Screening Twitter Users for Depression and PTSD with Lexical Decision Lists

Ted Pedersen
Department of Computer Science
University of Minnesota
Duluth, MN, 55812, USA
tpederse@d.umn.edu

Abstract

This paper describes various systems from the University of Minnesota, Duluth that participated in the CLPsych 2015 shared task. These systems learned decision lists based on lexical features found in training data. These systems typically had average precision in the range of .70 – .76, whereas a random baseline attained .47 – .49.

1 Introduction

The Duluth systems that participated in the CLPsych Shared Task (Coppersmith et al., 2015) explore the degree to which a simple Machine Learning method can successfully identify Twitter users who suffer from Depression or Post Traumatic Stress Disorder (PTSD).

Our approach was to build decision lists of Ngrams found in training Tweets that had been authored by users who had disclosed a diagnosis of Depression or PTSD. The resulting lists were applied to the Tweets of other Twitter users who served as a held–out test sample. The test users were then ranked based on the likelihood that they suffered from Depression or PTSD. This ranking depends on the number of Ngrams found in their Tweets that were associated with either condition.

There were eight different systems that learned decision lists plus one random baseline. The resulting lists are referred to as DecisionList_1 – DecisionList_9, where the system that produced the list is identified by the associated integer. Note that system 9 is a random baseline and not a decision list.

2 Data Preparation

The organizers provided training data that consisted of Tweets from 327 Twitter users who self–reported a diagnosis of Depression, and 246 users who reported a PTSD diagnosis. Each of these users had at least 25 Tweets. There were also Control users identified who were of the same gender and similar age, but who did not have a diagnosis of Depression or PTSD. While each control was paired with a specific user with Depression or PTSD, we did not make any effort to identify or use these pairings.

If a Twitter user has been judged to suffer from either Depression or PTSD, then all the Tweets associated with that user belong to the training data for that condition. This is true regardless of the contents of the Tweets. Thus for many users relatively few Tweets pertain to mental illness, and the rest focus on more general topics. All of the Tweets from the Control users are also collected in their own training set as well.

Our systems only used the text portions of the Tweets, no other information such as location, date, number of retweets, etc. was incorporated. The text was converted to lower case, and any non–alphanumeric characters were replaced with spaces. Thus, hashtags became indistinguishable from text, and emoticons were somewhat fragmented (since they include special characters) but still included as features. We did not carry out any spell checking, stemming, or other forms of normalization.

Then, the Tweets associated with each of the conditions was randomly sorted. The first eight million words of Tweets for each condition were included

Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 46–53,
Denver, Colorado, June 5, 2015.

in the training data for each condition. Any Tweets beyond that were discarded. This cut–off was selected since after pre-processing the smallest portion of the training data (PTSD) included approximately 8,000,000 words. We wanted to have the same amount of training data for each condition so as to simplify the process of feature selection.

3 Feature Identification

The decision lists were made up of Ngrams. Ngrams are defined as sequences of N contiguous words that occur within a single tweet.

Decision lists 3, 6, 7, and 8 used bigram (N == 2) features, while 1, 2, 4, and 5 used all Ngrams in size between 1 and 6. All of the Tweets in the training data for each condition were processed separately by the Ngram Statistics Package (Banerjee and Pedersen, 2003). All Ngrams of the desired size were identified and counted. An Ngram must have occurred at least 50 times more in one condition than the other to be included as a feature. Any Ngram made up entirely of stop words was removed from decision lists 2, 5, 6, and 8. The stoplist comes from the Ngram Statistics Package and consists of 392 common words, as well as single character words.

The task was to rank Twitter users based on how likely they are to suffer from Depression or PTSD. In two cases this ranking is relative to the Control group (DvC and PvC), and in the third case the ranking is between Depression and PTSD (DvP). A separate decision list is constructed for each of these cases as follows. For the condition DvC, the frequencies of the Ngrams from the Depression training data are given positive values, and the Ngrams from the Control data are given negative values. Then, the decision list is constructed by simply adding those values for each Ngram and recording the sum as the weight of the Ngram feature.

