Confronting Depression with Machine Learning

David Hall

May 8, 2012

Clinical depression is one of the most common mental health diseases in developed countries, often called the “common cold” of mental health diseases. According the CDC, in 2011 nearly 1 in 10 American adults currently suffers from some form of clinical depression, with rates in some states as high as 15%. Nearly half of those suffer from major depression. However, more than 30% of those with major depression do not seek treatment [AHRQ], while nearly 80% of those with some form of depression do not seek treatment at all. [Murray and Fortinberry, 2011]

However, even though it is so common, depression exacts a very real human cost. Suicide rates among those with major depression is as high as 15%. Depression also imposes a drag on the American economy. In 2004, losses in productivity due to depression were estimated at $51 billion in the United States alone [Rand].

Given the human and economic cost of depression and the fact that a shockingly high percentage of people do not seek treatment, it makes sense to develop tools that can automatically assist in the detection, prevention, and cure of depression. In particular, with the rise of “social” blogging platforms like LiveJournal, Twitter, and Facebook, we now have hundreds of millions of people disclosing their behaviors on an unprecedented scale. Thus, we have the resources (and indeed, moral imperative) to seek to prevent or ease the very real suffering of tens of millions of people.

In this report, we describe preliminary results in our effort to employ data mining to automatically detect depression, as well as to understand what factors lead to depression and its cessation. For data, we use a large scrape of the popular blogging platform LiveJournal, which includes self-reported emotional state in a large number of cases. by examining what factors contribute to both the onset and cessation of depression, we can hopefully detect and prevent depression before it becomes a problem.

1 Cognitive Behavioral Therapy

Our approach is inspired by Cognitive Behavioral Therapy (CBT). Cognitive Behavioral Therapy has deep roots in several traditions of therapy. CBT has increasingly become the preferred form of non-drug therapy for many cases of depression. CBT can be usefully applied to many other conditions, though our focus is on depression.
The key insight behind CBT is that there are certain negative beliefs about oneself that tend to be self-reinforcing. For instance, someone with low self-esteem might believe they have no friends, leading to them avoiding social situations, which in turns fuels the negative belief and sense of isolation.

Viewed this way, CBT is about identifying negative feedback cycles and counteracting them. This comprises two phases. The first, functional analysis or guided discovery, gets the patient to identify thoughts and beliefs that lead to maladaptive behaviors. The second, behavior change, helps the patient to cope with stressful situations in a manner that is not maladaptive.

From a data mining perspective, we are concerned with the first phase. Can we identify negative thoughts before they become feedback cycles? Moreover, can we identify positive behaviors or thoughts that are indicative of recovery?

2 The Dataset

We obtained a crawl of around 29 million posts from the popular blogging platform LiveJournal.\(^1\) This includes 448000 users, and around 2 billion tokens of content words.

In addition, there is a significant amount of metadata. First, posts are tagged with their user id, a title, the self-reported user’s location, and a current music choice. Locations were typically not true locations, but instead personally meaningful locations like “desk,” or “bedroom.” We include these metadata (except for user id) as features.

Perhaps most interestingly, unlike other platforms, users are encouraged to self-report their “mood” via one of 132 emoticons.\(^2\) Many of these “moods” differ subtly: there is a mood for “grateful” as well as “thankful.” Thus, We binned the moods into four coarse emotional states, along with an additional “Other” category: Happy, Sad, Angry, Anxious and Other.

We are interested in depressed episodes. We considered two ways of describing a depressed mood. The simplest to define depression as 7 or more consecutive posts with “Sad” affect. (We call 7 or more posts “a streak.”) Alternatively, we can look for posts spanning a period of 2 or more weeks. Streaks of “Angry” and “Anxious” were very rare, as were streaks of “Other.”

3 Methods

From our overarching goal of determining what factors contribute to the onset and cessation of depression we can create five questions. First, what are the differences in the dynamics of users who self-report negative affect mood. Second, does depression in fact “settle in,” as we would expect from CBT? Third, can we distinguish negative affect posts from positive affect posts? Fourth, can we predict that a user will continue to experience depressed mood from a few negative affect posts? Fifth, can we predict that a user is about to emerge from depression? This last question will

\(^1\)Thanks to Keng-Hao Chang for providing this dataset.

