics in Tables 5 and 6 are understandably similar because they are the strongest, the informed prior still assisted in identifying new useful topics and improving the ones identified by uninformed sLDA.

The "control topics" in Table 6 appeared to be influenced more by the prior than those on the other end of the scale. Perhaps this is because it is easy to identify coherent themes to cluster the group of depressed users but the control users could be discussing any variety of things. In this case the prior-improved model appears to do better than uninformed sLDA by identifying topics dealing with *activities* and *relationships*, which seem to be relevant to good mental health. The *sleep* topic at response value -1.595 is fairly similar to what it was before, though it is interesting to see it again associated with the control group when our clinician advisor in Resnik et al. (2015b) identified it as bearing a strong association with depression and our model on the Pennebaker data put it on the side of neuroticism (albeit more towards the neutral end). A possible explanation for this behavior is given in analysis of a similar *sleep* topic in the Twitter portion of Section 5.3 below.

For topics with strong positive regression parameters there were fewer changes introduced by the informed prior, but we again see some improvement. For instance, note the appearance in Table 6 of a topic at regression value 4.204 dealing with *relationships* and *mental health*, which is very strong and replaces the less-impressive topic at regression value 4.318 in Table 5. Uninformed sLDA found no such topic, even though each user of the depression set had at least one tweet in which they indicated they were diagnosed with depression.[5]

5.3 SNLDA

We ran SNLDA with parameters: $\alpha = 1$, topic-word Dirichlet hyperparameter for each level $(\beta_0, \beta_1, \beta_2) = 0.25, 0.1, 0.05$, $\rho = 1$, Gaussian variance for topic's regression parameters for each level $(\sigma_0, \sigma_1, \sigma_2) = 0.01, 0.5, 2.5$, $\mu = 0.0$, scaling factor for the β priors $(\gamma_1, \gamma_2) = 100, 10$, means of the β priors at each level (probability that a token stays at current node) $(\pi_1, \pi_2) = 0.2, 0.2$, and again using a continuous response value for both sets. The more complex structure of SNLDA makes it more difficult to determine k-values to use for each level of the hierarchical tree, so we experimented with a variety of different values, the results of which are better compared in Section 6. However, for ease of comparison with the other models, the below sections use values: number of topics in layer 1 (k_1) $= 50$, number of topics in layer 2 (k_2) $= 2$. Thus any improvements from adding the additional layer are more apparent and we were able to use the same informed prior as the other models.

Pennebaker essays Again the Pennebaker essays proved to be a good candidate for building intuitive topic posteriors with SNLDA, as highlighted in Table 7. Of special interest among the control-related topics is topic 17 and its children that seem to deal

[5]These tweets were pivotal in defining the dataset (see Section 3)