For example, if *feel tired* occurred 4000 times in the Depression training data, and 1000 times in the Control data, the final weight of this feature would be 3000. Ngrams with positive values are then indicative of Depression, whereas those with negative values point towards the Control group. An Ngram with a value of 0 would have occurred exactly the same number of times in both the Depression and Control group and would not be indicative of either condition. The same process is followed to create decision lists for PvC and DvP.

system	stoplist?	Ngrams	weights
3	N	2	binary
7	N	2	frequency
1	N	1–6	binary
4	N	1–6	frequency
6	Y	2	binary
8	Y	2	frequency
2	Y	1–6	binary
5	Y	1–6	frequency

Table 1: System Overviews.

Four of the systems limited the Ngrams in the decision lists to bigrams, while four systems used the Ngrams 1–6 as features. In the latter case, the smaller Ngrams that are also included in a longer Ngram are counted both as a part of that longer Ngram, and individually as smaller Ngrams. For example, if the trigram *I am tired* is a feature, then the bigrams *I am* and *am tired* are also features, as are *I*, *am*, *tired*.

4 Running the Decision List

After a decision list is constructed, a held out sample of test users can be evaluated and ranked for the likelihood of Depression and PTSD. The Tweets for an individual user are all processed by the Ngram Statistics Package to identify the Ngrams. Then the Ngrams in a user's Tweets are compared to the decision list and any time a user's Ngram matches the Decision List the frequency associated with that Ngram is added to a running total. Keep in mind that features for one class (e.g., Depression) will add positive values, while features for the other (e.g., Control) will add negative values. This sum is kept as all of an individual user's Tweets are processed, and in the end this sum will have either a positive or negative value that will determine the the class of the user. The raw score is used to rank the different users relative to each other.

There is also a binary weighting variation. In this case when a user's Ngram is encountered in the Decision list, if the frequency is positive a value of 1 is added to the running together, and if it is negative a value of -1 is added. This is done for all of a user's

	DvP		DvC		PvC	
rank	id	prec	id	prec	id	prec
1	2	.769	2	.736	1	.721
2	5	.764	1	.731	2	.720
3	4	.761	3	.718	3	.708
4	1	.760	8	.718	6	.704
5	8	.738	6	.718	7	.607
6	7	.731	7	.713	8	.572
7	6	.730	4	.713	4	.570
8	3	.724	5	.710	5	.539
9	9	.471	9	.492	9	.489

Table 2: System Precision per Condition.

system	DvC	DvP	PvC
1	20,788	23,552	19,973
4	20,788	23,552	19,973
2	18,617	21,145	17,936
5	18,617	21,145	17,936
3	5,704	6,385	6,068
7	5,704	6,385	6,068
6	4,442	4,998	4,747
8	4,442	4,998	4,747

Table 3: Number of Features per Decision List.

Tweets, and then whether this value is positive or negative indicates the class of the user.

Table 1 briefly summarizes the eight decision list systems. These systems vary in three respects :

- Whether the stoplist is used (Y or N),
- the length of the Ngrams used (2 or 1–6), and
- the type of weighting (binary or frequency).

All eight possible combinations of these settings were utilized.

5 Results

Table 2 shows the average precision per system for each of the three conditions.

Table 4 shows the average rank and precision attained by each system across all three conditions. It also lists the characteristics of each decision list.

When taken together, Tables 2 and 4 clearly show that systems 2 and 1 are the most effective across the three conditions. These two systems are identical, except that 2 uses a stoplist and 1 does not. They both use the binary weighting scheme and Ngrams of size 1–6.

Table 3 shows the number of features per decision list. The systems that use the ngram 1–6 features (1, 2, 4, 5) have a much larger number of features than the bigram systems (3, 6, 7, 8). Note however that in Table 2 there is not a strong correlation between a larger number of features and improved precision. While systems 1 and 2 have the highest precision (and the largest number of features) systems 4 and 5 have exactly the same features and yet attain average precision that is quite a bit lower than systems with smaller numbers of features, such as 3 or 6.

Note that the pairs of systems that have the same number of features in the decision list only differ in their weighting scheme (bigram versus frequency) and so the number of features would be expected to be the same. Also note that the number of features per condition for a given system is approximately the same – this was our intention when selecting the same number of words (8,000,000) per condition from the training data.

