The University of Maryland CLPsych 2015 Shared Task System

Philip Resnik$^{2,4}$, William Armstrong$^{1,4}$, Leonardo Claudino$^{1,4}$, Thang Nguyen$^3$

$^1$Computer Science, $^2$Linguistics, $^3$iSchool, and $^4$UMIACS, University of Maryland

{resnik,armstrow}@umd.edu
{claudino,daithang}@cs.umd.edu

1 Introduction

The 2015 ACL Workshop on Computational Linguistics and Clinical Psychology included a shared task focusing on classification of a sample of Twitter users according to three mental health categories: users who have self-reported a diagnosis of depression, users who have self-reported a diagnosis of post-traumatic stress disorder (PTSD), and control users who have done neither (Coppersmith et al., 2015; Coppersmith et al., 2014). Like other shared tasks, the goal here was to assess the state of the art with regard to a challenging problem, to advance that state of the art, and to bring together and hopefully expand the community of researchers interested in solving it.

The core problem under consideration here is the identification of individuals who suffer from mental health disorders on the basis of their online language use. As Resnik et al. (2014) noted in their introduction to the first ACL Workshop on Computational Linguistics and Clinical Psychology, few social problems are more costly than problems of mental health, in every possible sense of cost, and identifying people who need help is a huge challenge for a variety of reasons, including the fear of social or professional stigma, an inability of people to recognize symptoms and report them accurately, and the lack of access to clinicians. Language technology has the potential to make a real difference by offering low-cost, unintrusive methods for early screening, i.e. to identify people who should be more thoroughly evaluated by a professional, and for ongoing monitoring, i.e. to help providers serve their patients better over the long-term continuum of care (Young et al., 2014).

This paper summarizes the University of Maryland contribution to the CLPsych 2015 shared task. More details of our approach appear in Resnik et al. (2015), where we also report results on preliminary experimentation using the CLPsych Hackathon data (Coppersmith, 2015).

2 System Overview

In our system, we build on a fairly generic supervised classification approach, using SVM with a linear or RBF kernel and making use of baseline lexical features with TF-IDF weighting.

2.1 Variations explored

The innovations we explore center on using topic models to develop features that capture latent structure in the dataset, going beyond “vanilla” latent Dirichlet allocation (Blei et al., 2003) to include supervised LDA (Blei and McAuliffe, 2008, sLDA) as well as a supervised variant of the “anchor” algorithm (Arora et al., 2013; Nguyen et al., 2015, sAnchor). Putting together various combinations in our experimentation — linear vs. RBF kernel, big vs. small vocabulary, and four feature configurations (namely sLDA, sAnchor, lexical TF-IDF, and all combined), we evaluated a total of 16 systems for each of the three shared tasks (discriminating depression vs. controls, depression vs. PTSD, and PTSD vs. controls) for a total of 48 systems in all.

In general below, systems are named simply by concatenating the relevant elements of the description. For example, combobigvocabSVMlinear is the name of the system that uses (a) an SVM with linear kernel (SVMlinear), (b) models computed using the big vocabulary (bigvocabulary, details below), and (c) the “all combined” feature configuration.
2.2 SLDA and SAnchor topic features

We briefly describe the features we used based on sLDA and sAnchor; see Resnik et al. (2015) for more details, as well as sample topics induced by these models on the closely related CLPsych Hackathon dataset. For both topic models, we used posterior topic distributions, i.e. the vector of \( \Pr(\text{topic}_k|\text{document}) \), \( k = 1..K \) in a \( K \)-topic model, as features for supervised learning.

SLDA posteriors with informed priors. To take full advantage of the shared task’s labeled training data in a topic modeling setting, we opted to use supervised topic models (sLDA, introduced by Blei and McAuliffe (2008), sLDA). These systems are simply referred to as SLDA Prediction.