Node Index	η	Top 20 words
35	-7.469	ride bike line rid half add horse lie speak happen worth save tempt amount completely earlier afraid train climb enjoy
35:0	-26.308	time realize feel day leave sit begin guess happen reason spend ready hit hold ago pas entire grow close imagine
35:1	0.98	time remember forget remind stuff memory list run reason begin lose start fast detail energy forgive stop direction notice stick
17	-7.264	plan stuff organize ahead expect worry normal couple list manage challenge tough surprisingly extra john fast fill listen regular due
17:0	-21.764	time hard start free day sit harder idea relax alright reason hand hear figure mind lucky constantly eventually real huge
17:1	-6.8	lot time fun enjoy easier mate learn forward busy figure bite easy tough trouble bad amount bother deal surprise habit
13	-5.234	care selfish honest trust fault secret doe nice mine concern decent insecure rude enemy lack mistake kill bug fit return
13:1	-22.4	people person understand act true idea judge reason talk huge smart complain opinion assume accept easily barely close million impress
13:0	5.112	guy lot guess talk stuff happen bad chris weird bring mind suppose sort haven start draw kevin story confuse sit
...	...	...
19	-0.392	fish die story beautiful robert death princess medium center diana tank fire tom accident stand gun poor series issue woman
19:1	-1.104	people world country kill american america war live government news culture happen power rule september bush president hear support law
19:0	1.517	mother die child father family life death doctor kid pain bear grandmother cancer age nurse happen baby hospital ago pass
...	...	...
41	1.473	monday friday tuesday thursday schedule wednesday week tomorrow due bad test catch set sunday late lunch watch interview quick noon
41:0	-4.036	class professor psychology teacher lecture easy student note experiment semester pennebaker psy grade teach material skip discussion sign rhetoric section
41:1	5.569	sleep tire night hour nap day bed sleepy study stay busy lay morning relax late exhaust time rest catch extremely
...	...	...
0	1.99	matt black asian white chinese mexican race call japanese stereotype surprise american culture hispanic body listen ago china power extremely
0:1	-8.858	friend miss houston home parent brother austin worry family hard haven leave time yesterday move visit close day bye expect
0:0	3.854	cry leave bad sad happen call depress upset lose day break tear care emotional hear angry time hard normal hand
...	...	...
34	3.863	water drink thirsty teeth swim bottle mouth taste coke brush pool soda super sip bird caffeine pepper cloudy eternity late
34:0	-2.622	cold hot rain weather air nice freeze walk warm foot sweat heat winter wear temperature window bad dry condition shower
34:1	23.689	start time sit leave stop wait lot doe bad homework fall guess walk hate run eat front feel lazy annoy
18	6.61	focus concentrate skill task topic distraction procrastinate improve distract period extremely rarely add difficult procrastinator frustrate quality easily fairly perfectionist
18:1	8.381	person mind idea situation deal matter control reason tend negative attention stay figure effect decide mention issue honestly concern single
18:0	23.153	time change realize spend waste continue sit worry pass mood happen busy affect effort period wrong start easily hear consume
43	10.485	plan absolutely travel south york begin surprise pick theatre frustrate hop disappoint visit chicago completely extremely explore italy europe italian
43:1	-2.462	austin school texas move town university live city experience attend dallas friend college decide extremely hometown graduate difficult huge entire
43:0	42.519	people love feel hard lot doe life excite reason crazy person amaze understand bad friend talk wait act true scar

Table 7: Most extreme and selected neutral (ranked by 1st-level nodes) SNLDA topics from Pennebaker training data. More positive regression parameters indicate a stronger association with a high Big-5 neuroticism score

with *organization*, *planning*, and *free time*, all of can be reasonably assumed to reduce the anxiety and frustration association with neuroticism in a person's life. We can see the added benefit of an additional layer of nodes in Topic 41, which seem to deal with *scheduling* and when coupled with *schoolwork* (41:0) are positive, but when coupled with *sleepiness* (41:1) could indeed be an indicator of neuroticism.[6] Similarly, though *missing home* in topic 0:1 could be an indicator of good relationships and good emotional health, it also appears to be a sign of separation anxiety and depression in 0:0. Topic 19 seems to deal with *accidents* and *death*, which can be a sign of strong public spirit in 19:1, or also cause a great deal of emotional trauma if involving a close family member, as appears in 19:0. Unlike sLDA, which mostly highlighted negativity on the extreme positive side, the SNLDA model appears to instead highlight as most associated with neuroticism topics involving *somatic complaints*, *frustrations with the essay task itself* and with *interpersonal relationships*.

Twitter In Table 8 we can see the advantage of the hierarchical topic structure from the 50 topics found by sLDA and LDA on the Twitter set. A strong example of this is Topic 10, which appears to involve *adolescence*, *music*, and *relationships*. Where sLDA identified similar topics as being neutral, SNLDA shows us that the topic can actually be polarized on both sides of the spectrum. Perhaps, for instance, music can be positive in a new and happy relationship (10:1), but negative after a breakup (10:0). In topic 48 we similarly see that anxiety and mental stress can be highly indicative of depression (48:1), but is much more normal when associated with regular college stress (48:0). Also of interest is Topic 44, which breaks out medical words that would usually be associated with poor mental health into a control topic of what appear to be individuals who merely work in a medical profession, and a depression valanced topic for those actually being treated by such.

