Multitask Learning for Mental Health Conditions with Limited Social Media Data

Adrian Benton
Johns Hopkins University
adrian@cs.jhu.edu

Margaret Mitchell
Microsoft Research[1]
mmitchellai@google.com

Dirk Hovy
University of Copenhagen
mail@dirkhovy.com

Abstract

Language contains information about the author's demographic attributes as well as their mental state, and has been successfully leveraged in NLP to predict either one alone. However, demographic attributes and mental states also interact with each other, and we are the first to demonstrate how to use them jointly to improve the prediction of mental health conditions across the board. We model the different conditions as tasks in a multitask learning (MTL) framework, and establish for the first time the potential of deep learning in the prediction of mental health from online user-generated text. The framework we propose significantly improves over all baselines and single-task models for predicting mental health conditions, with particularly significant gains for conditions with limited data. In addition, our best MTL model can predict the presence of conditions (neuroatypicality) more generally, further reducing the error of the strong feed-forward baseline.

1 Introduction

Mental health conditions, like depression or anxiety, are still one of the leading causes of death worldwide. Suicide, often the direct outcome of mental health conditions, is the 11th most frequent cause of death in the US (Anderson, 2001). Detecting mental health risk factors early is key to preventing many of these deaths. Unfortunately, traditional diagnosis methods require access to and willingness to talk with a psychologist, and rely mainly on impressions formed during short sessions. Consequently, conditions leading to preventable suicides can often not be accurately diagnosed.

Automated monitoring and risk assessment of patients' language have the potential to overcome the logistic and time constraints associated with traditional assessment methods. Written text carries implicit information about the author, a relationship that has been exploited in natural language processing (NLP) to predict author characteristics, such as age (Goswami et al., 2009; Rosenthal and McKeown, 2011; Nguyen et al., 2011; Nguyen et al., 2014), gender (Sarawgi et al., 2011; Ciot et al., 2013; Liu and Ruths, 2013; Alowibdi et al., 2013; Volkova et al., 2015; Hovy, 2015), personality and stance (Schwartz et al., 2013b; Schwartz et al., 2013a; Volkova et al., 2014; Plank and Hovy, 2015; Park et al., 2015; Preoţiuc-Pietro et al., 2015), or occupation (Preotiuc-Pietro et al., 2015a; Preoţiuc-Pietro et al., 2015b). The same signal has also been effectively used to predict mental health conditions, such as depression (Coppersmith et al., 2015b; Schwartz et al., 2014), suicidal ideation (Coppersmith et al., 2016; Huang et al., 2015), schizophrenia (Mitchell et al., 2015) or post-traumatic stress disorder (PTSD) (Pedersen, 2015), often more accurately than by traditional diagnoses.

However, these studies typically model each condition in isolation and ignore other author attributes that can improve prediction, thereby artificially limiting performance. Existing research, however, indicates that 1) incorporating demographic attributes can help text classification (Volkova et al., 2013; Hovy, 2015), and 2) learning several auxiliary tasks which share common structures (e.g., part-of-speech tagging, parsing, and NER) can improve performance, as the learning implicitly exploits interactions between the tasks (Caruana, 1993; Sutton et al., 2007; Rush et al.,

[1] Now at Google Research.

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 152–162,
Valencia, Spain, April 3-7, 2017.

2010; Collobert et al., 2011; Søgaard and Goldberg, 2016).

In this paper, we propose such a multitask learning (MTL) approach to mental health prediction. The main tasks of our model are predictions of *neurotypicality* (i.e., the absence of any mental health conditions), *anxiety*, *depression*, *suicide attempt*, *eating disorder*, *panic attacks*, *schizophrenia*, *bipolar disorder*, and *post-traumatic stress disorder (PTSD)*. All of the above, plus gender prediction, also serve as auxiliary tasks.

The auxiliary tasks reflect the observation that several conditions frequently occur together (comorbidity), and that they correlate with demographic factors. The MTL framework allows us to share information across predictions. We use a neural architecture that enables the inclusion of several loss functions with a common shared underlying representation. This experimental setup is flexible enough to extend this model to further factors than the ones shown here, provided suitable data.

We also explore the effect of auxiliary-task selection on model performance for a given prediction task. Similar to Caruana (1996), we find that choosing auxiliary tasks which are prerequisites or related to the main task is critical for learning a strong model.

