Beyond LDA: Exploring Supervised Topic Modeling for Depression-Related Language in Twitter

Philip Resnik[2,4], **William Armstrong**[1,4], **Leonardo Claudino**[1,4], **Thang Nguyen**[3], **Viet-An Nguyen**[1,4], **and Jordan Boyd-Graber**[3,5]
[1]Computer Science, [2]Linguistics, [3]iSchool, and [4]UMIACS, University of Maryland
[5]Computer Science, University of Colorado Boulder
{resnik,armstrow}@umd.edu
{claudino,daithang,vietan}@cs.umd.edu
{daithang,jbg}@umiacs.umd.edu

Abstract

Topic models can yield insight into how depressed and non-depressed individuals use language differently. In this paper, we explore the use of supervised topic models in the analysis of linguistic signal for detecting depression, providing promising results using several models.

1 Introduction

Depression is one of the most prevalent forms of mental illness: in the U.S. alone, 25 million adults per year suffer a major depressive episode (NAMI, 2013), and Katzman et al. (2014) observe that "[by] 2020, depression is projected to be among the most important contributors to the global burden of disease". Unfortunately, there are significant barriers to obtaining help for depression and mental disorders in general, including potential stigma associated with actively seeking treatment (Rodrigues et al., 2014) and lack of access to qualified diagnosticians (Sibelius, 2013; APA, 2013). When patients suffering from depression see a primary care physician, the rates of misdiagnosis are staggering (Vermani et al., 2011).

These considerations have helped to motivate a recent surge of interest in finding accessible, cost effective, non-intrusive methods to detect depression and other mental disorders. Continuing a line of thought pioneered by Pennebaker and colleagues (Pennebaker and King, 1999; Rude et al., 2004, and others), researchers have been developing methods for identifying relevant signal in people's language use, which could potentially provide inexpensive early detection of individuals who might require a specialist's evaluation, on the basis of their naturally occurring linguistic behavior, e.g. (Neuman et al., 2012; De Choudhury et al., 2013; Coppersmith et al., 2014). Critical mass for a community of interest on these topics has been building within the computational linguistics research community (Resnik et al., 2014).

To date, however, the language analysis methods used in this domain have tended to be fairly simple, typically including words or n-grams, manually defined word categories (e.g., Pennebaker's LIWC lexicon, Pennebaker and King (1999)), and "vanilla" topic models (Blei et al., 2003, latent Dirichlet allocation (LDA)). This stands in contrast to other domains of computational social science in which more sophisticated models have been developed for some time, including opinion analysis (Titov and McDonald, 2008), analysis of the scientific literature (Blei and Lafferty, 2007), and computational political science (Grimmer, 2010).

In this paper, we take steps toward employing more sophisticated models in the analysis of linguistic signal for detecting depression, providing promising results using supervised LDA (Blei and McAuliffe, 2007) and supervised anchor topic models (Nguyen et al., 2015), and beginning some initial exploration of a new supervised nested LDA model (SNLDA).

2 Data

Our primary experimental dataset is the Twitter collection created by Coppersmith et al. (2014)

Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 99–107,
Denver, Colorado, June 5, 2015.

and used in the CLPsych Hackathon (Coppersmith, 2015). The raw set contains roughly 3 million tweets from about 2,000 twitter users, of which roughly 600 self-identify as having been clinically diagnosed with depression (by virtue of having publicly tweeted "I was diagnosed with depression today" or similar, with manual validation by the individuals preparing the data). We grouped all tweets by an individual user into a single document, and a base vocabulary was created by pre-processing documents using standard NLP tools, specifically: (1) keeping alphanumeric words and word-encoded emoticons, (2) removing stopwords using the MALLET stopword list, and (3) lemmatizing using NLTK's WordNetLemmatizer. We then filtered out words that appeared in fewer than 20 documents, words only appearing in documents of fewer than 50 words (fewer than 10 users), and URLs. The resulting set of 1,809 documents was randomly divided into train/dev/test subsets to create a 60-20-20% split. We model documents from the Twitter datasets depression subset as having a regression value of 1 and those from the control subset as having a regression value of -1.