6 Decision Lists

Below we show the top 100 entries in each decision list created by system 2, which had overall the highest precision of our runs.

System 2 uses Ngrams of size 1–6 with stop words removed and binary weighting of features. The decision lists below show the Ngram feature and the frequency in the training data. Note that Ngrams that begin with u and are followed by numeric values (e.g., u2764, u201d, etc.) are emoticon encodings.

All of the decision lists include a mixture of standard English features and more Web specific features, such as portions of URLs and more notably emoticons. Our systems treated these like any other Ngram, and so a series of emoticons will appear as an Ngram, and URLs are broken into fragments which appears as Ngrams.

6.1 Decision List system 2, DvC

This decision list has 18,617 entries, the first 100 of which are shown below. This decision list attained average precision of 77%.

Features and positive counts in **bold** indicate De-

system	avg rank	ranks DvP, DvC, PvC	avg precision	stoplist?	Ngrams	weights
2	1.3	1, 1, 2	.742	Y	1–6	binary
1	2.3	4, 2, 1	.737	N	1–6	binary
3	4.7	8, 3, 3	.717	N	2	binary
6	5.3	7, 5, 4	.717	Y	2	binary
7	5.7	6, 6, 5	.684	N	2	frequency
4	5.7	3, 7, 7	.681	N	1–6	frequency
8	5.0	5, 4, 6	.676	Y	2	frequency
5	6.0	2, 8, 8	.671	Y	1–6	frequency
9	9.0	9, 9, 9	.484			

Table 4: Average Rank and Precision over all Conditions.

pression, while those in *italics* are negative counts that are associated with the Control.

http -26084; http t co -23935; http t -23906; co -22388; t co -22210; ud83d -20341; **ud83c 15764;** *lol -9429;* **please 8166;** *u2764 u2764 -8127; u2764 u2764 u2764 -8017; u2764 u2764 u2764 u2764 -7947; u2764 u2764 u2764 u2764 u2764 -7852; u2764 -7769; u2764 u2764 u2764 u2764 u2764 u2764 -7767; gt -7078;* **love 6041;** *u201c -5815; u201d -5635;* **follow 5578;** *amp -5420; gt gt -5237;* **ufe0f 5138; re 4875;** *ud83d ude02 -4841; ude02 -4839; photo -4791;* **fucking 4616; love you 4603; im 4542;** *u0627 -4412; rt -4132;* **udf38 4046; ud83c udf38 4046; udc95 4033; ud83d udc95 4033; u043e 3879; you re 3681; u0430 3666; ve 3624; pj3l408vwlgs3 3606; don t 3563; udf41 3543; ud83c udf41 3542; u0435 3530;** *ud83d ude02 ud83d -3529; ude02 ud83d -3528; gt gt gt -3459;* **fuck 3372; please follow 3359;** *check -3357; ud83d ude02 ud83d ude02 -3355; ude02 ud83d ude02 -3354;* **don 3298; i love 3284; u2661 3088; udf38 ud83c 3058; ud83c udf38 ud83c 3058; i don 3020; i don t 2976; i ve 2962; udc95 ud83d 2922; ud83d udc95 ud83d 2922; u0438 2905; feel 2818;** *u0644 -2733; check out -2703;* **udc95 ud83d udc95 2687; ud83d udc95 ud83d udc95 2687;** *photo http t co -2684; photo http -2684; photo http t -2683;* **u043d 2581; follow me 2517; udc95 ud83d udc95 ud83d 2511; ud83d udc95 ud83d udc95 ud83d 2511; udc95 ud83d udc95 ud83d udc95 2464; ud83d udc95 ud83d udc95 ud83d udc95 2464; u0442 2405;** *lt lt -2376;* **i love you 2371;** *today -2365;* **udc95 ud83d udc95 ud83d udc95 ud83d 2322; u0440 2289; b4a7lkokrkpq 2260; udf38 ud83c udf38 2236; ud83c udf38 ud83c udf38 2236; inbox 2218; mean 2172; udf0c 2148; ud83c udf0c 2148;** *ud83d ude02 ud83d ude02 ud83d -2147; ude02 ud83d ude02 ud83d -2146;* **ni 2142; oh 2114;** *ud83d ude02 ud83d ude02 ud83d ude02 ud83d ude02 -2101; ude02 ud83d ude02 ud83d ude02 -2100;* **u0441 2075; udf41 ud83c 2074; ud83c udf41 ud83c 2074;**

6.2 Decision List system 2, PvC

This decision list has 17,936 entries, the first 100 of which are shown below. This decision list attained average precision of 74%.