Though the hierarchy seems to be helpful in the qualitative interpretation of the results, without the informed prior the most extreme positive topics identified by SNLDA don't appear to be those that are most associated with depression. Instead the 1st-level topics that were highest on the depression scale seem to have a lot to do with *TV*, *hair*, and *pets/family*. These topics that deal more with discussing things and events instead of a person's internal state of being.

[6] Notice the inclusion of Pennebaker himself in 41:0

Node Index	η	Top 20 words
30	-3.994	dont ill ive youre didnt wont whats doesnt shes havent isnt wasnt idk wouldnt cuz theyre arent reply couldnt youve
30:0	-1.868	hate people care talk ugly friend school call stop wanna shut boyfriend mad girl annoy anymore stupid mood sick funny
30:1	2.104	fuck shit bitch fuckin wtf dick damn piss libra suck sex hell bullshit kill stupid cunt slut asshole fucker pussy
27	-3.379	lmao lol lmfao drake nigga damn wtf dude kendrick shut bro idk ratchet twerk talk black kanye lmaooo lmfaoooo lakers
27:1	-8.179	girl love guy damn people country shit stop kid boy white doe bitch crazy real wait tweet true fight hit
27:0	1.742	EMOJI EMOJI EMOJI EMOJI EMOJI EMOJI EMOJI EMOJI EMOJI wanna EMOJI EMOJI EMOJI EMOJI maria phone EMOJI EMOJI EMOJI EMOJI
20	-2.093	dog eat puppy house phone pet ball cat bee suck till kid nephew fire turtle neighbor smell lazy bark meow
20:1	-12.574	bad time mine baby call walk day run hand mouth watch eye forever stick sit teeth finger shut feel dead
20:0	1.637	sleep bed wake tire night hour feel asleep sick hate awake fall home nap cold eat lay stay cuddle hungry
...	...	...
44	-0.569	nurse hospital care doctor patient chat twitter read health social tweet topic enjoy medical medium lawyer dont app issue error
44:0	-5.601	question call answer hear start move happen stuff nice leave change bad late head guess phone wrong remember lot job
44:1	6.745	gonna wanna time gotta talk life doe kinda anymore guess wait hell hurt care leave stop sick run mind crazy
...	...	...
10	0.929	miley taylor song cyrus swift people wanna sing doe hannah shut music wreck die vmas vma lana zac drake hate
10:1	-3.665	girl text relationship person boyfriend kiss perfect girlfriend hate fall guy feel smile stop cuddle single jealous cute call night
10:0	25.513	girl life boy friend break heart cry lose date forever people hate time beautiful remember change kill walk miss god
...	...	...
48	2.354	mental comment link depression health submit anxiety feel illness disorder advice friend issue deal suicide life doe experience suffer diagnose
48:0	-2.642	class college study day test school exam student homework math semester campus teacher final finish walk fail start week sit
48:1	3.932	stop cut start reason care alive strong life hope day hold literally suicide inspiration leave world sick bad people die
...	...	...
38	4.137	castle pic season watch episode mind reid scene criminal fan night yep agree matthew hug start xoxo week sweet hun
38:0	-1.026	sound hope lovely time bite hear lot excite idea mine piece doe music tea day listen wonderful busy forward enjoy
38:1	0.83	rain train australia watch nice morning weather sun cold sydney storm play awesome dear manchester fly enjoy wow twin funny
07	5.245	hair eye cute cody black color brown dye cut laugh meet date blonde wear ignore pretty red age pink dark
07:0	-4.744	lol lmao yea crazy hell smh lolol idk soooo funny gotta ppl voice hungry sooo ready aww sound min awww
07:1	26.825	love baby miss watch wait girl leave sleep night doe sister call brother feel mom month boyfriend till dad ugh
47	7.278	cat ago week tho kitten buy month recently surgery pain doe book date plan family couple move child read wife
47:0	-1.136	christmas nail snow color thanksgiving blue holiday beauty polish spring tree flower beautiful fall winter merry santa hand eve perfect
47:1	4.583	movie watch film episode doctor star horror batman series character awesome comic war trailer dead marvel hug black america review

Table 8: Most extreme and selected neutral (ranked by 1st-level nodes) SNLDA topics from Twitter training data without informed prior. More positive regression parameters indicate a stronger association with the self-reported depression set