In building some of our models, we also use a collection of 6,459 stream-of-consciousness essays collected between 1997 and 2008 by Pennebaker and King (1999), who asked students to think about their thoughts, sensations, and feelings in the moment and "write your thoughts as they come to you". As discussed in Section 3.1, running LDA on this dataset provides informative priors for sLDA's learning process on the Twitter training data. The student essays average approximately 780 words each, and Resnik et al. (2013) showed that unsupervised topic models based on this dataset can produce very clean, interpretable topical categories, a number of which were viewed by a clinician as relevant in the assessment of depression, including, for example, "vegetative" symptoms (particularly related to sleep and energy level), somatic symptoms (physical discomfort, e.g. headache, itching, digestive problems), and situational factors such as homesickness.

For uniformity, we preprocessed the stream-of-consciousness dataset with the same tools as the Twitter set.[1] We created a shared vocabulary for our models by taking the union of the vocabularies from the two datasets, leading to a roughly 6% increase in vocabulary size over the Twitter dataset alone.

3 Models

3.1 LDA

LDA (Blei et al., 2003) uncovers underlying structure in collections of documents by treating each document as if it was generated as a "mixture" of different topics. As a useful illustration, replicating Resnik et al. (2013), we find that using LDA with 50 topics on the Pennebaker stream-of-consciousness essays produces many topics that are coherent and meaningful. We had a licensed clinical psychologist review these to identify the topics most likely to be relevant in assessing depression, shown in Table 1.[2] This step exploiting domain expertise can be viewed as a poor-man's version of interactive topic modeling (Hu et al., 2014), which we intend to explore in future work.

3.2 Supervised LDA

Basic (sometimes referred to as "vanilla") LDA is just the entry point when it comes to characterizing latent topical structure in collections of documents, and extensions to LDA have proven valuable in other areas of computational social science. *Supervised* topic models (sLDA, introduced by Blei and McAuliffe (2007)), extend LDA in settings where the documents are accompanied by labels or values of interest, e.g. opinion analysis (reviews accompanied by k-star ratings) or political analysis (political speeches accompanied by the author's political party). The advantage of supervised topic modeling is that the language in the documents and the accompanying values are modeled jointly — this means that the unsupervised topic discovery process seeks to optimize not just the coherence of the topics underlying the discourse, but the model's ability to predict the associated values. So, for example, in modeling Amazon reviews, vanilla LDA might discover a topic containing opinion words (*great*, *enjoy*, *dislike*, etc.) but sLDA would be more likely to separate these out into a positive opinion-word topic

[1] With the exception of the document count filters, due to the different number and sizse of documents; instead, we allowed all non stopwords that appear in more than one document.

[2] Many other topics were coherent and meaningful, but were judged as falling below the clinician's intuitive threshold of relevance for assessing depression.

Notes	Valence	Top 20 words
high emotional valence	e	life live dream change future grow family goal mind rest decision marry chance choice successful career set regret support true
high emotional valence	e	love life happy heart amaze hurt perfect crazy beautiful lose smile cry boy true fall real sad relationship reason completely
relationship problems	n	time boyfriend friend relationship talk person break doe happen understand hard trust care spend reason san situation antonio date leave
transition to college	n	school college student semester university experience hard grade parent graduate freshman campus learn texas attend teacher expect challenge adjust education
self-doubt	n	question realize understand completely idea sense level bring issue concern simply situation lack honestly admit mention fear step feeling act
poor ego control	n	yeah suck wow haha stupid funny hmm crap crazy blah freak type ugh weird lol min gosh hey bore hmmm
feeling ignored/annoyed *	n	call talk phone doe stop bad ring message loud head homework answer cell mad forget annoy sound hurt suppose mine
somatic complaints	n	cold hot feel sick smell rain walk start weather bad window foot freeze nice wait throat day heat hate warm
emotional distress *	n	feel happy day sad depress feeling cry scar afraid lonely head moment emotion realize confuse hurt inside guilty fear upset
family of origin issues	n	mom dad family sister parent brother kid child mother father grow doctor baby hard cousin die age cry proud husband
negative affect *	n	damn hell doe shit fuck smoke woman hate drink piss sex drug kid god bitch time real break screw cigarette
anxiety over failure	n	worry hard study test class lot grade focus mind start nervous stress concentrate trouble reason easier hop harder fail constantly
negative affect*	n	hate doe bad stupid care understand time suck happen anymore mad don mess scar horrible smart matter hat upset fair
sleep disturbance*	n	sleep tire night morning wake bed day time late stay hour asleep nap fall start tomorrow sleepy haven awake lay
somatic complaints	n	hurt eye hear itch hand air sound tire nose arm loud leg leave noise finger smell neck stop light water
social engagement	p	game football team win ticket excite school weekend week texas run lose night season saturday sport dallas longhorn coach fan
exercise, good self-care	p	run day feel walk class wear lose weight buy gym gain short fat dress shop exercise campus clothe body shirt

