CLPsych 2015 Shared Task: Depression and PTSD on Twitter

Glen Coppersmith
Qntfy
glen@qntfy.io

Mark Dredze
Johns Hopkins University
mdredze@cs.jhu.edu

Craig Harman
Johns Hopkins University
charman@jhu.edu

Kristy Hollingshead
IHMC
kseitz@ihmc.us

Margaret Mitchell
Microsoft Research
memitc@microsoft.com

Abstract

This paper presents a summary of the Computational Linguistics and Clinical Psychology (CLPsych) 2015 shared and unshared tasks. These tasks aimed to provide apples-to-apples comparisons of various approaches to modeling language relevant to mental health from social media. The data used for these tasks is from Twitter users who state a diagnosis of depression or post traumatic stress disorder (PTSD) and demographically-matched community controls. The unshared task was a hackathon held at Johns Hopkins University in November 2014 to explore the data, and the shared task was conducted remotely, with each participating team submitted scores for a held-back test set of users. The shared task consisted of three binary classification experiments: (1) depression versus control, (2) PTSD versus control, and (3) depression versus PTSD. Classifiers were compared primarily via their average precision, though a number of other metrics are used along with this to allow a more nuanced interpretation of the performance measures.

1 Introduction

Language is a major component of mental health assessment and treatment, and thus a useful lens for mental health analysis. The psychology literature has a long history of studying the impact of various mental health conditions on a person's language use. More recently, the computational linguistics community has sought to develop technologies to address clinical psychology challenges. Some of this work has appeared at the Computational Linguistics and Clinical Psychology workshops (Resnik et al., 2014; Mitchell et al., 2015).

The 2015 workshop hosted a shared and unshared task. These tasks focused on fundamental computational linguistics technologies that hold promise to improve mental health-related applications; in particular, detecting signals relevant to mental health in language data and associated metadata. Specifically, technologies that can demonstrably separate community controls from those with mental-health conditions are extracting signals relevant to mental health. Examining the signals those techniques extract and depend on for classification can yield insights into how aspects of mental health are manifested in language usage. To that end, the shared and unshared tasks examined Twitter users who publicly stated a diagnosis of depression or PTSD (and age- and gender-matched controls).

Shared tasks are tools for fostering research communities and organizing research efforts around shared goals. They provide a forum to explore new ideas and evaluate the best-of-breed, emerging, and wild technologies. The 2015 CLPsych Shared Task consisted of three user-level binary classification tasks: PTSD vs. control, depression vs. control, and PTSD vs. depression. The first two have been addressed in a number of settings (Coppersmith et al., 2015; Coppersmith et al., 2014b; Coppersmith et al., 2014a; Resnik et al., 2013; De Choudhury et al., 2013; Rosenquist et al., 2010; Ramirez-Esparza et al., 2008), while the third task is novel. Organizing this shared task brought together many teams to consider the same problem, which had the benefit of establishing a solid foundational understanding, common standards, and a shared deep understanding of both task and data.

Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 31–39,
Denver, Colorado, June 5, 2015.

The unshared task (affectionately the "hackathon") was a weekend-long event in November 2014 hosted by Johns Hopkins University. The hackathon provided data similar to the shared task data and encouraged participants to explore new ideas. In addition to starting new research projects, some of which were subsequently published in the CLPsych workshop, the event laid the foundation for the shared task by refining task definitions and data setup.

This paper summarizes both the shared and unshared tasks at the 2015 Computational Linguistics and Clinical Psychology workshop. We outline the data used for these tasks, and summarize the methods and common themes of the shared task participants. We also present results for system combination using the shared task submissions.

2 Shared Task Data

Data for the shared task are comprised of public tweets collected according to the procedures of Coppersmith et al. (2014a). We briefly describe the procedure here, and refer interested readers to Coppersmith et al. (2014a). for details.

Users of social media may publicly discuss their health for a variety of reasons, such as to seek treatment or health advice. More specifically to mental health, users may choose a public forum to fight the societal stigma associated with mental illness, or to explain certain behaviors to friends. Many users tweet statements of diagnosis, such as "I was just diagnosed with *X* and ...", where *X* is a mental health condition. While this can include a large variety of mental health conditions (Coppersmith et al., 2015), the shared task considered two conditions: depression or PTSD. We chose these conditions since they are among the most common found in Twitter and have relatively high prevalence compared to other conditions. A human annotator evaluates each such statement of diagnosis to remove jokes, quotes, or any other disingenuous statements. For each user, up to their most recent 3000 public tweets were included in the dataset. Importantly, we removed the tweet in which the genuine statement of diagnosis was found, to prevent any artifact or bias created from our data sampling technique. However, some of these users do mention their condition in other tweets, and some approaches may be influenced by this phenomenon. To ensure that each included user has a sufficient amount of data, we ensured that each user has at least 25 tweets and that the majority of them are English (75% according to the Compact Language Detector[1]).