Our contributions

1. We are the first to apply MTL to predict mental health conditions from user content on Twitter – a notoriously difficult task (Coppersmith et al., 2015a; Coppersmith et al., 2015b).
2. We explore the influence of auxiliary-task selection on prediction performance, including the effect of gender
3. We show how to model tasks with a *large* number of positive examples to improve the prediction accuracy of tasks with a *small* number of positive examples.
4. We increase the True Positive Rate at 10% false alarms by up to 9.7% absolute (for anxiety), a result with direct impact for clinical applications.

2 Model Architecture

We opt for a neural architecture to exploit the synergies between mental conditions. Our choice is based on practical more than ideological reasons: previous work (Collobert et al., 2011; Caruana, 1996; Caruana, 1993) has indicated that this is a promising model architecture, which allows us to share parameters across tasks, can be trained on large amounts of data, and accounts for varying degrees of annotation across tasks.[1]

Even within the neural model framework, however, there are many variations to consider. In the following, we outline some attributes and decisions.

Previous approaches have shown considerable improvements over single task models by using MTL (Caruana, 1993). The arguments are convincing: predicting multiple related tasks should allow us to exploit any correlations between the predictions.

However, we note that the benefit of using a MTL model is only one possible explanation, and that another, more salient factor might have been overlooked: the difference in the general model class, i.e., neural architectures vs. discriminative or generative models, or, more generally, the *expressivity* of the model. Some comparisons might therefore have inadvertently compared apples to oranges.

We compare the multitask demographics and risk prediction with models with equal expressivity. We evaluate the performance of a standard logistic regression model (a standard approach to text-classification problems), a multilayer perceptron single-task learning (STL) model, and a neural MTL model, the latter two with equal numbers of parameters. This ensures a fair comparison by isolating the unique properties of MTL from the dimensionality-reduction aspects of deep architectures in general.

The neural models we evaluate come in two forms. The first, depicted in plate notation in Figure 1, is the STL model. These are feedforward networks with two hidden layers, trained independently to predict each task. On the right of Figure 1 is the MTL model, where the first hidden layer from the bottom is shared between all tasks. An additional per-task hidden layer is used to give the model flexibility to map from the task-agnostic representation to a task-specific one. Each hidden layer uses a rectified linear unit as non-linearity. The output layer uses a logistic non-linearity, since all tasks are binary predictions.

[1] We also experimented with a graphical model architecture, but found that it did not scale as well and provided less traction.

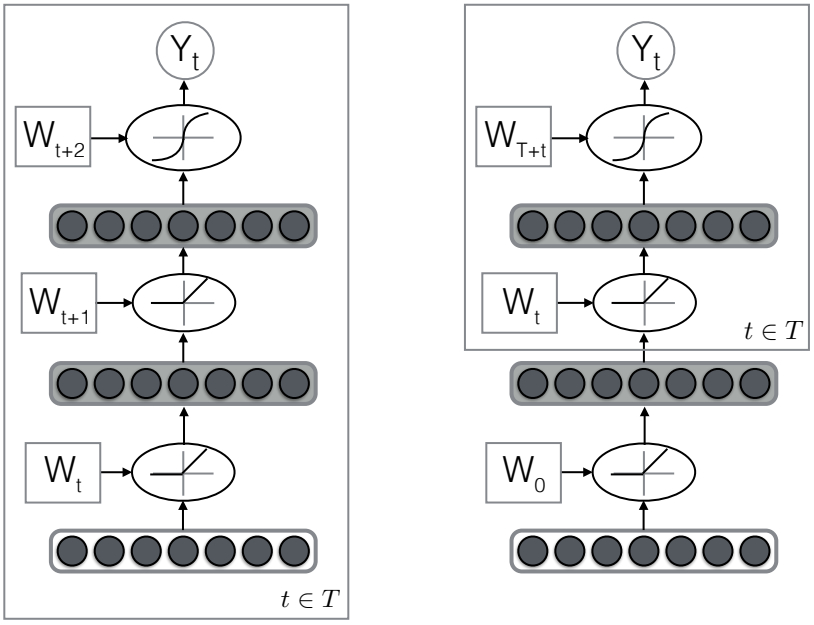

Figure 1: STL model in plate notation (left): weights trained independently for each task t (e.g., anxiety, depression) of the T tasks. MTL model (right): shared weights trained jointly for all tasks, with task-specific hidden layers.
Curves in ovals represent the type of activation used at each layer (rectified linear unit or sigmoid). Hidden layers are shaded.