Table 1: LDA topics from Pennebaker stream-of-consciousness essays identified by a clinician as most relevant for assessing depression. Topics with negative valence (n) were judged likely to be indicators for depression, those with positive valence (p) were judged likely to indicate absence of depression, and those labeled (e) have strong emotional valence without clearly indicating likely assessment. Asterisked topics were viewed as the strongest indicators.

(*great*, *enjoy*, etc.) predicting higher star ratings and a negative opinion-word topic (*dislike*, *sucks*, etc.) predicting lower ratings.

Table 2 illustrates topics we obtained by running 50-topic sLDA on the Pennebaker stream-of-consciousness dataset, using, as each essay's regression variable, the student's degree of neuroticism — a personality trait that can be a risk factor for internalizing disorders such as depression and anxiety — as assessed using the Big-5 personality inventory (John and Srivastava, 1999). The neuroticism scores are Z-score normalized, so the more positive (negative) a topic's regression value, the more (less) the supervised model associates the topic with neuroticism. As was done for Table 1, we had a clinician identify the most relevant topics; these were presented in random order without the neuroticism regression values in order to avoid biasing the judgments. The sLDA neuroticism values for topics in Table 2 pattern nicely with the clinician judgments: negative neuroticism scores are associated with clinician-judged positive valence topics, and positive neuroticism scores with negative valence. Scores for the p and n valence items differ significantly according to a Mann-Whitney U test ($p < .005$).

Table 3 shows topics derived using sLDA on the Twitter training data; owing to space limitations, we show the topics with the 5 highest and 5 lowest Z-normalized regression scores.

We also derive topics on Twitter training data using a "seeded" version of sLDA in which the 50 topics in Section 3.1 provide informative priors; recall that these came from the Pennebaker stream-of-consciousness data. We were motivated by the hypothesis that many of the topics emerging cleanly in Pennebaker's population of college students would be relevant for the Twitter dataset, which also skews toward a younger population but is significantly messier. Although the sLDA runs with and without informative priors produce many similar topics, Table 4 shows a number of topics identified by sLDA with informative priors, that were not among the topics found without them.

3.3 Supervised Anchor Model

As another extension to LDA-based modeling, we explore the use of the the anchor algorithm (Arora et al., 2013, hence ANCHOR), which provides a fast way to learn topic models and also enhances interpretability by identifying a single "anchor" word associated with each topic. Unlike sLDA, which examines every document in a dataset, ANCHOR requires only a V by V matrix Q of word cooccurrences, where V is the size of the vocabulary, to discover topics. Nguyen et al. (2015) introduces a *supervised* anchor algorithm (hence SANCHOR), which, like sLDA, takes advantage of joint modeling with document-level metadata to learn better topics and enable prediction of regression variables.