Tables 3 to 5 show the top words in the sLDA topics with the 5 highest and 5 lowest \( Z \)-normalized regression scores, sLDA having been run with parameters: number of topics \( (k) = 50 \), document-topic Dirichlet hyper-parameter \( (\alpha) = 1 \), topic-word Dirichlet hyper-parameter \( (\beta) = 0.01 \), Gaussian variance for document responses \( (\rho) = 1 \), Gaussian variance for topic’s regression parameters \( (\sigma) = 1 \), Gaussian mean for topic’s regression parameters \( (\mu) = 0.0 \), using binary variables for the binary distinction in each experimental task.

Supervised anchor (SAnchor) posteriors. The anchor algorithm (Arora et al., 2013) provides a fast way to learn topic models and also enhances interpretability by automatically identifying a single “anchor” word associated with each topic. For example, one of the topics induced from the Hackathon tweets is associated with the anchor word fat and is characterized by the following most-probable words in the topic:

\[
\text{fat eat hate body sleep weight girl bed skinny cry fast beautiful die perfect cross hair ugh week sick care}
\]

Nguyen et al. (2015) introduce sANCHOR, a supervised version of the anchor algorithm which, like sLDA, jointly models text content along with a per-document regression variable. We did not explore informed priors with SANCHOR in these experiments; this is left as a question for future work.

2.3 Classifier details

The majority of our experiments used SVM classifiers with either a linear or an RBF kernel. Specifically, we used the python scikit-learn module (sklearn.svm.SVC), which interfaces with the widely-used libsvm. Default parameters were used throughout except for the class_weight parameter, which was set to None.

In the SLDA Prediction experiments, we followed Blei and McAuliffe (2008) in computing the response value for each test document from \( \eta^\top \bar{z} \) where \( \bar{z} \) is the document’s posterior topic distribution and the \( \eta_s \) are the per-topic regression parameters. SLDAPrediction_1 and SLDAPrediction_2 were conducted with small and big vocabularies, respectively.

2.4 Data Preparation

Data organization: weekly aggregation. To overcome potential problems for topic modeling with documents that are too small (individual tweets) or too large (all tweets for an author) we grouped tweets together by the week they were posted. Thus each author corresponded to several documents, one for each week they tweeted one or
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. Many more of the 50 topics from this model were intuitively coherent but not judged as particularly relevant for the depression-assessment task. This table is reproduced from Resnik et al. (2015).

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

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. Many more of the 50 topics from this model were intuitively coherent but not judged as particularly relevant for the depression-assessment task. This table is reproduced from Resnik et al. (2015).

more times; each document was treated as being labeled by the author’s individual-level label. In preliminary experimentation, we found that this temporal grouping greatly improved the performance of our models, though it should be noted that organizing the data in this way fails to account for the fact that an author’s mental health can vary greatly from week to week. For instance, a user identified as having depression at some point may not be experiencing symptoms in any given week, yet that week’s document would still be labeled as positive for depression. This could potentially be mitigated in future work by attempting to identify the time of diagnosis and increasing the label weight on documents near that time.

Token pre-processing and vocabularies. All systems went through the same basic pre-processing: we first removed words with non-alphanumeric characters, emoticon character codes, and stop words.1 The remaining tokens were further lemmatized.

For SVM classification we explored using systems with both small and big vocabularies. For the former, we first filtered out documents with less than 50 tokens and then selected tokens that appeared more than 100 documents; the latter was obtained in a similar fashion, except setting the word-per-document cutoff to 10 rather than 50, and the document-per-word cutoff to 30 instead of 100.2

For sLDA prediction, we used an external vocabulary produced from the set 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 2, running LDA on this dataset provided informative priors for sLDA’s learning process on the Twitter training data. The student essays averaged approximately 780 words each, and for uniformity, we pre-processed them with the same tools as the Twitter set.3 We created a shared vocabulary for our models by taking the union of the vocabularies from the two datasets, resulting in a roughly 10-20% increase in vocabulary size over the Twitter dataset alone.

Author-level features. In order to arrive at a single feature vector for each author based on documents aggregated at the weekly level, we weight-averaged features across weeks, where weights corresponded to the fraction of the author’s tweets associated with each week alone. In other words, the more an author posted in a week, the more important that week’s features would be, compared to the other weeks.