The MTL model can easily be extended to a stack of shared hidden layers, allowing for a more complicated mapping from input to shared space.[2]

As noted in (Collobert et al., 2011), MTL benefits from mini-batch training, which both allows optimization to jump out of poor local optima, and take more stochastic gradient steps in a fixed amount of time (Bottou, 2012). In that paper, the authors use a randomized selection over the tasks to train. In our paper, we create mini-batches by sampling from the users in our data. Each of these users has some subset of the mental conditions we are trying to predict, and may or may not be annotated with gender. At each mini-batch gradient step we update weights for all tasks. This not only allows for randomization and faster convergence, it also provides a speed-up over the individual selection process in (Collobert et al., 2011).

One of the advantages of our setup is that we do not need complete information for every instance: instead, learning can proceed with asynchronous updates dependent on what the data in each batch has been annotated for, while sharing representations throughout. This effectively learns a joint model with a common representation for several different tasks, and allows the use of several "disjoint" data sets, some with limited annotated instances.

Optimization and Model Selection Even in a relatively simple neural model, there are a number of parameters that can (and have to) be tuned to achieve good performance. We perform a line search for every model we use, sweeping over L_2 regularization and hidden layer width. We select the best model based on the development loss. Figure 5 shows the performance on the corresponding test sets (plot smoothed by rolling mean of 10 for visibility).

In our experiments, we sweep over the L2 regularization constant applied to all weights in $\{10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.5, 1.0, 5.0, 10.0\}$, and hidden layer width (same for all layers in the network) in $\{16, 32, 64, 128, 256, 512, 1024, 2048\}$. We fix the mini-batch size to 256, and 0.05 dropout on the input layer. We found that choosing a small mini-batch size and the model with low-

[2]We tried training a 4-shared-layer MTL model to predict targets on a separate dataset, but did not see any gains over the standard 1-shared-layer MTL model in our application.

Task	# of users
Neurotypicality	4820
Anxiety	2407
Depression	1400
Suicide attempt	1208
Eating disorder	749
Schizophrenia	349
Panic disorder	263
Bipolar disorder	234
PTSD	191
Female | Male	788 | 248
total	9611

Table 1: Number of users with each self-stated condition and human-annotated gender in the joined dataset.

est development loss was sufficient to account for overfitting.

We train each model for 5,000 iterations, jointly updating all weights in our models. After this initial joint training, we select each task separately, and only update the task-specific layers of weights independently for another 1,000 iterations (selecting the set of weights achieving lowest development loss for each task individually). Weights are updated using mini-batch Adagrad (Duchi et al., 2011) – we found this to converge more quickly than other optimization schemes we experimented with. We evaluate the tuning loss every 10 epochs, and evaluate the model with the lowest tuning loss.

3 Data

We train our models on a union of multiple Twitter user datasets: 1) users identified as having anxiety, bipolar disorder, depression, panic disorder, eating disorder, PTSD, or schizophrenia (Coppersmith et al., 2015a), 2) those who had attempted suicide (Coppersmith et al., 2015c), and 3) those identified as having either depression or PTSD from the 2015 Computational Linguistics and Clinical Psychology Workshop shared task (Coppersmith et al., 2015b), along with neurotypical gender-matched controls (Twitter users not identified as having a mental condition). Users were identified as having one of these conditions if they stated explicitly they were diagnosed with this condition on Twitter (verified by a human annotator). For a subset of 1,101 users, we also have manually-annotated gender. The final dataset contains 9,611 users in total, with an average of 3521 tweets per user. The number of users with each condition is included in Table 1. Users in this joined dataset may be tagged with multiple conditions, thus the counts in this table do not sum to the total number of users.

We use the entire Twitter history of each user as input to the model, and split it into character 1-to-5-grams, which have been shown to capture more information than words for many Twitter text classification tasks (Mcnamee and Mayfield, 2004; Coppersmith et al., 2015a). We compute the relative frequency of the 5,000 most frequent n-gram features for $n \in \{1, 2, 3, 4, 5\}$ in our data, and then feed this as input to all models. This input representation is common to all models, allowing for fair comparison.

4 Experiments

Our task is to predict any number of mental conditions for each of the users in these data, possibly using gender prediction as an auxiliary task to improve our prediction performance.

We evaluate three classes of models: a baseline logistic regression over character n-gram features (LR), feed-forward multilayer perceptrons trained to predict each task separately (STL), and a multi-task network predicting a set of conditions simultaneously (MTL). We also perform ablation experiments, to see which subsets of auxiliary tasks help us learn an MTL model that predicts a particular mental condition best. For all experiments, data were divided into five equal-sized folds, three for training, one for tuning, and one for testing (we report the performance on this).

All our models are implemented in Keras[3] with Theano backend and GPU support. We train the models for a total of up to 15,000 epochs, using mini-batches of 256 instances. Training time on all five training folds ranged from one to eight hours on a machine with Tesla K40M.