Briefly, the anchor algorithm assumes that each

Notes	Valence	Regression value	Top 20 words
social engagement	p	-1.593	game play football team watch win sport ticket texas season practice run basketball lose soccer player beat start tennis ball
social engagement	p	-1.122	music song listen play band sing hear sound guitar change remind cool rock concert voice radio favorite awesome lyric ipod
social engagement	p	-0.89	party night girl time fun sorority meet school house tonight lot rush drink excite fraternity pledge class frat hard decide
social engagement	p	-0.694	god die church happen day death lose doe bring care pray live plan close christian control free hold lord amaze
high emotional valence	e	-0.507	hope doe time bad wait glad nice happy worry guess lot fun forget bet easy finally suck fine cat busy
somatic complaints	n	-0.205	cold hot hair itch air light foot nose walk sit hear eye rain nice sound smell freeze weather sore leg
poor ego control; immature	n	0.177	yeah wow minute haha type funny suck hmm guess blah bore gosh ugh stupid bad lol hey stop hmmm stuff
relationship issues	n	0.234	call talk miss phone hope mom mad love stop tonight glad dad weird stupid matt email anymore bad john hate
homesick; emotional distress	n	0.34	home miss friend school family leave weekend mom college feel parent austin stay visit lot close hard boyfriend homesick excite
social engagement	p	0.51	friend people meet lot hang roommate join college nice fun club organization stay social totally enjoy fit dorm conversation time
negative affect*	n	0.663	suck damn stupid hate hell drink shit fuck doe crap smoke piss bad kid drug freak screw crazy break bitch
high emotional valence	e	0.683	life change live person future dream realize mind situation learn goal grow time past enjoy happen control chance decision fear
sleep disturbance*	n	0.719	sleep night tire wake morning bed day hour late class asleep fall stay nap tomorrow leave mate study sleepy awake
high emotional valence	e	0.726	love life happy person heart cry sad day feel world hard scar perfect feeling smile care strong wonderful beautiful true
memories	n	0.782	weird talk doe dog crazy time sad stuff funny haven happen bad remember day hate lot scar guess mad night
somatic complaints*	n	0.805	hurt type head stop eye hand start tire feel time finger arm neck move chair stomach bother run shoulder pain
anxiety*	n	1.111	feel worry stress study time hard lot relax nervous test focus school anxious concentrate pressure harder extremely constantly difficult overwhelm
emotional discomfort	n	1.591	feel time reason depress moment bad change comfortable wrong lonely feeling idea lose guilty emotion confuse realize top comfort happen
homesick; emotional distress*	n	2.307	hate doe sick feel bad hurt wrong care happen mess horrible stupid mad leave worse anymore hard deal cry suppose

Table 2: sLDA topics from Pennebaker stream-of-consciousness essays identified by a clinician as most relevant for assessing depression. Supervision (regression) is based on Z-scored Big-5 neuroticism scores.

Regression value	Top 20 words
2.923	eat fat cut hate fuck weight cross line body sleep scar die food cry fast ugh gym skinny boyfriend week
1.956	omg cry school god cute literally hair gonna hate mom ugh idk wow sleep omfg laugh wear picture tbh sad
1.703	book write read episode twitter story tweet fan cover movie awesome win doctor alex season character yeah film happen week
1.676	fuck shit bitch gonna wanna hate damn man dick wtf suck dude smoke god drink gay sex girl hell piss
1.602	pls depression donate kindly film support mental word ocd health package producer hour anxiety mind tomorrow hun teamfollowback disorder visit
-1.067	game win team play coach season run player state tonight fan football baseball lead brown dodger ohio score red week
-1.078	game man win play team damn fan lebron tonight dude gonna football heat ball bro nba hell boy basketball bull
-1.354	man goal fan win unite game arsenal play team player league score season madrid football match manchester cup sign chelsea
-1.584	EMOJI EMOJI EMOJI EMOJI EMOJI EMOJI EMOJI EMOJI EMOJI EMOJI EMOJI EMOJI EMOJI birthday EMOJI bae EMOJI lmao EMOJI
-2.197	birthday class tonight week literally hour tomorrow weekend summer college home break party favorite excite game die beach drive study

Table 3: Most extreme sLDA topics from Twitter training data

Regression value	Top 20 words
4.119	happiness cut line depression post cross anxiety mental read view eat suicide scar die ago family connect month account hospital
1.68	brain episode doctor fan season week movie link tumblr comment finally read story ago scene buy gaga write order hey
0.054	eat sleep morning hour home food bed drink week run dinner tomorrow wake dog fat coffee tire buy tonight lunch
0.039	girl baby boy hot beautiful kiss date heart sexy dance babe week sweet hair marry birthday lady retweet nice miley
-0.641	tonight dress beautiful fashion style cute party beauty hair nail black shop lady free beach vip bottle outfit buy ticket
-1.199	wanna baby sleep phone hate home mad bore tire bitch text morning hurt play man ready tomorrow leo stay ima