\(^2\)We found a list of 132 emoticons, though some posts had identifiers for emoticons we were not able to find.
hopefully give us insight into the kinds of thoughts that are characteristic of “positive” thinking, while the fourth question will hopefully give us insight into the kinds of thoughts that lead to long-term depressed mood.

The first two questions can best be answered by shallow analysis of statistics about the self-reported mood. The final three questions require classification. For these tasks, we employ the logistic classifier, though we also tried Naive Bayes and SVMs. The logistic classifier seemed to outperform the others in our early experiments. We made extensive use of the ScalaNLP toolkit\(^3\) for tokenizing, preprocessing, and classification. We use 10-fold cross validation for all experiments. When reporting accuracies, we report both macro- and micro-averaged precision, recall and F1, because of the unbalanced condition of many of our problems.

4 Analysis

Episodes of elevated happiness are more frequent than negative affect posts: 92% of “streaks” of 7 or more posts with the same affect were Happy. In what follows, we analyze coarse transition statistics to get a sense of the dynamics of LiveJournal users’ affect.

4.1 Transitions

We created in Figure 1 a transition diagram of user self-reported moods, including high probability transitions between the most common X% of moods. We used the “twopi” layout tool that is part of graphviz for layout. Based on the research of Wills [1997], “twopi” lays out a graph in concentric circles based on distance to a center node.

Notice that “happy” and “bored” are hubs in the network, with many in-bound transitions. “Bored” itself is the center, which given the assumed nature of the blogging platform’s audience is not terribly surprising.

Some moods—including “happy,” “bored” and “depressed”—are sticky: once inside these moods, users are likely to remain their for consecutive posts. Most positive moods sticky to some degree, while most negative affect moods (except, notably, “depressed”) are not sticky.

“Depressed” in particular is an island of sorts, with few inbound transitions. However, once there, users tend to remain in “depressed” for some time. Other sad affect posts are not sticky: users who are “sad” or “exhausted” typically transition out of those states quickly.

4.2 The Length of Depression

Next, we plot the number of sad streaks binned by their length on a log plot, in Figure 2.\(^4\)

\(^3\)http://www.scalanlp.org
\(^4\)I feel compelled to reference this comic from “Pictures for Sad Children:” http://27.media.tumblr.com/ BTY54d2GDn1hng113NotSrauo1_500.png
Figure 1: Mood Transition Diagram. Thickness of the arrows indicates probability of transition. Larger states are more common. Notice that “happy” and “bored” are central.
Figure 2: Length of Depression, in number of posts. The y-axis is the natural log of the count. Notice that the graph is roughly (log)linear until around a streak of ten.

Note that the graph is roughly log-linear until around 10 consecutive posts, at which point it becomes heavy-tailed. This result indicates that true “depression” does not simply disappear at a constant rate. That is, it does not follow an exponential or geometric distribution. Otherwise, the graph would appear linear on a log-plot. Thus, there is something about long bouts of depression that are inherently different from short term sadness: it is just as easy to exit a negative affect streak on post 7 as it is on post 2, but as depression settles in, it is harder and harder to escape. This is support for the CBT explanation of depression: depression is self-reinforcing, meaning that the longer one is in depression, the harder it is to get out. Therefore, we have some hope of figuring out what distinguishes long depressions from short depressions.

5 Classification Experiments

We now turn to classification. Our aim here is not performance, per se, but analysis: what factors are predictive in determining the length of depression?

5.1 Streak Classification

Our first task is the relatively easy task of streak identification. Can we distinguish positive affect streaks from negative affect streaks? The baseline is fairly high here, with just over 92% of streaks being positive. We built a bag-of-words classifier on the binary classification task of distinguishing negative from positive affect streaks. The results are in Table 1. We get a reasonable improvement over random guessing, though the prediction of depression is surprisingly low. This is likely due to the bias term, since most streaks are just overwhelmingly positive. However, precision is lower than one would anticipate, indicating that the difference between the words one uses when “sad” are surprisingly similar to non-depressed.
<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Micro-Averaged</td>
<td>.953</td>
<td>.953</td>
<td>.953</td>
</tr>
<tr>
<td>Macro-Averaged</td>
<td>.836</td>
<td>.770</td>
<td>.798</td>
</tr>
<tr>
<td>Happy</td>
<td>.968</td>
<td>.983</td>
<td>.975</td>
</tr>
<tr>
<td>Sad</td>
<td>.704</td>
<td>.557</td>
<td>.662</td>
</tr>
</tbody>
</table>

Table 1: Classification performance on the “Streak Classification” task. The majority class gets around 0.92 micro-averaged F1 (i.e. accuracy).