2.1 Age- and Gender-Matched Controls

A goal of the shared task is to differentiate users with a mental health diagnosis from those who do not. To that end, the shared task data included a set of randomly selected Twitter users.

Age and gender play a significant role in many mental health conditions, making certain segments of the population more or less likely to be affected or diagnosed with them. When possible, demographic variables such as age and gender are controlled for when doing clinical psychology or mental health research. Few studies looking at social media and clinical psychology have done analysis with explicit matched samples, though some have done this implicitly by examining a segment of the population, (e.g., college students (Rude et al., 2004)). Some work in social media analysis has considered the effect of matched samples (Dos Reis and Culotta, 2015).

To create age- and gender-matched community controls, we estimated the age and gender of each user in our sample through analysis of their language. We used the demographic classification tool from the World Well-Being Project (Sap et al., 2014)[2]. For each depression and PTSD user we estimated their gender, forcing the classifier to make a binary decision as to whether the user was 'Female' or 'Male', and used the age estimate as-is (an ostensibly continuous variable). We did the same for a pool of control users who tweeted during a two week time period in early 2013 and met the criteria set out above (at least 25 Tweets and their tweets were labeled as at least 75% English). To obtain our final data set, for each user in the depression or PTSD class, we sampled (without replacement) a paired community control user of the same estimated gender with the closest estimate age.

We expect (and have some anecdotal evidence)

[1] https://code.google.com/p/cld2/

[2] http://wwbp.org/

that some of the community controls suffer from depression or PTSD, and made no attempt to remove them from our dataset. If we assume that the rate of contamination in the control users is commensurate with the expected rate in the population, that would mean that this contamination makes up a small minority of the data (though a nontrivial portion of the data, especially in the case of depression).

2.2 Anonymization

Per research protocols approved by the Johns Hopkins University Institutional Review Board, the data was anonymized to protect the identity of all users in the dataset. We used a whitelist approach to allow only certain kinds of information to be maintained, as they posed minimal risk of inadvertently exposing the identity of the user. We kept unedited the timestamp and the language identification of the text. For metadata about the user, we kept the number of friends, followers, and favorites the user has, the time zone the user has set in their profile, and the time their account was created. Screen names and URLs were anonymized (via salted hash), so they were replaced with a seemingly-random set of characters. This procedure was applied to the text content and all the metadata fields (to include embedded tweets such as retweets and replies). This was done systematically so the same set of random characters was used each time a given screen name or URL was used. This effectively enabled statistics such as term frequency or inverse document frequency to be computed without revealing the identity of the user or URL (which sometimes provided a link to an identifiable account name, within or outside of Twitter). Some of Twitter's metadata uses character offsets into the text to note positions, so our anonymized hashes were truncated to be the same number of characters as the original text (e.g., `@username` became `@lkms23sO`). For URLs, we left the domain name, but masked everything beyond that: (e.g., `http://clpsych.org/shared_task/` became `http://clpsych.org/sijx0832aKxP`). Any other metadata that did not match the whitelisted entries or the fields subject to anonymization was removed altogether – this includes, for example, any geolocation information and any information about what devices the user tweets from.

Shared task participants each signed a privacy agreement and instituted security and protective measures on their copy of the data. Participants were responsible for obtaining ethics board approval for their work in order to obtain the shared task data. Data was distributed in compliance with the Twitter terms of service.

3 Shared Task Guidelines

The shared task focused on three binary classification tasks.

1. Identify depression users versus control users.
2. Identify PTSD users versus control users.
3. Identify depression users versus PTSD users.

Twitter users were divided into a train and test partition that was used consistently across the three tasks. The train partition consisted of 327 depression users, 246 PTSD users, and for each an age- and gender-matched control user, for a total of 1,146 users. The test data contained 150 depression users, 150 PTSD users, and an age- and gender-matched control for each, for a total of 600 users. Shared task participants were provided with user data and associated labels (depression, PTSD, or control) for the users contained in the train partition. Participants were given user data without labels for the test partition.