Evaluation Setup We compare the accuracy of each model at predicting each task separately.

In clinical settings, we are interested in minimizing the number of false positives, i.e., incorrect diagnoses, which can cause undue stress to the patient. We are thus interested in bounding this quantity. To evaluate the performance, we plot the false positive rate (FPR) against the true positive rate (TPR). This gives us a receiver operating characteristics (ROC) curve, allowing us to inspect the

[3] http://keras.io/

performance of each model on a specific task at any level of FPR.

While the ROC gives us a sense of how well a model performs at a fixed true positive rate, it makes it difficult to compare the individual tasks at a low false positive rate, which is also important for clinical application. We therefore report two more measures: the area under the ROC curve (AUC) and TPR performance at FPR=0.1 (TPR@FPR=0.1). We do not compare our models to a majority baseline model, since this model would achieve an expected AUC of 0.5 for all tasks, and F-score and TPR@FPR=0.1 of 0 for all mental conditions – users exhibiting a condition are the minority, meaning a majority baseline classifier would achieve zero recall.

5 Results

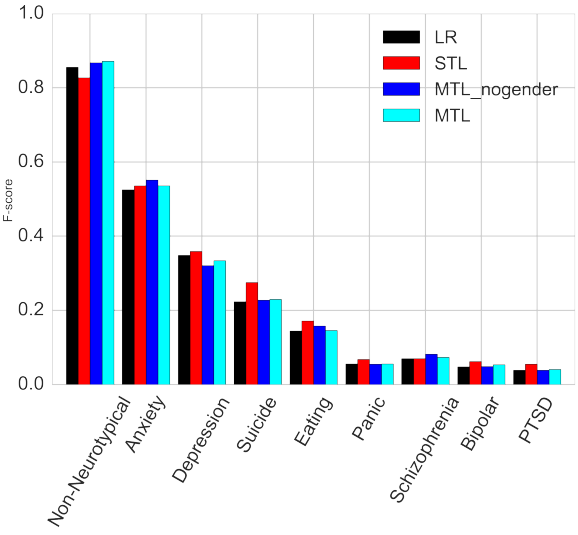

Figure 2: F1-score for predicting each condition.

Figure 2 shows the F1-score of each model at predicting each task separately, Figure 3 shows the AUC-score, and Figure 4. Precision-recall curves for each of model/task are in Figure 6.

STL corresponds to a multilayer perceptron with two hidden layers (with a similar number of parameters as the proposed MTL model). The MTL_nogender and MTL models predict all tasks simultaneously, but are only evaluated on the main respective task.

MTL often outperforms the LR baseline in terms of AUC and TPR@F=0.1, but the difference is less clear when comparing F1-scores.

In terms of AUC and TPR@F=0.1, STL models do not perform nearly as well as MTLor LR. This is likely because the neural networks learned by STL cannot be guided by the inductive bias provided by MTL training. Note, however, that STL and MTL are oftentimes comparable in terms of F1-score.