---

1Unicode emoticons were left in, converted to the token EMOJI.

2When referring to vocabulary size, we use the terms short and small interchangeably.

3With the exception of the document count filters, due to the different number and sizes of documents, which were adjusted accordingly.
other weeks.

**Data splits.** We did an 80-20 partition into train-
ing and development sets, respectively. Since we
did not do any hyper-parameter tuning, the dev set
was used primarily for sanity checking and to get a
preliminary sense of system performance. We report
test set results based on models that were trained on
the training set alone.4

3 Results

3.1 Overall results and ROCs

The ROC curves for all our submitted systems
on the shared tasks (Section 2) are shown in Figure 1. The area under curve (AUC) scores for TFIDF (baseline) and all configurations of combined
features (best systems) are shown in Table 2, from
which we see that the 8 best-performing feature con-
fugurations achieved an average AUC of about 0.84.
We obtained the best overall results when we used
a big vocabulary, combined all features, and trained
a linear SVM. We saw that bigger vocabularies
improved performance of linear SVMs but not RBF
SVMs, and that, in general, linear SVMs did better.

The order of difficulty for these discrimination
problems seems to be DvP > DvC > PvC, judg-
ing from the performance of our top 8 systems.
This suggests that there may be greater overlap of
linguistic signal between tweets from people who
have self-reported PTSD and those who have self-
reported depression, which is not entirely surprising
since the two conditions often co-occur. According
to Tull (2015), “Depression is one of the most com-
monly occurring disorders in PTSD... Among peo-
ple who have or have had a diagnosis of PTSD, ap-
proximately 48% also had current or past depression
People who have had PTSD at some point in their
life are almost 7 times as likely as people without
PTSD to also have depression.”

3.2 Qualitative discussion for sLDA

To get a sense of the role that supervised topic
modeling may be playing, we take a brief qualita-
tive look at the topics induced by sLDA on the train-
ing set. Tables 3, 4, and 5 show the most polarized
topics resulting from the sLDA models constructed
on the DvC, DvP and PvC tasks respectively, where
polarization is measured by the value of the sLDA
regression variable for the topic. The topics we see
are not as clean and coherent as the topics in Ta-
ble 1, which is unsurprising since the latter topics
came from LDA run on individually coherent doc-
uments (stream-of-consciousness essays) collected
from a more uniform population (UT Austin col-
lege students) (Pennebaker and King, 1999), as con-
trasted with aggregations of tweets over time from a
sample of Twitter users.

At the same time, there does seem to be inter-
pretable signal distinguishing the high versus low
polarity topics, at least in comparisons against con-
trols. Comparing depression vs. control (Table 3),
we see subdivisions of negative affect — for ex-
ample, the most depression-oriented topic, as identified
using positive regression values, is dominated by
negatively oriented interjections (f***, sh*t, damn,
etc.), and the next most depression oriented topic
appears to largely capture relationship discussion
(omg, cute, cry, guy, feel, hot, pretty). Conversely,
the least depression-oriented topics in the table, i.e.
with the most negative regression values, contain
not only many positive affect terms (lol, hahaha, etc.)
but also activities related to family (car, weekend,
home) and social activity (food, tonight, party, din-
ner, weekend).

In PTSD vs. control (Table 5), we see, among the
topics more oriented toward PTSD users, topics that
may be related to attention to veteran issues (sign,
support, homeless, petition, marine), and possibly