Participants were asked to produce systems using only the training data that could provide labels for each of the three tasks for the test data. Participants used their systems to assign a numeric real-valued score for each test user for each of the three tasks. Each participating team submitted three ranked lists of the 600 test users, one list for each task. Given that machine-learning models often have a number of parameters that alter their behavior, sometimes in unexpected ways, participants were encouraged to submit multiple parameter settings of their approaches, as separate ranked lists, and the best-performing of these for each task would be taken as the "official" figure of merit.

Evaluation was conducted by the shared task organizers using the (undistributed) labels for the test users. During evaluation, irrelevant users were removed; i.e., for PTSD versus control, only 300 users

were relevant for this condition: the 150 PTSD users and their demographically matched controls. The depression users and their demographically matched controls were removed from the ranked list prior to evaluation.

Each submission was evaluated using several metrics. Our primary metric was average precision, which balances precision with false alarms, though this only tells a single story about the methods examined. We also evaluated precision at various false alarm rates (5%, 10%, and 20%) to provide a different view of performance. The reader will note that the highest-performing technique varied according to the evaluation measure chosen – a cautionary tale about the importance of matching evaluation measure to the envisioned task.

3.1 Data Balance

We decided to distribute data that reflected a balanced distribution between the classes, rather than a balance that accurately reflects the user population, i.e., one that has a larger number of controls. This decision was motivated by the need for creating a dataset maximally relevant to the task, as well as limitations on data distribution from Twitter's terms of service. A balanced dataset made some aspects of the shared task easier, such as classifier creation and interpretation. However, it also means that results need to be examined with this caveat in mind. In particular, the number of false alarms expected in the general population is much larger than in our test sample (7-15 times as frequent). In effect, this means that when examining these numbers, one must remember that each false alarm could count for 7-15 false alarms in a more realistic setting. Unfortunately, when this fact is combined with the contamination of the training data by users diagnosed (but not publicly stating a diagnosis of) depression or PTSD, it quickly becomes difficult or impossible to reliably estimate the false alarm rates in practice. A more controlled study is required to estimate these numbers more accurately. That said, the relative rankings of techniques and approaches is not subject to this particular bias: each system would be affected by the false alarm rates equally, so the relative ranking of approaches (by any of the metrics investigated) does provide a fair comparison of the techniques.

4 Shared Task Submissions

We briefly describe the approaches taken by each of the participants, but encourage the reader to examine participant papers for a more thorough treatment of the approaches.

4.1 University of Maryland

UMD examined a range of supervised topic models, computed on subsets of the documents for each user. Particularly, they used a variety of supervised topic-modeling approaches to find groups of words that had maximal power to differentiate between the users for each classification task. Moreover, rather than computing topics over two (typical) extreme cases – treating each tweet as an individual document or treating each users's tweets collectively as a single document (concatenating all tweets together) – they opted for a sensible middle ground of concatenating all tweets from a given week together as a single document (Resnik et al., 2015).

4.2 University of Pennsylvania, World Well-Being Project

The WWBP examined a wide variety of methods for inferring topics automatically, combined with binary unigram vectors (i.e., "did this user ever use this word?"), and scored using straightforward regression methods. Each of these topic-modeling techniques provided a different interpretation on modeling what groups of words belonged together, and ultimately may provide some useful insight as to which approaches are best at capturing mental health related signals (Preotiuc-Pietro et al., 2015).

4.3 University of Minnesota, Duluth

The Duluth submission took a well-reasoned rule-based approach to these tasks, and as such provides a point to examine how powerful simple, raw language features are in this context. Importantly, the Duluth systems allow one to decouple the power of an open vocabulary approach, quite independent of any complex machine learning or complex weighting schemes applied to the open vocabulary (Pedersen, 2015).

4.4 MIQ – Microsoft, IHMC, Qntfy

We include a small system developed by the organizers for this shared task to examine the effect of pro-

viding qualitatively different information from the other system submissions. In this system, which we will refer to as the MIQ[3] (pronounced 'Mike') submission, we use character language models (CLMs) to assign scores to individual tweets. These scores indicate whether the user may be suffering from PTSD, depression, or neither.

The general approach is to examine how likely a sequence of characters is to be generated by a given type of user (PTSD, depression, or control). This provides a score even for very short text (e.g., a tweet) and captures local information about creative spellings, abbreviations, lack of spaces, and other textual phenomena resulting from the 140-character limit of tweets (McNamee and Mayfield, 2004). At test time, we search for sequences of tweets that look "most like" the condition being tested (PTSD or depression) by comparing the condition and control probabilities estimated from the training data for all the n-grams in those tweets.