Node Index	η	Top 20 words
32	-5.222	text call phone talk message texting friend send answer stop lose ring feel texted EMOJI mad listen mine hope sound
32:0	-20.506	hate love feel girl bad annoy funny kid god sad anymore happy awkward hungry stop pretty weird sing mad hilarious
32:1	-1.285	girl mom love boyfriend hate phone talk friend mad picture text stop play dumb EMOJI cuddle retweet funny girlfriend EMOJI
01	-4.69	lol yea stupid funny guess wow ugh awwww suck min fun yeah crazy cool hmmm minute mess yay haha freak
01:0	-3.214	haha hahaha yeah hahahaha hahah watch awww hahahah omg phone ahhh suck alright hehe ohh freak nah weird wow aww
01:1	4.128	lol bruh idk boo watch nah chill cool hell funny dude ima crazy movie guess bad start hella bore lil
44	-4.308	hurt care relationship person love trust feeling feel mistake happy strong fight treat advice fall doe forgive reason insecure worst
44:0	-14.46	love smile cry perfect heart beautiful kiss deserve happy true omg hug tear fall hold feel sweet forever sad forget
44:1	11.458	love baby life heart forever fall sagittarius promise forget fight friend trust pain beautiful understand stand single relationship smile hide
...	...	...
12	-0.641	post facebook pretty comment doe bad tag yesterday move cool scan lady stuff nice low deal guy idea feel write
12:1	-0.314	cool guy super dude stuff pretty nice play doe yep apparently sweet terrible pick bet joke cheer surprise seahawks huh
12:0	9.814	ugh cute picture shit tattoo guy suck makeup omg gonna stupid pierce nude freak annoy hair sexy pretty fuck life
...	...	...
45	1.165	night sleep late bed hour tomorrow morning wake bus start nap class alexis relax fall earlier rest yesterday asleep catch
45:0	-2.404	morning tomorrow excite wait friday night weekend saturday monday till sunday EMOJI weather yesterday shop gym start wednesday date finish
45:1	1.561	sleep wake bed feel hour asleep nap fall awake mood finally lay wait sleepy haven leave shower friend till gud
...	...	...
24	3.329	friend hang close meet person realize friendship talk school awe lot ignore shy stay mine short touch college hardest fun
24:0	-3.926	school wear hate EMOJI summer ugly senior talk tho stay stupid homecoming week stfu music sick lolol lunch fml swear
24:1	16.224	people talk word hear lot fun learn hug stupid lesson speak grow meet touch hang dream reason deal past music
...	...	...
23	4.711	gay fake alex doe josh confirm hate boob stupid suck vagina funny bra drink call amanda whore lesbian bad weird
23:0	-2.407	time photoset capricorn cont call kris matt ben wow kevin chris mike joe ryan eric tony trade don hat katie
23:1	-0.416	smoke drink hoe hit weed drug time call black bad mind bring till throw alcohol party suck crazy cigarette shower
22	4.978	feel stress scar nervous understand mood exhaust afraid anxiety anymore confuse emotion calm normal frustrate dread empty relax guilty bother
22:1	0.662	android aww hun start complete piss bed EMOJI load yay collect pack blah shop coin reply decide bath fit worse
22:0	2.039	omg bieber tweet guy gonna remember meet beliebers damn cody ariana bye fuck selena proud amaze ready account pic dad
13	9.977	people person fake care don meet ignore stranger act rude lot feel automatically judge tend personality assume genuinely easily accept
13:1	0.126	wanna gonna time cut literally drive honestly car feel care kill EMOJI hate person realize hold EMOJI wrong perfect waste
13:0	48.712	feel time hate gonna cry happen bad people sleep nice care literally cute person send hard hell perfect true date

Table 9: Most extreme and selected neutral SNLDA topics from Twitter training data without informed prior. More positive regression parameters indicate a stronger association with the self-reported depression set