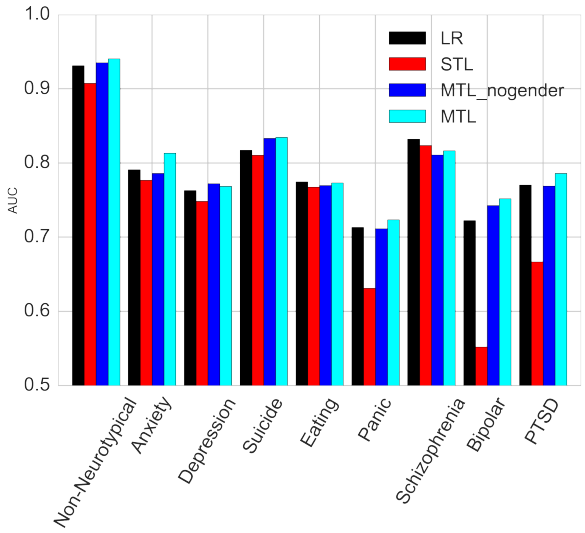

Figure 3: AUC for predicting different tasks

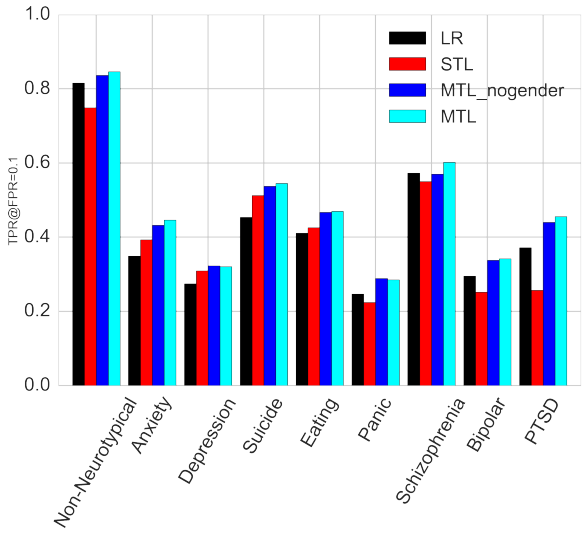

Figure 4: TPR at 0.10 FPR for predicting different tasks

MTL Leveraging Comorbid Conditions Improves Prediction Accuracy We find that the prediction of the conditions with the least amount of data – *bipolar disorder* and *PTSD* – are significantly improved by forcing the model to also predict comorbid conditions which have substantially more data: *depression* and *anxiety*. We are able to increase the AUC for predicting PTSD to 0.786 by MTL, from 0.770 by LR, whereas STL fails to perform as well with an AUC of

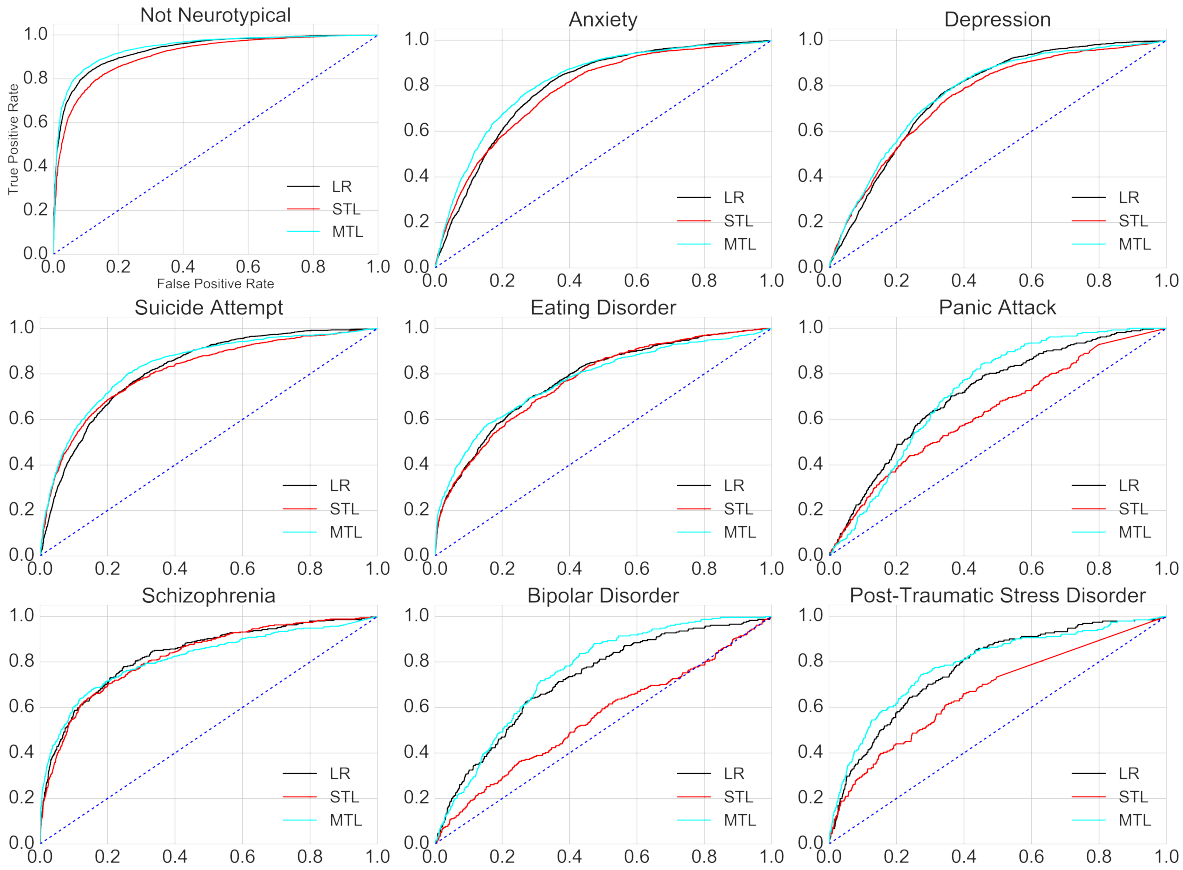

Figure 5: ROC curves for predicting each condition. The precision (diagnosed, correctly labeled) is on the y-axis, while the proportion of false alarms (control users mislabeled as diagnosed) is on the x-axis. Chance performance is indicated by the dotted diagonal line.