In more detail, we build a CLM for each condition using the training data. For each user at test time, we score each tweet based on the character n-grams in the tweet C with the CLMs for conditions A and B as $\frac{\sum_C \log p(c_A) - \log p(c_b)}{|C|}$, where $p(c_A)$ is the probability of the given n-gram c according to the CLM model for condition A, and $p(c_B)$ is the probability according to the CLM for condition B. We then compute a set of aggregate scores from a sliding window of 10 tweets at a time, where the aggregate score is either the mean, median, or the proportion of tweets with the highest probability from the CLM for condition A ('proppos'). To compute a single score for a single user, we take the median of the aggregate scores. This follows previous work on predicting depression and PTSD in social media (Coppersmith et al., 2014a; Coppersmith et al., 2014b). We also experimented with excluding or including tweets that heuristically may not have been authored by the Twitter account holder – specifically, this exclusion removes all tweets with URLs (as they are frequently prepopulated by the website hosting the link) and retweets (as they were authored by another Twitter user). We created 12 system submissions using: n-grams of length 5 and 6 (two approaches) crossed with the mean, median, and proppos aggregation approaches (three approaches), and with or without exclusion applied (two approaches).

The top systems for Depression versus Control used 5-grams, proppos and 5-grams, mean. The top system for PTSD versus Control used 5-grams, median, no exclusion. And the top systems for Depression versus PTSD used 6-grams, mean and 6-grams, proppos.

5 Results

We examine only the best-performing of each of the individual system submissions for each binary classification task, but again encourage the reader to examine the individual system papers for a more detailed analysis and interpretation for what each of the teams did for their submission.

5.1 Individual Systems

The results from the four submitted systems are summarized in Figure 1. The top two rows show the performance of all the parameter settings for all the submitted systems, while the bottom two rows show receiver operating characteristic (ROC) curves for only the best-performing parameter settings from each team. Each column in the figure denotes a different task: 'Depression versus Control' on the left, 'PTSD versus Control' in the middle and 'Depression versus PTSD' on the right. Chance performance is noted by a black dotted line in all plots, and all systems performed better than chance (with the exception of a system with deliberately random performance submitted by Duluth).

In the panels in the top two rows of Figure 1, each dot indicates a submitted parameter setting, arranged by team. From left to right, the dots represent Duluth (goldenrod), MIQ (black), UMD (red), and WWBP (blue). The best-performing system for each team is denoted by a solid horizontal line, for ease of comparison. The top row shows performance by the "official metric" of average precision, while the second row shows performance on precision at 10% false alarms.

The bottom two rows of Figure 1 show the results of each team's top-performing system (according to average-precision) across the full space of false alarms. The third row shows precision over the

[3]M-I-Q for the three authors' three institutions. Interestingly and coincidentally, 'MIQ' is also Albanian for 'Friends.'