Features and positive counts in **bold** indicate PTSD, while those in *italics* are negative counts that are associated with the Control.

ud83d -82824; rt -20230; ude02 -14516; ud83d ude02 -14516; **u2026 12941;** *gt -12727; u2764 -10630; lol -9932; u201c -9736; ude02 ud83d -9112; ud83d ude02 ud83d -9112; u201d -8962; gt gt -8947; u2764 u2764 -8753; u2764 u2764 u2764 -8425; u2764 u2764 u2764 u2764 -8217; u2764 u2764 u2764 u2764 u2764 -8064; u2764 u2764 u2764 u2764 u2764 u2764 -7940; ude02 ud83d ude02 -7932; ud83d ude02 ud83d ude02 -7932;* **co 7291; t co 7140;** *ud83c -6306; gt gt gt -6171; love -5322; ude02 ud83d ude02 ud83d -5165; ud83d ude02 ud83d ude02 ud83d -5165; ude0d -5058; ud83d ude0d -5056; ude02 ud83d ude02 ud83d ude02 -4901; ud83d ude02 ud83d ude02 ud83d ude02 -4901;* **u043e 4877; u0430 4485;** *u0627 -4251;* **u0435 4241; thank 4109; thank you 4079;**

gt gt gt gt -3936; im -3843; ude18 -3617; ud83d ude18 -3617; **please 3533; u0438 3526;** *shit -3337; don -3288;* **health 3277;** *don t -3262; lt -3259; haha -3175; lt lt -3172; ude02 ud83d ude02 ud83d ude02 ud83d -3094;* **u043d 3074; u0442 3065; answer 2998; my answer 2963; http 2937;** *ude29 -2932; ud83d ude29 -2932;* **answer on 2930; tgtz to 2929; tgtz 2929; on tgtz to 2929; on tgtz 2929; my answer on tgtz to 2929; my answer on tgtz 2929; my answer on 2929; answer on tgtz to 2929; answer on tgtz 2929;** *ude2d -2911; ud83d ude2d -2911; wanna -2873; day -2869; miss -2868;* **u0440 2855;** *nigga -2798; gt gt gt gt gt -2673; u0644 -2632; udc4c -2607; ud83d udc4c -2607;* **u0441 2581;** *ude0d ud83d -2574; ud83d ude0d ud83d -2572;* **ptsd 2550; amp 2534; bqtn0bi 2510; help 2459;** *ude12 -2438; ud83d ude12 -2438; bitch -2433; girl -2398; school -2395; ass -2355; lmao -2288;* **u0432 2274;** *hate -2267; ain -2259; ain t -2258; i love -2256; lt lt lt -2242;* **nhttp 2226;**

6.3 Decision List system 2, DvP

This decision list has 21,145 entries, the first 100 of which are shown below. This decision list attained average precision of 72%.

Features and positive counts in **bold** indicate Depression, while those in *italics* are negative counts that are associated with PTSD.