0.667. Similarly for predicting bipolar disorder (MTL:0.723, LR:0.752, STL:0.552) and panic attack (MTL:0.724, LR:0.713, STL:0.631).

These differences in AUC are significant at $p = 0.05$ according to bootstrap sampling tests with 5000 samples. The wide difference between MTL and STL can be explained in part by the increased feature set size – MTL training may, in this case, provide a form of regularization that STL cannot exploit. Further, modeling the common mental health conditions with the most data (depression, anxiety) helps in pulling out more rare conditions that also manifest in these common health conditions.

This is clear evidence that an MTL model provides strong gains for predicting elusive conditions by using large data for common conditions, and only a small amount of data for the related, small-data conditions.

Utility of Authorship Attributes Figures 3 and 4 both suggest that adding gender as an auxiliary task leads to more predictive models, even though the difference is not statistically significant for most tasks. This is in line with the suggestions in Volkova et al. (2013), Hovy (2015). Interestingly, though, the MTL model is worse at predicting gender itself. While this could be a direct result of data sparsity (recall that we have only a small subset annotated for gender), which could be remedied by annotating additional users for gender, this appears unlikely given the other findings of our experiments, where MTL helped in specifically these sparse scenarios.

However, it has been pointed out by Caruana (1996) that not all tasks benefit from a MTL setting in the same way, and that some tasks serve purely auxiliary functions. Here, gender prediction does not benefit from including mental conditions, but helps vice versa. In other words, predicting gender is qualitatively different from predicting mental health conditions: it seems likely that the signals for anxiety ares much more similar to the ones for depression than for, say, being male, and can therefore add to detecting de-