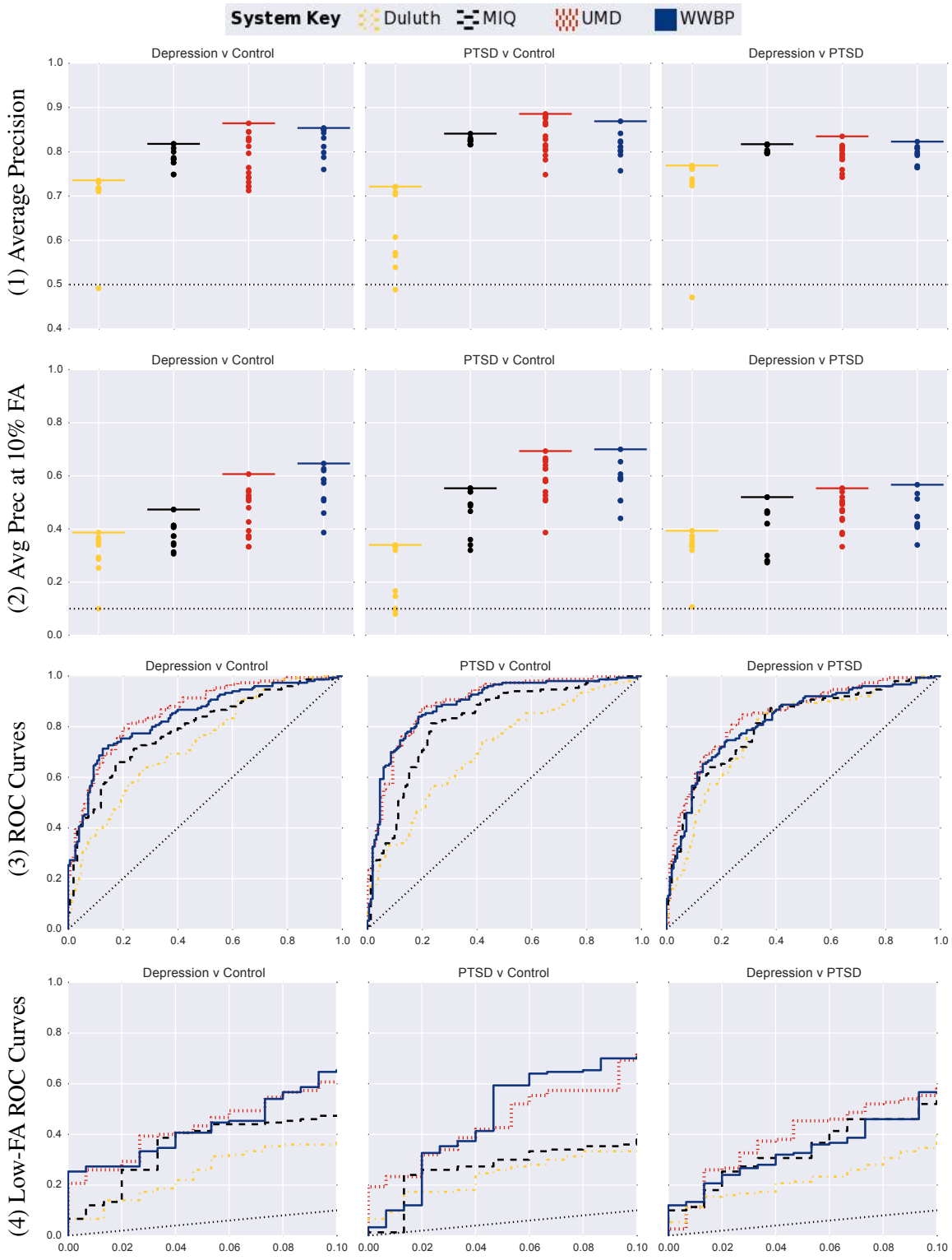

Figure 1: From top to bottom: (1) average precision and (2) precision at 10% false alarms (3) the ROC curve for each institution with the highest average precision, (4) same ROC curves, focused on the low false alarm range. For (1) and (2) the submissions are collected and colored by group. Each submitted parameter setting is represented with a single dot, with the top-scoring submission for each group in each experiment denoted with a horizontal line. The best ROC curve (according to average precision) for each institution, colored by group are shown in (3) and (4). (3) covers the range of all false alarms, while (4) is the same ROCs focused on the low false alarm range. Chance in all plots is denoted by the dotted line.

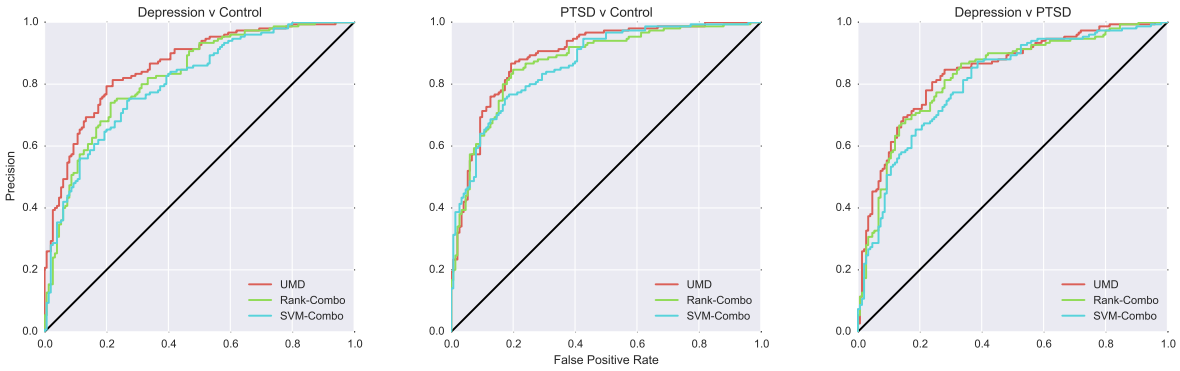

Figure 2: ROC curves for system combination results.

whole space of false alarms, while the bottom row "zooms in" to show the precision at low (0-10%) false alarm rates. These bottom two rows are shown as ROC curves, with the the false alarm rate on the x-axis and the precision on the y-axis. Performance at areas of low false alarms are particularly important to the envisioned applications, since the number of control users vastly outnumber the users with each mental health condition.