Twitter with informed prior Table 9 shows the results of running SNLDA with the 50-topic informed prior resulting from running LDA on the Pennebaker dataset, which show improvement over the uninformed version of SNLDA. As with informed LDA and sLDA, we see more topics from the prior appearing in the resulting topics that highlight discussion of *sleep*, *emotionality*, and *relationships*. The strongest evidence of this is in the topics with the most positive regression value, which should be most indicative of depression. Although these topics were not scored high by uninformed SNLDA (Table 8), they were definitely among the topics most intuitively associable with depression once the prior was introduced–in particular topics 22 and 13 with their subtopics show strong *emotional stress*, *anxiety*, and *negativity*. On the opposite end of the spectrum are topics about *relationships* and *joviality*, as expected for the control group. Overall, these topics appear to be the strongest, qualitatively speaking, of any of the models for the Twitter dataset.

A few other interesting groups are highlighted in the center of the chart, which are insightful because they show how a single parent topic can yield subtopics on both ends of the regression scale. For instance, topic 12 deals with Facebook comments which are fairly neutral in themselves, yet the negative ones in 12:0 show a definite correlation with depression, while 12:1 is more positive and on the control side of the scale. The placing of the *sleep* topic among the control topics in Tables 6 and 5 is better explained here in topic 45. *Sleepiness* and *lethargy* (45:1) are identified as indicators of depression while sleep in association with *anticipation* (45:0-presumably referring to the next day) can actually be a good thing. Similarly, topic 24 and its children (also topic 44 with its subtopics on the control side) show how interpersonal relationships can be either good or bad for one's mental health.

6 Quantitative Results

To determine the ability of these models to meet our second goal of automatically identifying mental health disorders in individuals, we used each of them to predict whether an unseen set of user documents belonged to someone from the Depressed or Control group of the Twitter dataset. This classification was done first with a standard baseline system, which we improved on using sLDA and SNLDA. LDA is an inherently unsupervised model and is therefore not suitable for prediction by itself.[7]

Model	AUC
TF-IDF SVM Baseline	0.805
sLDA	0.824
sLDA with prior	0.818
SNLDA (k=15,8)	0.794
SNLDA with prior (k=15,8)	0.806
SNLDA (k=30,4)	0.790
SNLDA with prior (k=30,4)	0.820
SNLDA (k=50,2)	0.802
SNLDA with prior (k=50,2)	0.802

Table 10: Area under curve (AUC) of selected feature configurations for depression vs. control prediction on the Twitter Dataset. Baseline reproduced from Resnik et al. (2015a)

Since the dataset used in these experiments is identical to that of Resnik et al. (2015a), we were able to directly reference that baseline, which uses a standard TF-IDF model, for comparison of our results.[8] These results are included in Table 10.

As with the qualitative results in Section 5, the sLDA and SNLDA models were created using 500 iterations of sampling, the first 100 of which were burn-in. The testing was then done by iterating the new data over the model 250 times, averaging the resulting regression values from every 25 iterations except for the first 50 which were used as burn-in. It was shown in Nguyen et al. (2014) that averaging across iterations produces better results than merely accepting the final state and our experiments confirmed that this testing method performed about as well or slightly better than merely the final state in all instances.

As discussed in Section 3, the input documents for each model were grouped by the week in which they were tweeted. Since the original classification of the data is based on the author it was therefore necessary to re-aggregate the resulting features across weeks.

[7] Although it is common practice to simply use the topic posteriors as features, we explored this with little success in Resnik et al. (2015b) so it is not further addressed here

[8] Baseline features were created by Thang Nguyen and Leonardo Claudino from the work in Resnik et al. (2015a), and trained and predicted using a linear SVM classifier