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Micro-Averaged</td>
<td>.771</td>
<td>.771</td>
<td>.771</td>
</tr>
<tr>
<td>Macro-Averaged</td>
<td>.703</td>
<td>.757</td>
<td>.692</td>
</tr>
<tr>
<td>Happy</td>
<td>.728</td>
<td>.948</td>
<td>.824</td>
</tr>
<tr>
<td>Sad</td>
<td>.781</td>
<td>.342</td>
<td>.475</td>
</tr>
</tbody>
</table>

Table 2: Classification performance on the “Depression Prediction” task. The majority class gets around 0.68 micro-averaged F1 (i.e. accuracy).

## 5.2 Depression Prediction

Given a string of three consecutive negative affect posts where the user had not previously been depressed, can we predict the chance that the user will have negative affect for a majority of posts for the next week? We built a dataset for the condition, for just under 102,000 instances, where the predictors were bag-of-words and metadata features based on the three negative affect posts. The dataset was somewhat imbalanced. In 32% of instances, the user continued to experience negative affect in more than half of their posts for the next week. As before, we built a bag-of-words classifier using logistic regression.

The results are in Table 2. As before, the classifier errs on the side of classifying too many posts as positive affect.

While weights from discriminative classifiers are notoriously hard to interpret, nevertheless we note that the 20 predictors with the highest weight (besides the intercept) were all fine-grained self-reported mood features. “Blank,” “drained,” “blah,” and “depressed,” and “numb” were the most negative of these, while “morose,” “rejected,” “lethargic,” and “pessimistic” were least negative. This matches our intuitions about depression: depression is characterized as much by the absence of feeling as by decidedly negative feelings. No moods had positive weight, and the word features do not seem interpretable.

## 5.3 Cessation Prediction

Given a user who has been in a sad streak of 10 or more posts, can we predict if he or she will have positive affect? This task is surprisingly balanced: 5000 transition to at least half positive while 4852 stay depressed. Again, we trained a bag of words classifier. The results are in Table 3.
<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Micro-Averaged</td>
<td>.779</td>
<td>.779</td>
<td>.779</td>
</tr>
<tr>
<td>Macro-Averaged</td>
<td>.703</td>
<td>.670</td>
<td>.682</td>
</tr>
<tr>
<td>Happy</td>
<td>.828</td>
<td>.887</td>
<td>.853</td>
</tr>
<tr>
<td>Sad</td>
<td>.577</td>
<td>.452</td>
<td>.507</td>
</tr>
</tbody>
</table>

Table 3: Classification performance on the “Cessation Prediction” task. The majority class gets 0.507 micro-averaged F1 (i.e. accuracy).

Ideally, the weights from this classifier should have been indicative of what leads to cessation. However, again the weights are not terribly interpretable. Similar patterns are observed in terms of the mood indicators.

6 Things that didn’t work

For the sake of completeness, we briefly list a few things we tried that did not work, which is to say, those things whose performance did not significantly outperform chance.

- Depression Prediction: Predicting onset of depression streak from 1-3 negative affect posts, with depression alternately defined as a streak of consecutive negative affect posts using either number of posts and length in time. Different lengths of time/number of posts were tried.

- Cessation Prediction: Predicting end of depression from a context of 1-5 posts, where cessation is defined as no negative affect posts for at least 7 posts.

These likely did not work because they were too black and white: depression is something that goes away slowly, not all at once. After all, even depressed people have good days and bad days. Perhaps more sophisticated modeling may lead to a better conception of what constitutes depression. We leave that for future work.

7 Conclusion

We presented preliminary results in our investigation of the dynamics of depression in online blogs. We examined challenges to doing this kind of analysis, and examined briefly what factors contributed to depression, at least as analyzable using logistic regression.

In the future, we plan to develop more sophisticated models of the dynamics of depression. In particular, while many users self-report mood, around half do not. We hypothesize that depressed users are more likely to not disclose their “mood,” if only for the simple fact that depression is largely regarded as a weakness by most of the population. Moreover, there are likely false self-reports, and so it might be useful to treat self-reported mood as a lossy indicator of underlying mood, necessitating more complex learning algorithms.
References

AHRQ. Substance abuse and mental health services administration, office of applied studies, national survey on drug use and health., 2011.