ud83d 62483; *co -29679; t co -29350; http -29021; http t -26110; http t co -24404;* **ud83c 22070; rt 16098;** *u2026 -13855;* **love 11363; ude02 9677; ud83d ude02 9675; im 8385;** *amp -7954;* **follow 6927; don t 6825; don 6586; love you 6330; gt 5649; ude02 ud83d 5584; ud83d ude02 ud83d 5583; i love 5540; ufe0f 5069; pj3l408vwlgs3 4806; please 4633; ude02 ud83d ude02 4578; udc95 4577; ud83d ude02 ud83d ude02 4577; ud83d udc95 4577; ude0d 4564; ud83d ude0d 4564; fuck 4474; re 4247; udf38 4112; ud83c udf38 4112; i don t 3939; u201c 3921; i don 3882; you re 3770; gt gt 3710; shit 3695; udf41 3604; ud83c udf41 3603; follow me 3547; please follow 3506;** *news -3499;* **fucking 3499; hate 3491; u2661 3483; wanna 3410;** *thanks -3370;* **u201d 3327; i love you 3276; school 3262;** *answer -3108;* **udc95 ud83d 3104; ud83d udc95 ud83d 3104; gonna 3103; udf38 ud83c 3068; ud83c udf38 ud83c 3068;** *health -3025;* **ude02 ud83d ude02 ud83d 3019; ud83d ude02 ud83d ude02 ud83d 3018; feel 2987;** *my answer -2977;* **people 2932;** *answer on -2930; tgtz to -2929; tgtz -2929; on tgtz to -2929; on tgtz -2929; my answer on tgtz to -2929; my answer on tgtz -2929; my answer on -2929; answer on tgtz to -2929; answer on tgtz -2929;* **b4a7lkokrkpq 2875; u2764 2861; omg 2852; ude02 ud83d ude02 ud83d ude02 2801; ud83d ude02 ud83d ude02 ud83d ude02 2800; udc95 ud83d udc95 2782; ud83d udc95 ud83d udc95 2782;** *thank -2759; photo -2749;* **gt gt gt 2712;** *great -2623;* **ude2d 2618; ud83d ude2d 2616; udc95 ud83d udc95 ud83d 2590; ud83d udc95 ud83d udc95 ud83d 2590;** *thank you -2587;* **ude0d ud83d 2541; ud83d ude0d ud83d 2541; udc95 ud83d udc95 ud83d udc95 2535; ud83d udc95 ud83d udc95 ud83d udc95 2535;** *bqtn0bi -2533; nhttp -2525;* **harry 2506;** *ptsd -2502;*

7 Indicative Features

The following results show the top 100 most frequent Ngram features from the training data that were also used in the Tweets of the user with the highest score for each of the conditions. Recall that for system 2 the weighting scheme used was binary, so these features did not have any more or less value than others that may have been less frequent in the training data. However, given that each decision list had thousands of features 3, this seemed like a reasonable way to give a flavor for the kinds of features that appeared both in the training data and in users' Tweets. While not definitive, this will hopefully provide some insight into which of the decision list features play a role in determining if a user may have a particular underlying condition. Note that the very long random alpha strings are anonymized Twitter user ids.

7.1 Decision List system 2, DvC

This user used 3,267 features found in our decision list, where 2,360 of those were indicative of Depression, and 907 for Control. This gives this user a score of 1,453 which was the highest among all users for Depression. What follows are the 100 most frequent features from the training data that are indicative of Depression that this user also employed in a tweet at least one time.

ud83c; please; love; follow; re; fucking; love you; im; udf38; ud83c udf38; udc95; ud83d udc95; you re; ve; don t; fuck; please follow; don; i love; u2661; udf38 ud83c; ud83c udf38 ud83c; i don; i don t; i ve; udc95 ud83d; ud83d udc95 ud83d; feel; i love you; udf38 ud83c udf38; ud83c udf38 ud83c udf38; mean; ni; oh; think; why; actually; guys; i ll; omg; ll; lt 3; n ud83c; people; hi; 3; udf38 ud83c udf38 ud83c; ud83c udf38 ud83c udf38 ud83c; https; https t; https t co; udf38 ud83c udf38 ud83c udf38; ud83c udf38 ud83c udf38 ud83c udf38; sorry; okay; gonna; love you so; thank you; i feel; bc; this please; otygg6_yrurxouh; would mean; i hope; loves; thank; love you so much; pretty; friend; u2022; xx; cute; hope; hate; boys; depression; life; udf38 ud83c udf38 ud83c udf38 ud83c; a lot; she loves; perfect; u2014; oh my; lot; i think; thing; help; literally; u2661 u2661; the world; ve been; yeah; they re; still; it would mean; my life; friends; the fuck; crying; nplease

7.2 Decision List system 2, PvC

This user used 3,896 features found in our decision list, where 2,698 of those were indicative of PTSD, and 1,198 of Control. This gives this user a score of 1,500 which was the highest among all users for PTSD. What follows are the 100 most frequent features from the training data that are indicative of PTSD that this user also employed in a tweet at least one time.