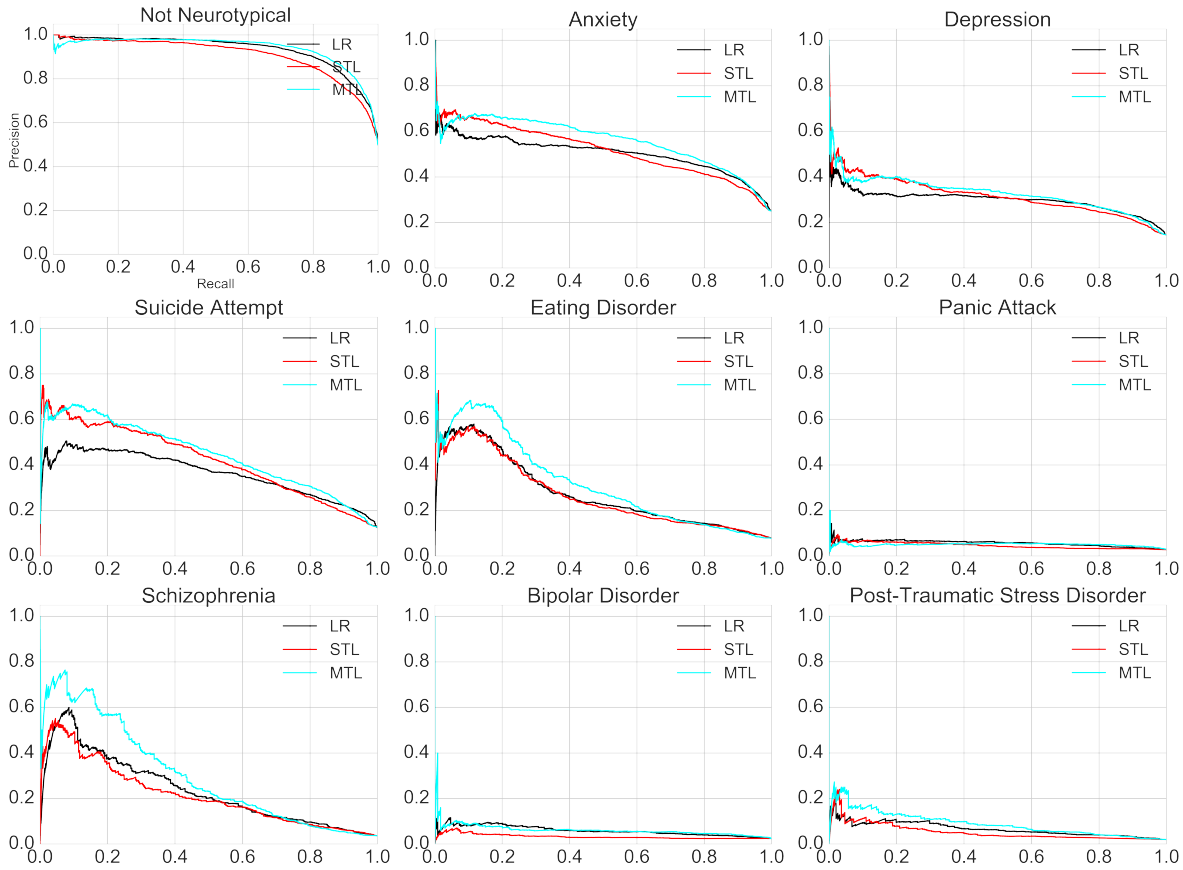

Figure 6: Precision-recall curves for predicting each condition.

pression. However, the distinction between certain conditions does not add information for the distinction of gender. The effect may also be due to the fact that these data were constructed with inferred gender (used to match controls), so there might be a degree of noise in the data.

Choosing Auxiliary Tasks Although MTL tends to dominate STL in our experiments, it is not clear whether auxiliary tasks just introduce a beneficial bias in MTL models in general, or if there exists a specific subset of auxiliary tasks for predicting each condition. We perform ablation experiments by training MTL models on only a subset of the tasks, and evaluate them at predicting a single target. We focus on four conditions we want to predict well: anxiety, depression, suicide attempts, and bipolar disorder. For each task, we vary the auxiliary tasks we train the MTL model with, and evaluate how well it predicts the main task. Since considering all possible subsets of tasks as auxiliary tasks is combinatorily unfeasible, we choose the following task subsets as auxiliary:

- *all*: all mental conditions along with gender
- *all conds*: only all mental conditions (gender omitted)
- *neuro*: only neurotypicality
- *neuro+mood*: neurotypicality, depression, and bipolar disorder (mood disorders)
- *neuro+anx*: neurotypicality, anxiety, and panic attack (anxiety conditions)
- *neuro+targets*: neurotypicality, anxiety, depression, suicide attempt, and bipolar disorder
- *none*: no auxiliary tasks, equivalent to STL model

Table 2 shows AUC for the four prediction tasks with different subsets of auxiliary tasks. Statistically significant improvements over the respective LR baselines are denoted by superscript. Restricting the auxiliary tasks to a small subset tends to hurt performance for most tasks. This suggests that the biases induced by predicting any mental condition are all mutually beneficial – e.g., models that predict depression, are also useful at predicting anxiety.

It is thus best not to think of MTL as one single "black box" model that can predict all mental con-

Auxiliary Tasks	Main Task			
	anxiety	bipolar	depression	suicide attempt
all	$0.813^{*\dagger}$	$0.752^{*\dagger}$	$0.769^{\dagger}$	$0.835^{*\dagger}$
all conds	0.786	$0.743^{\dagger}$	$0.772^{\dagger}$	$0.833^{*\dagger}$
neuro	0.763	$0.740^{\dagger}$	0.759	0.797
neuro+mood	0.756	$0.742^{\dagger}$	0.761	0.804
neuro+anx	0.770	$0.744^{\dagger}$	0.746	0.792
neuro+targets	0.750	$0.747^{\dagger}$	0.764	0.817
none (STL)	0.777	0.552	0.749	0.810
LR	0.791	$0.723^{\dagger}$	0.763	0.817

Table 2: Test AUC when predicting *Main Task* after training to predict a subset of auxiliary tasks. Significant improvement over LR baseline at $p = 0.05$ is denoted by *, and over no auxiliary tasks (STL) by †.

ditions at the same time, but a framework to exploit auxiliary tasks as regularization to effectively combat data paucity and less-than-trustworthy labels.

6 Discussion

Our results indicate that the proposed MTL setting results in significant gains for the prediction of mental health conditions with limited data, benefiting from predicting related mental conditions and demographic attributes simultaneously.

We experimented with all the optimizers that Keras provides, and found that Adagrad seems to converge fastest to a good optimum, although all the adaptive learning rate optimizers (such as Adam, etc.) tend to converge quickly. This indicates that the gradient is significantly steeper along certain parameters than others. Default stochastic gradient descent (SGD) was not able to converge as quickly, since it is not able to adaptively scale the learning rate for each parameter in the model – taking too small steps in directions where the gradient is shallow, and too large steps where the gradient is steep. We further note an interesting behavior: all of the adaptive learning rate optimizers yield a strange "step-wise" training loss learning curve, which hits a plateau, but then drops after about 900 iterations, only to hit another plateau, and so on. Obviously, we would prefer to have a smooth training loss curve. We can indeed achieve this using SGD, but it takes much longer to converge than, for example, Adagrad. This suggests that a well-tuned SGD would be the best optimizer for this problem, a step that would require some more experimentation and is left for future work.