Table 4: Selected sLDA topics from Twitter training data with informative priors

Anchor	Top 20 words
business	business market plan lead build birmingham car city support social pay company system legal financial deal service design creative control
college	college school class girl week student study hour test learn summer parent high hate sit tomorrow senior mom wear teacher
dance	dance girl school amaze tonight wear song funny movie picture beautiful pretty fun sing omg hot high drink hair boy
fat	fat eat hate body sleep weight girl bed skinny cry fast beautiful die perfect cross hair ugh week sick care
friday	friday tonight weekend week tomorrow party monday saturday morning thursday tuesday sunday club meet drink hour wednesday queen card movie
fuck	fuck shit hate bitch girl wanna gonna sleep care school drink damn die suck yeah break kill text stupid phone
god	god heart man jesus lord bless pray person men mind church trust woman care truth girl walk hear matter true
haha	haha yeah tomorrow gonna bed pretty omg xx nice sleep excite tweet fun week hour yay mum amaze hate tonight
music	music song album awesome single grey rock hear justin meet band gonna light sound tour grab concert artist tonight amaze
play	play game tonight man fan team radio hey season sound hour yeah episode nice buy hear football ball beat player
win	win game team fan tonight vote season player goal football man chance final card coach score week luck usa top

Table 5: Examples of topics identified by SANCHOR on Twitter training data.

topic has at least one anchor word that unambiguously identifies that topic — when you see an anchor in a document, you know for sure that that topic is relevant somewhere in it.[3] For instance, *fifa* might be an anchor word for the soccer topic. Words such as *ball*, *net*, or *player* are related to the soccer topic, but they cannot be anchor words because they are also mentioned in topics such as baseball or networking. The supervised anchor algorithm (SANCHOR) extends ANCHOR by expanding the word co-occurrence data to include word-level conditional probabilities for the regression variable of interest (Nguyen et al., 2015). Table 5 illustrates a number of the topics discovered by SANCHOR in the Twitter training data.[4] See the Appendix for more details.

3.4 Supervised Nested Latent Dirichlet Allocation

Like all topic models, SNLDA is based on a generative model in which each document is created by selecting a probability distribution over topics it will contain, and then selecting words based on that topic distribution; that is, every document can be viewed as coming from a mixture of topics. Like sLDA (Section 3.2), SNLDA allows us to connect each topic with a regression variable of interest; however, in SNLDA we additionally assume that the underlying topics are organized into a tree. The additional hierarchy is intended to improve our ability to represent more complicated text and account for the fact that a single topic can contribute to either side of the regression parameter depending on its subcontext.

The input of SNLDA is identical to that of sLDA, namely a collection of D documents, each associated with a response variable. The output is a tree $\mathcal{T}$, with fixed height L and a pre-defined number of children K_l for each level l of the tree. At each node, we have a process similar to sLDA: we draw (a) a topic ϕ_k specifying what this node k is about and (b) a regression parameter η_k specifying the weight of k in capturing the response variable. A child node is connected with its parent node, topically, by drawing its topic distribution from a Dirichlet prior Dir$(\beta_{l_k}\phi_{p_k})$ whose mean vector ϕ_{p_k} is the topic of the parent node p_k. See the Apendix for more details.

[3]This assumption can be violated, but the truer it is, the better the model.

[4]Note that SANCHOR does not produce regression values for each topic in the way that sLDA does.

Features	P, R=0.5	P, R=0.75	P, R=1
(A) Unigrams	0.607	0.483	0.342
(B) LIWC	0.571	0.479	0.344
(C) LDA-50 (Mallet)	0.447	0.402	0.349
(D) sLDA features, uninformative priors	0.308	0.352	0.341
(E) sLDA **features, informative priors**	**0.648**	**0.584**	**0.353**
(F) **SANCHOR**	**0.638**	**0.529**	**0.348**
(G) sLDA prediction, uninformative priors	0.568	0.479	0.271
(H) sLDA prediction, informative priors	0.643	0.436	0.303
(I) Combining A+B+C+E+F	0.632	0.526	0.342

Table 7: Evaluation on Twitter test set, showing precision at three levels of recall.