u2026; co; t co; thank; thank you; please; health; answer; http; ptsd; amp; bqtn0bi; help; nhttp; ve; http t; https; nhttp t; https t; nhttp t co; https t co; read; medical; thanks; women; obama; i ve; ebola; oxmljtykruvsnpd; tcot; think; http u2026; curp4uo6ffzn2x1qckyok78w2hl u2026; news; thanks for; fbi; ferguson; children; support; mental; mentalhealth; story; curp4uo6ffzn2x1qckyok78w2hl; fucking; hope; living; http http t co; http http t; http http; auspol; sign; war; veterans; police; freemarinea; i think; bbc; god; woman; men; 2014; white; great; found; child; ago; drugs; kind; book; report; thank you for; n nhttp; agree; healthy; military; ppl; sure; n nhttp t; dvfrpdjwn4z; n nhttp t co; please check; care; writing; please check out; america; israel; tcot http; law; please check out my; bqtn0bi tcot; lot; son; kids; tcot http t; uk; isis; homeless; petition; the fbi; daughter

7.3 Decision List system 2, DvP (Depression)

This user used 3,797 features found in our decision list, where 2,945 of those were indicative of Depression, and 852 for PTSD. This gives this user a score of 2,093 which was the highest among all users for Depression when gauged against PTSD. Note that this is a different user than scored highest in DvC. What follows are the 100 most frequent features from the training data that are indicative of Depression as opposed to PTSD that this user also employed in a tweet at least one time.

ud83d; ud83c; rt; love; ude02; ud83d ude02; im; follow; don t; don; love you; gt; ude02 ud83d; ud83d ude02 ud83d; i love; ufe0f; please; ude02 ud83d ude02; udc95; ud83d ude02 ud83d ude02; ud83d udc95; ude0d; ud83d ude0d; fuck; re; udf38; ud83c udf38; i don t; u201c; i don; you re; gt gt; shit; udf41; ud83c udf41; follow me; fucking; hate; u2661; wanna; u201d; i love you; school; udc95 ud83d; ud83d udc95 ud83d; gonna; ude02 ud83d ude02 ud83d; ud83d ude02 ud83d ude02 ud83d; feel; people; u2764; omg; ude02 ud83d ude02 ud83d ude02; ud83d ude02 ud83d ude02 ud83d ude02; gt gt gt; ude2d; ud83d ude2d; ude0d ud83d; ud83d ude0d ud83d; happy; guys; oh; girl; mean; cute; i hate; girls; okay; why; ude18; ud83d ude18; udf41 ud83c; ud83c udf41 ud83c; n ud83c; boys; udf42; ud83c udf42; ude02 ud83d ude02 ud83d ude02 ud83d; bitch; bc; gt gt gt gt; perfect; miss; love you so; sleep; ude0d ud83d ude0d; ud83d ude0d ud83d ude0d; ude12; ud83d ude12; night; ni; u2022; life; i feel; wait; my life; ur; day; u263a; hi

7.4 Decision List system 2, DvP (PTSD)

This user used 4,167 features found in our decision list, where 2,885 of those were indicative of PTSD, and 1,282 for Depression. This gives this user a score of 1,603 which was the highest among all users for Depression when gauged against PTSD. Note that this is the same user that scored highest in PvC. What follows are the 100 most frequent features from the training data that are indicative of PTSD as opposed to Depression that this user also employed in a tweet at least one time.

co; t co; http; http t; http t co; u2026; amp; news; thanks; answer; health; thank; photo; great; thank you; bqtn0bi; nhttp; ptsd; obama;

nhttp t; nhttp t co; thanks for; medical; u2019s; read; women; tcot; curp4uo6ffzn2x1qckyok78w2hl; curp4uo6ffzn2x1qckyok78w2hl u2026; oxmljtykruvsnpd; check; fbi; http u2026; ebola; today; ppl; help; support; ferguson; check out; police; sign; book; veterans; work; blog; children; war; 2; country; gop; living; thanks for the; report; freemarinea; auspol; u2019t; military; media; bbc; woman; house; men; u2026 http; truth; white; u2026 http t; u2026 http t co; http http; http http t; http http t co; posted; n nhttp; son; story; a great; photo http; n nhttp t; photo http t; photo http t co; law; n nhttp t co; healthy; america; dvfrpdjwn4z; state; tcot http; agree; mt; government; please check; god; kids; share; please check out; tcot http t; way; please check out my; case; bqtn0bi tcot