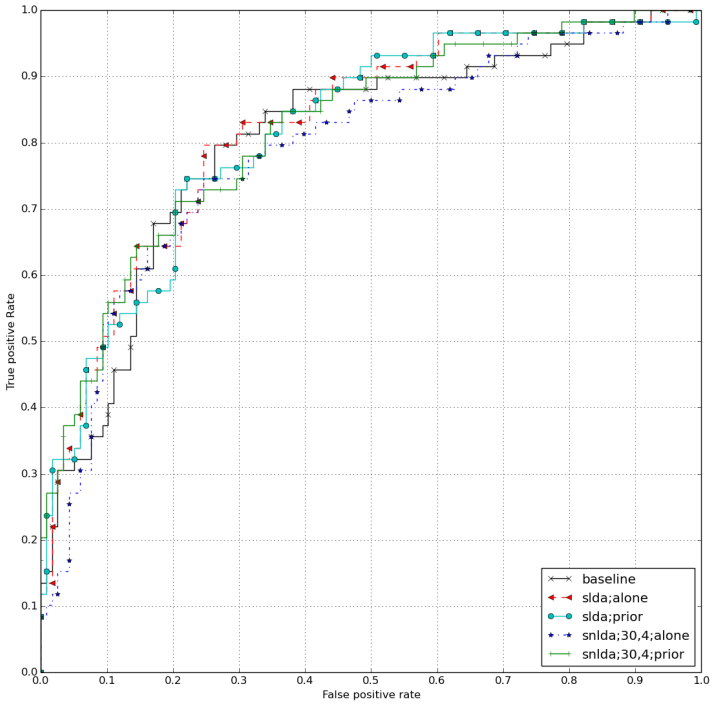

Figure 5: ROC curves of best performing systems.

We did so by averaging the predicted regression values across weeks, weighting each by the number of tweets in that week so that the more an author posted in a week the more important it would be in the final classification. All models were trained on the training set and tested against the unseen test set (see Section 3)

Since we used regression values instead of the binary labels from the data set, we can assume the predicted response values correlate with the confidence the model has in a prediction. A binary prediction was then made by selecting a threshold and counting the results on either side of it. Thus we were able to fine-tune our results to get a good idea of the the trade-offs between misses and false-positives. These Receiver Operator Characteristic (ROC) curves are shown in Figure 5.

6.1 SLDA

As discussed in Section 4, we used our sLDA models to directly predict the response values of unseen documents. Resnik et al. (2015b) compared the results of feature-based approaches and model-based approaches for predicting a slightly different version of the dataset with sLDA. In those experiments using the topic posteriors performed about as well as "vanilla" LDA since this technique does little to leverage the added benefits of sLDA. Since every instance of predicting with the model directly performed about as well as or better than the feature-based classification, we will focus on the former.

As can be seen in Table 10 and Figure 5, both sLDA models performed well on the test set and surpassed the baseline, yet uninformed sLDA unexpectedly performed slightly better than the informed version. The fact that the prior did not improve the results show that the uninformed model in this case successfully identified crucial topics to make a decision even if those topics were not necessarily as desirable by a human. This suggests that the computer-generated topics from an uninformed prior could be useful to provide insight into diagnosing mental health in ways that are not currently understood. However, the results of both were close together and other experiments have shown improvement from the use of an informed prior (see SNLDA results below, as well as Resnik et al. (2015b) and Resnik et al. (2015a)) so we believe it is still a useful approach

to the task.

6.2 SNLDA

For predicting regression values with the SNLDA model it again made sense to use the model directly in order to make full use of the hierarchical structure instead of flattening the topics into a feature-based system. However, it was not able to very well outperform sLDA with the additional structure, except partially in the best case (with the informed prior and a topic structure of 30 layer-1 nodes with 4 layer-2 nodes each outperforming informed sLDA). One possible reason for this is the increased difficulty of determining a good structure to use in the model, which is harder to experimentally determine in this case since there is a degree of freedom at each level, including the determination of number of levels, instead of just one variable in sLDA. As such, we did more experimentation with these models in an attempt to find a good structure, the best of which results are shown in Table 10. Note that for the prior of the $k = 30, 4$ model, a new prior was created by running LDA on the Pennebaker dataset with $k = 30$, and for the prior of the $k = 15, 8$ model 15 topics were manually extracted from the 50 topic LDA prior, based on the topics highlighted by the clinician in Resnik et al. (2015b).

Here we see more definitive improvement from the introduction of the informed prior, most notably in the $k = 30, 4$ model. Since the same 30-topic prior was detrimental in the case of sLDA ($AUC = 0.800$), somehow the structure of that SNLDA tree was better suited for being influenced by the prior.