The structure of this model is similar in spirit to SHLDA (Nguyen et al., 2013), and it is intended to serve a similar purpose, namely inducing structure in such a way that sub-topics meaningfully specialize their parent nodes. Nguyen et al. illustrate how this can be useful in the political domain — for example, in an analysis of Congressional floor debates, the model identifies taxation as a first-level topic, with one child node that captures Democrats' framing of the subject (with terms like *child support*, *education*, *students*, and *health care*, i.e. the social services that taxes pay for) and another child node capturing Republican framing (with terms like *death tax*, *jobs*, *family businesses*, and *equipment*, related to the implications of taxation for businesses). Here our goal is to use a similarly structured model, but jointly modeling authors' language with their depression status as the regression variable rather than their political affiliation.

Tables 6 provide some illustrative examples of SNLDA topics induced from the Twitter training data. The hierarchical organization is apparent in, for example, Topic 8, where a sports topic is subdivided into subtopics related to, among others, soccer and professional wrestling; Topic 9 on politics/news, subdividing into, among others, education, India, Britain, and controversies involving race and law enforcement (Ferguson, the Travyon Martin shooting); and Topic 6, which our clinician characterizes as issues that tend to be discussed on social media by women, e.g. relationships, body issues, parenting, and physical maladies.

Topic:Subtopic	Regression value	Top 20 words
8	-3.279	game win team play player fan season football coach basketball score lebron nfl baseball nba ball beat lead ohio brown
8:3	-0.15	goal dodger cup madrid match brazil usa chris soccer germany worldcup ronaldo messi spain ucla ger fifa orlando oscar att
8:5	-0.021	spur wrestle match wwe raw danny podcast wrestler fantastic batman title fan cont cena nxt wrestlemania corbin debut manu kick
9	-1.874	obama vote news report support government police bob president tax plan obamacare labour campaign business law leader election birmingham city
9:1	-0.244	student art education teach college teacher visa africa university scholarship mandela literacy typhoon science digital haiyan nelson child phot
9:2	-0.23	india medium hindu saint allegation conspiracy indian follower delhi fake diwali expose police sai rape truth false support jail fir
9:3	-0.056	manchester tory bbc ukip lib britain cut british event dems council library thatcher clegg guardian dem england farage unite mail
9:7	0	ferguson black williams prison crochet police topic false morning zimmerman trayvon chicago woman angeles family community ebay guest sxsw discus
6	0.093	lol sleep haha hate wanna omg ugh eat mom tire gonna baby idk bed yeah tomorrow wake hurt bore hair
6:0	0.102	anxiety vlog stress weightloss anxious panda migraine tire guinea therapy shift interview EMOJI remedy mind relief irritable chil
6:1	0.171	skype husband lols hubby dream reply week meet edit youi nowplaying owner instagram steam beautiful yup birthday notice amaze admin
6:4	0.972	fat eat cut weight cross calorie skinny fast line body burn workout account food water weigh gain exercise leg healthy

Table 6: Selected SNLDA topics

4 Quantitative results

An established use of topic models in predictive modeling is to create a K-topic model using some relevant document collection (which might or might not include the training set), and then, for training and test documents, to use the posterior topic distribution $\Pr(z_k|d), k = 1..K$ as a set of K features (Resnik et al., 2013; Schwartz et al., 2014). These features can be useful because the automatically discovered topics sometimes capture higher-level properties or "themes" in authors' language that have predictive value beyond individual words or phrases. Our experimentation used these features from LDA, sLDA, and sANCHOR; using topic posteriors from SNLDA is left for future work.

To assess the ability of the models/features and how they compare to baseline methods, we trained a linear support vector regression (SVR) model on the union of the Twitter train and dev sets, evaluated on the test set. We chose regression over classification despite having binary labels in our data in order to more easily evaluate precision at various levels of recall, which can be done simply by thresholding the predicted value at different points in order to obtain different recall levels. In addition, SVR has been shown to be an adequate choice to other similar text regression problems (Kogan et al., 2009), and in future analyses the use of the linear kernel will allow us to further see the contributions of each feature from the weights assigned by the regression model. We follow standard practice in using unigram features and LIWC categories as baseline feature sets, and we also use topic posteriors from a 50-topic LDA model built on the Twitter training data.[5]

As shown in Table 7, we evaluated alternative models/feature sets by fixing the percentage of recalled (correctly classified) depression subjects at levels R=1, 0.75, and 0.5 and looking at precision, or, equivalently, the rate of misdiagnosed control subjects.[6] When R=1, it means the classification threshold was set to the smallest value such that all depressed subjects were correctly classified. The results show that all methods perform similarly badly at 100% recall: when required to identify all depressed individuals, two thirds or so of the flagged individuals are false positives. When allowed to trade off recall for improved precision, sLDA performs well *if* provided with informative priors, and the supervised anchor method (without informative priors) is not far behind.