5.2 System Combination

As each of the submitted systems used what appeared to be very complementary feature sets, we performed several system combination experiments. However, as can be seen in Figure 2, system combination failed to outperform the best-performing system submitted for the shared task (UMD).

As features for system combination, we used either system ranks or scores. For each system combination experiment, we included all scores from each of the submitted systems, for a total of 47 systems (9 from Duluth, 12 from MIQ, 16 from UMD, and 10 from WWBP), without regard for how well that system performed on the classification task; future work may examine subsetting these scores for improved combination results. Since the range of the scores output by each system varied significantly, we applied a softmax normalization sigmoid function to bring all scores for each system to range from zero to one.

We explored a simple 'voting' scheme as well as a machine learning method, using Support Vector Machines (SVM). For the SVM, shown in Figure 2 as the lower blue 'SVM-Combo' curve, we experimented with using raw scores or normalized scores as features, and found the normalized scores performed much better. The SVM model is the result of training ten SVMs on system output using 10-fold cross-validation, then normalizing the SVM output prediction scores and concatenating to obtain the final result. For the voted model, which can be seen in Figure 2 as the middle green 'Rank-Combo' curve, we simply took the rank of each Twitter user according to each system output, and averaged the result. Future work will examine other methods for system combination and analysis.

6 Discussion & Conclusion

This shared task served as an opportunity for a variety of teams to come together and compare techniques and approaches for extracting linguistic signals relevant to mental health from social media data. Perhaps more importantly, though, it established a test set upon which all participating groups are now familiar, which will enable a deeper level of conversation.

Two of the classification tasks examined were previously attempted, and the techniques indicate improvement over previously-published findings. Past results did differ in a number of important factors, most notably in not examining age- and gender-matched controls, so direct comparisons are unfortunately not possible.

From these submitted systems we can take away a few lessons about classes of techniques and their relative power. There are clear benefits to using topic-

modeling approaches, as demonstrated by two of the groups (UMD and WWBP) – these provide strong signals relevant to mental health, and some intuitive and interpretable groupings of words without significant manual intervention. Simple linguistic features, even without complicated machine learning techniques, provide some classification power for these tasks (as demonstrated by Duluth and MIQ). Looking forward, there is strong evidence that techniques can provide signals at a finer-grained temporal resolution than previously explored (as demonstrated by UMD and MIQ). This may open up new avenues for applying these approaches to clinical settings.

Finally, the results leave open room for future work; none of these tasks were solved. This suggests both improvements to techniques as well as more work on dataset construction. However, even at this nascent stage, insight from the mental health signals these techniques extract from language is providing new directions for mental health research.

References

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014a. Quantifying mental health signals in Twitter. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*.

Glen Coppersmith, Craig Harman, and Mark Dredze. 2014b. Measuring post traumatic stress disorder in Twitter. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Denver, Colorado, USA, June. North American Chapter of the Association for Computational Linguistics.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Virgile Landeiro Dos Reis and Aron Culotta. 2015. Using matched samples to estimate the effects of exercise on mental health from Twitter. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Paul McNamee and James Mayfield. 2004. Character n-gram tokenization for European language text retrieval. *Information Retrieval*, 7(1-2):73–97.

Margaret Mitchell, Glen Coppersmith, and Kristy Hollingshead, editors. 2015. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. North American Association for Computational Linguistics, Denver, Colorado, USA, June.

Ted Pedersen. 2015. Screening Twitter users for depression and PTSD with lexical decision lists. 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.

Daniel Preotiuc-Pietro, Maarten Sap, H. Andrew Schwartz Schwartz, and Lyle Ungar. 2015. Mental illness detection at the World Well-Being Project for the CLPsych 2015 shared task. 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.

Nairan Ramirez-Esparza, Cindy K. Chung, Ewa Kacewicz, and James W. Pennebaker. 2008. The psychology of word use in depression forums in English and in Spanish: Testing two text analytic approaches. In *Proceedings of the 2nd International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1348–1353.

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, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 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.

J. Niels Rosenquist, James H. Fowler, and Nicholas A. Christakis. 2010. Social network determinants of depression. *Molecular psychiatry*, 16(3):273–281.

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

Maarten Sap, Greg Park, Johannes C. Eichstaedt, Margaret L. Kern, David J. Stillwell, Michal Kosinski, Lyle H. Ungar, and H. Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151.