Overall SNLDA models are well-suited for prediction as they all performed nearly as well or better than the baseline model and further experimentation with the tree structure could improve the results even further. Since running with an informed prior improved or matched the uninformed results in all SNLDA instances, we conclude that an informed prior is useful for classification with this model.

7 Conclusions and Future Directions

Topic modeling is a useful tool for extracting information related to mental health from a large dataset of naturally occurring text. The ability of these models to predict depression is on-par with the state-of-the art models (as shown in Coppersmith et al. (2015)), especially when using a version of topic-modeling beyond "vanilla" LDA or when combining these techniques with additional features in a supervised learning paradigm. As a future effort, we would like to incorporate these successes into larger feature-sets that have been shown to be successful in related work (see Section 2), and take advantage of data available in social media that were not leveraged in our topic modeling (i.e. timestamps, friend-lists and geotags).

The reduction of impossibly large social-media datasets to ones that are more readable makes topic modeling attractive for improving the human-interpretability of the data. This holds true for all of the techniques explored, with each adding its own value. A clinician's input regarding the quality of our results in running LDA on the Pennebaker essays provided some feedback into our model and eventually let to improved results through the use of informative priors. Since this is the principle behind Interactive Topic Modeling (Hu et al. (2014)), we intend to explore this technique more explicitly in future work.

Though the predictive accuracy of these techniques is not at the level of replacing psychologist's jobs in diagnosing depression, topic modeling can easily bring large amounts of new data to bear on treating and diagnosing depression in a clinical setting. We hope to facilitate the introduction of such tools into psychology domains to supplement and extend the work of clinicians in treating depression and other mental health disorders.

Acknowledgments

Special thanks and recognition is due to Philip Resnik for his extensive advice and direction throughout this research as well as his comprehensive reviewing of the content of this paper. Without his advisor-ship none of this work would have happened and he is responsible for many of the ideas expressed and implemented here. Additionally I recognize and appreciate Viet-An Nguyen who provided the source code for the models as well as frequent technical advice for using it, plus a substantial contribution to the technical description of SNLDA in Section 4. I also extend gratitude to Leonardo

Claudino and Thang Nguyen for their assistance and collaboration in this and our related work as well as Jordan Boyd-Graber for his frequent consultation. Credit is due to Glen Coppersmith, Mark Drezde, Jamie Pennebaker and their colleagues for providing the data used in these experiments. I also appreciate Rebecca Resnik, Brett Holden, and Camille Armstrong for providing consultation in the psychological analysis of our generated topics. Financial support was provided in part by NSF awards 1320538, 1018625, and 1211153, though the opinions expressed here do not necessarily reflect the views of the sponsor.

References

David M Blei and Jon D McAuliffe. 2007. Supervised topic models. In *Proceedings of Advances in Neural Information Processing Systems*, pages 121–128.

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

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Denver, Colorado, USA, June. North American Chapter of the Association for Computational Linguistics.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *ICWSM*.

Gregor Heinrich. 2005. Parameter estimation for text analysis. Technical report, Technical report.

Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. Interactive topic modeling. *Machine Learning*, 95(3):423–469.

Oliver P John, Laura P Naumann, and Christopher J Soto. 2008. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, 3:114–158.

Gerald Matthews, Ian J Deary, and Martha C Whiteman. 2003. *Personality traits*. Cambridge University Press.

Viet-An Nguyen, Jordan L. Boyd-Graber, and Philip Resnik. 2014. Sometimes average is best: The importance of averaging for prediction using MCMC inference in topic modeling. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1752–1757. ACL.

Viet-An Nguyen. 2015. *AGuided Probabilistic Topic Models for Agenda-Setting and Framing*. Ph.D. dissertation, University of Maryland.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Pew Research Center. 2014. Demographics of key social networking platforms. `http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms`. Accessed: 2015-04-26.

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

Philip Resnik, William Armstrong, Leonardo Claudino, and Thang Nguyen. 2015a. The university of maryland clpsych 2015 shared task system. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015b. Beyond LDA: exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*.

H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Hanna M Wallach, David Minmo, and Andrew McCallum. 2009. Rethinking lda: Why priors matter.