For completeness, we also used the sLDA models directly for prediction, i.e. computing the expected response value for a test document from $\eta^\top \bar{z}$ where $\bar{z}$ is the document's posterior topic distribution and the ηs are the per-topic regression parameters. These results are shown as "sLDA prediction" (lines G and H) in the table. The utility of this technique is illustrated on the model without informative priors (G), where it yielded a substantial improvement over the use of the posterior topics as features for both LDA (line C) and sLDA with uninformative priors (line D). This suggests that sLDA-based features (D) may have performed so poorly because they failed to sufficiently leverage the added value of the regression parameter, making them no better than vanilla LDA (C). SNLDA models can similarly be used to predict a test document's expected

[5]Not to be confused with the LDA model built using the stream-of-consciousness dataset in Section 3.1, which was used to provide informative priors for sLDA.

[6]Owing to an error discovered late in the writing process, 4 out of 396 test items were excluded from the SANCHOR evaluation. If accepted, this will be corrected in the final version.

response value; we will explore this in future work.

To the extent that this test set is representative of the real world, the results here seem promising: with R=0.75, 3 of 4 depressed individuals are detected at the cost of roughly 1 false positive per 3 individuals predicted. The representativeness of the experiment, however, depends heavily on the true prevalence of depression. On the one hand, the prevalence in the Coppersmith (2015) dataset — in the vicinity of 30% — is consistent with Vermani et al. (2011), who cite four prior studies when stating that "major depressive disorder has been shown to be one of the most common mental disorders seen in primary care patients, with prevalence rates ranging from 23% to 35%". In their own study of 840 primary care patients in Canada, they found that 27.2% met criteria for major depressive disorder. On the other hand, those numbers seem quite high: Vermani et al. also cite a WHO study finding that 10.4% of screened patients met criteria for current depression, and that number is more in line with NIMH's 12-month prevalence figures.[7]

Although it introduces a mismatch between training and test data prevalence, therefore, we experimented with randomly down-sampling the number of positive examples in the test data (but not the training set) to get a test-set prevalence of 10%. Table 8 shows the mean $\pm$ standard deviation results.[8] The absolute numbers are significantly lower, but the same trend persists in the comparison across models/features.

Elsewhere in this volume, a companion paper describes our participation in the CLPsych 2015 Shared Task (Coppersmith et al., 2015), providing experimentation on shared task datasets and further discussion and analysis (Resnik et al., 2015).

5 Conclusions

Our goal in this paper has been to go beyond simple, "vanilla" topic models to explore the potential utility of more sophisticated topic modeling in the automatic identification of depression. Qualitative examples have confirmed that LDA, and now additional LDA-like models, can uncover meaningful and potentially useful latent structure, and our quantitative experimentation using the CLPsych Hackathon dataset has shown more sophisticated topic models exploiting supervision, such as SLDA and SANCHOR, can improve on LDA alone.

Features	P, R=0.5	P, R=0.75	P, R=1
Uni	0.239 ± 0.047	0.165 ± 0.042	0.108 ± 0.010
SANCHOR	0.271 ± 0.045	0.189 ± 0.033	0.126 ± 0.015
sLDA-inf	0.267 ± 0.042	0.216 ± 0.035	0.119 ± 0.022

Table 8: Mean $\pm$ stdev precision (P) and recall (R) scores of linear SVR for the3 best-performing models/features in Table 7 (SLDA with informative priors, SANCHOR and unigrams) on test sets where the prevalence of depression was randomly downsampled to 10%.

One of the additional take-aways here is that informative priors can make a meaningful difference in performance; we plan to pursue this further using interactive topic modeling (Hu et al., 2014) with our domain expert, and also by providing informative priors for anchor methods.

Another important observation is that prevalence matters, and therefore further work is needed exploring the sensivity of early screening approaches to changes in the proportion of the target signal represented in the data.