<table>
<thead>
<tr>
<th>Feature configuration / Problem</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DvC</strong></td>
<td></td>
</tr>
<tr>
<td>tfidfshortvocabSVMlinear</td>
<td>0.824</td>
</tr>
<tr>
<td>tfidfbigvocabSVMlinear</td>
<td>0.845</td>
</tr>
<tr>
<td>tfidfshortvocabSVMrbf</td>
<td>0.831</td>
</tr>
<tr>
<td>tfidfbigvocabSVMrbf</td>
<td>0.815</td>
</tr>
<tr>
<td><strong>DvP</strong></td>
<td></td>
</tr>
<tr>
<td>comboshortvocabSVMlinear</td>
<td>0.841</td>
</tr>
<tr>
<td>combobigvocabSVMlinear</td>
<td>0.860</td>
</tr>
<tr>
<td>comboshortvocabSVMrbf</td>
<td>0.835</td>
</tr>
<tr>
<td>combobigvocabSVMrbf</td>
<td>0.830</td>
</tr>
<tr>
<td><strong>PvC</strong></td>
<td></td>
</tr>
<tr>
<td>tfidfbigvocabSVMlinear</td>
<td>0.860</td>
</tr>
<tr>
<td>tfidfshortvocabSVMrbf</td>
<td>0.815</td>
</tr>
</tbody>
</table>

Table 2: Area under curve (AUC) of selected feature con-
fugurations in Fig. 1 per each problem: depression vs.
control (DvC), depression vs. PTSD (DvP) and PTSD
vs. control (PvC). Boldface: big vocabulary, combined
features, SVM linear. This setting was the best for all
three tasks.

---

4It is possible that modest improvements could be obtained
by folding the dev set back into the training data, but we wished
to avoid inspecting the dev set so that we can continue to use it
for further development.
Figure 1: ROC curves of submitted systems.

Table 3: Most extreme sLDA topics from Twitter training data (Depression (1) vs. Control (-1))
Consistent with the lower performance on depression vs. PTSD (DvP), in Table 4 no topics jump out quite as forcefully as being polarized toward one condition or the other, except for the most PTSD-oriented topic, which appears as if it may concern efforts to draw attention to PTSD (ptsd, learn, fear, speak, positive, step, battle, join, voice, awareness). It may be, however, that in incorporating the depression vs. PTSD distinction, the model is actually capturing broader characteristics of relevant subpopulations: particularly in this dataset, people self-reporting a PTSD diagnosis may well be older on average than people self-reporting a depression diagnosis, if not chronologically than in terms of life experience. The topic with the most positive regression value in the table, i.e. leaning toward depression rather than PTSD, includes terms most likely related to youth/pop culture: Niall Horan, Harry Styles, Liam Payne, and Louis Tomlinson are the members of the pop boy band One Direction. Other positive- (i.e. depression-leaning) topics in the table also have a quality of disinhibition more characteristic of younger people. In contrast, the negative- (i.e. PTSD-leaning) topics in the table tend toward more mature topics, including, for example, politics and current affairs (obama, tcot (top conservatives on Twitter), vote, ebola).

Although our efforts are still in an early stage, our hope is that more sophisticated topic models can not only enhance predictive accuracy, as in Table 2, but also that observations like these about topics or themes might help create insight for clinicians. Examples like the ones in Tables 1 and 3-5 can help establish face validity with clinicians by showing that these models can capture things they already know about. Others can potentially lead to new questions worth asking, e.g. in Table 3, might the topic relating to entertainment (watch, movie, episode, read, write, season, book) suggest a closer look at social isolation (staying in watching movies, reading books) as a linguistically detectable online behavior that might correlate with increased likelihood of depression? If true, this would be consistent with, and complement, Choudhury et al. (2013), who look at non-linguistic measures of social engagement in Twitter among their potential depression-related attributes.\(^5\)

### 4 Conclusions and Future Directions

In this paper we have briefly described the University of Maryland contribution to the CLPsych 2015 shared tasks. We found that TF-IDF features alone...
performed very well, perhaps surprisingly well, on all three tasks; TF-IDF combined with supervised topic model posteriors resulted in an even more predictive feature configuration.

In future work, we plan to conduct a thorough error analysis to see where the models go astray. We also plan to look at the extent to which our data organization may have influenced performance; in preliminary experimentation in Resnik et al. (2015), we found suggestive evidence that aggregating tweets by week, rather than as a single document per user, might make a significant difference, and that is the strategy we adopted here. This may not just be a question of document size — other time-based aggregations may be worth exploring, e.g. grouping tweets by time of day.

Acknowledgments

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.

References