8 Discussion and Conclusions

This was our first effort at analyzing text from social media for mental health indicators. Our system here was informed by our experiences in other shared tasks for medical text, including the i2b2 Smoking Challenge (Pedersen, 2006; Uzuner et al., 2008), the i2b2 Obesity Challenge (Pedersen, 2008; Uzuner, 2009), and the i2b2 Sentiment Analysis of Suicide Notes Challenge (Pedersen, 2012; Pestian et al., 2012).

In those shared tasks we frequently observed that rule based systems fared reasonably well, and that machine learning methods were prone to over–fitting training data, and did not generalize terribly well. For this shared task we elected to take a very simple machine learning approach that did not attempt to optimize accuracy on the training data, in the hopes that it would generalize reasonably well.

However, this task is quite distinct in that the data is from Twitter. In the other shared tasks mentioned data came either from discharge notes, or suicide notes, all of which were generally written in standard English. We did not attempt to normalize abbreviations or misspellings, and we did not handle emoticons or URLs any differently than ordinary text. We also did not utilize any of the information available from Tweets beyond the text itself. These are all issues we plan to investigate in future work.

While it was clear that the Ngram 1–6 features performed better than bigrams, it would be interesting to know if the increased accuracy came from a particular length of Ngram, or if all the different Ngrams contributed equally to the success of Ngram 1–6. In particular we are curious as to whether or not the unigram features actually had a positive impact, since unigrams may tend to be both noisier and more semantically ambiguous.

Likewise, the binary weighting was clearly superior to the frequency based method. It seems important to know if there are a few very frequent features that are skewing these results, or if there are other reasons for the binary weighting to result in such better performance.

While is it difficult to generalize a great deal from these findings, there is some anecdotal evidence that these results have some validity. First, the user that was identified as most prone to Depression when compared to Control (in DvC) was different from the user identified as most prone to Depression when compared to PTSD (in DvP). This seems consistent with the idea that a person suffering from PTSD may also suffer from Depression, and so the DvC case is clearly distinct from the DvP since in the latter there may be confounding evidence of both conditions.

In reviewing the decision lists created by these systems, as well as the features that are actually found in user's Tweets, it seems clear that there were many somewhat spurious features that were included in the decision lists. This is not surprising given that features were included simply based on their frequency of occurrence - any Ngram that occurred 50 times more in one condition than the other would be included as a feature in the decision list. Moving forward having a more selective method for including features would surely help improve results, and provide greater insight into the larger problem of identifying mental illness in social media postings.

Acknowledgments

My thanks go to the CLPsych 2015 organizers for creating a very interesting and compelling task. This was not only a lot of fun to work on, but really presented some new and exciting challenges that will no doubt inspire a great deal of future work.

References

Satanjeev Banerjee and Ted Pedersen. 2003. The design, implementation, and use of the Ngram Statistics Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 370–381, Mexico City, February.

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.

Ted Pedersen. 2006. Determining smoker status using supervised and unsupervised learning with lexical features. In *Working Notes of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, Washington, DC, November.

Ted Pedersen. 2008. Learning high precision rules to make predictions of morbidies in discharge summaries. In *Proceedings of the Second i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, Washington, DC, November.

Ted Pedersen. 2012. Rule-based and lightly supervised methods to predict emotions in suicide notes. *Biomedical Informatics Insights*, 2012:5 (Suppl. 1):185–193, January.

John Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, Kevin Cohen, John Hurdle, and Chris Brew. 2012. Sentiment analysis of suicide notes: A shared task. *Biomedical Informatics Insights*, 2012:5 (Suppl. 1):3–16, January.

Ozlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2008. Identifying patient smoking status from medical discharge records. *Journal of the Medical Informatics Association*, 15(1):14–24.

Ozlem Uzuner. 2009. Recognizing obesity and co–morbidities in sparse data. *Journal of the Medical Informatics Association*, 16(4):561–570.