Finally, a third interesting observation coming out of our experimentation was that aggregation might matter a great deal. Rather than aggregating by author, we defined a *set* of documents for each author as their tweets aggregated on a weekly basis, i.e. one document per author per week. Although just a preliminary experiment with one model, we found with SANCHOR that the weekly grouping improved precision at R=0.5 to 74% and precision at R=0.75 to 62%. The improvement makes intuitive sense, since topics and emotional state vary over time and language samples grouped on a weekly basis are likely to have more internal coherence than samples aggregated over long periods. This led us to adopt weekly aggregation in the CLPsych 2015 shared task, with good results (Resnik et al., 2015), and other forms of aggregation therefore seem like a fruitful area for further exploration.

[7] http://www.nimh.nih.gov/health/statistics/prevalence/major-depression-among-adults.shtml

[8] To obtain means and standard deviations we down-sampled 100 times.

Acknowledgments

We appreciate the helpful comments of the anonymous reviewers, we are grateful to Rebecca Resnik for contributing her comments and clinical expertise, and we thank Glen Coppersmith, Mark Dredze, Jamie Pennebaker, and their colleagues for kindly sharing data and resources. This work was supported in part by NSF awards 1320538, 1018625, and 1211153. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

Appendix

The Appendix is available online at `http://www.umiacs.umd.edu/~resnik/pubs/clpsych2_appendix.pdf`.

References

APA. 2013. The critical need for psychologists in rural America. http://www.apa.org/about/gr/education/rural-need.aspx, Downloaded September 16, 2013.

Sanjeev Arora, Rong Ge, Yoni Halpern, David M. Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2013. A practical algorithm for topic modeling with provable guarantees.

David M Blei and John D Lafferty. 2007. A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35.

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

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *The 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.

Glen Coppersmith. 2015. [Un]Shared task: Computational linguistics and clinical psychology. `http://glencoppersmith.com/papers/CLPsych2015_hackathon_shared_task.pdf`.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the 2013 ACM annual conference on Human factors in computing systems*, pages 3267–3276. ACM.

Justin Grimmer. 2010. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, 18(1):1–35.

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

Oliver P John and Sanjay Srivastava. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2:102–138.

Martin A Katzman, Leena Anand, Melissa Furtado, and Pratap Chokka. 2014. Food for thought: understanding the value, variety and usage of management algorithms for major depressive disorder. *Psychiatry research*, 220:S3–S14.

Shimon Kogan, Dimitry Levin, R. Bryan Routledge, S. Jacob Sagi, and A. Noah Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280. Association for Computational Linguistics.

NAMI. 2013. Major depression fact sheet, April. http://www.nami.org/Template.cfm?Section=depression.

Yair Neuman, Yohai Cohen, Dan Assaf, and Gabbi Kedma. 2012. Proactive screening for depression through metaphorical and automatic text analysis. *Artif. Intell. Med.*, 56(1):19–25, September.

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. 2013. Lexical and hierarchical topic regression. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1106–1114.

Thang Nguyen, Jordan Boyd-Graber, Jeffrey Lund, Kevin Seppi, and Eric Ringger. 2015. Is your anchor going up or down? Fast and accurate supervised topic models. In *North American Chapter of the Association for Computational Linguistics*.

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.

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, Rebecca Resnik, and Margaret Mitchell, editors. 2014. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Baltimore, Maryland, USA, June.

Philip Resnik, William Armstrong, Leonardo Claudino, and Thang Nguyen. 2015. The University of Maryland CLPsych 2015 shared task system. 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.

Stephanie Rodrigues, Barbara Bokhour, Nora Mueller, Natalie Dell, Princess E Osei-Bonsu, Shibei Zhao, Mark Glickman, Susan V Eisen, and A Rani Elwy. 2014. Impact of stigma on veteran treatment seeking for depression. *American Journal of Psychiatric Rehabilitation*, 17(2):128–146.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

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.

Kathleen Sibelius. 2013. Increasing access to mental health services, April. http://www.whitehouse.gov/blog/2013/04/10/increasing-access-mental-health-services.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, pages 111–120. ACM.

Monica Vermani, Madalyn Marcus, and Martin A Katzman. 2011. Rates of detection of mood and anxiety disorders in primary care: a descriptive, cross-sectional study. *The primary care companion to CNS disorders*, 13(2).