Learning Rate	Loss	L2	Loss	Hidden Width	Loss
10^{-4}	5.1	10^{-3}	2.8	32	3.0
$5 * 10^{-4}$	2.9	$5 * 10^{-3}$	2.8	64	3.0
10^{-3}	2.9	10^{-2}	2.9	128	2.9
$5 * 10^{-3}$	2.4	$5 * 10^{-2}$	3.1	256	2.9
10^{-2}	2.3	0.1	3.4	512	3.0
$5 * 10^{-2}$	2.2	0.5	4.6	1024	3.0
0.1	20.2	1.0	4.9		

Table 3: Average dev loss over epochs 990-1000 of joint training on all tasks as a function of different learning parameters. Optimized using Adagrad with hidden layer width 256.

We also found that feature counts have a pronounced effect on the loss curves: relative feature frequencies yield models that are much easier to train than raw feature counts.

As indicated by the effect of raw vs. relative counts, feature representations are another area of optimization, such as different ranges of character n-grams (e.g., $n > 5$) and word unigrams. We decided on character 1-to-5-grams, since we believe that these features generalize better to a new domain (e.g., Facebook) than word unigrams. However, there is no fundamental reason not to choose longer character n-grams, other than time constraints in regenerating the data, and sufficiently accounting for overfitting with proper regularization.

Initialization is often listed as a decisive factor in neural models, and Goldberg (2015) recommends repeated restarts with differing initializations to find the optimal model. In an earlier experiment, we tried initializing a MTL model (albeit without task-specific hidden layers) with pretrained word2vec embeddings of unigrams trained on the Google News n-gram corpus. However, we did not notice an improvement in F-score. This could be due to the other factors, though, such as feature sparsity.

Table 3 displays sweeps over learning parameters with hidden layer width 256, training the MTL model to predict multiple mental conditions jointly for the Qntfy self-stated data (character trigrams as input features). The sweet spots in this table are probably good starting points for training models.

7 Related Work

MTL was introduced by Caruana (1993), based on the observation that humans rarely learn things in isolation, and that it is the similarity between related tasks that helps us get better.

Some of the first works on MTL have been motivated by medical risk prediction (Caruana et al., 1996), and it is now being rediscovered for this purpose (Lipton et al., 2016). The latter use a long short-term memory (LSTM) structure to provide several medical diagnoses from health care features (yet no textual or demographic information), and find small, but probably not significant improvements over a structure similar to the STL we use here.

However, in both cases, the target was medical conditions as detected in patient records, not mental health conditions in online data. The main focus in this work has been on the correlation between individual conditions and linguistic markers, to establish the possibility of detecting risk in written data. While some of the approaches have looked at more than one condition, none of them have done so in an MTL framework, foregoing the possibility of modeling comorbidity and correlation with demographic factors.

The framework proposed by Collobert et al. (2011) allows for predicting any number of NLP tasks from a convolutional neural network (CNN) representation of the input text. The model we present is much simpler, just a feed-forward network with n-gram input layer. Our contribution is to show that constraining n-gram embeddings to be predictive of various mental health condition also helps. We chose to experiment with a feed-forward network against independent logistic regression models since this was the simplest way to test our hypothesis. Comparing more complicated models is possible, but distracts from the question whether or not MTL training with extra-linguistic targets helps us.

8 Conclusion and Future Work

In this paper, we develop neural MTL models for 10 prediction tasks (eight mental health conditions, neurotypicality, and gender). We compare their performance with STL models trained to predict each task independently.

Our results show that the most complex MTL model performs significantly better than independent LR models, reaching 0.846 TPR where FPR=0.1 and reducing the error rate in identifying anxiety by up to 11.9%. We also investigate the influence of the depth of the model, by comparing to progressively deeper STL feed-forward networks with the same number of parameters. We find: (1) Most of the modeling power stems from the expressiveness conveyed by deep architectures. (2) Choosing the correct set of auxiliary tasks for a given mental condition can yield a significantly more predictive model. (3) The MTL model dramatically improves for conditions with the smallest amount of data. (4) Gender prediction does not follow the two previous points, but improves performance when added as an auxiliary task.

Accuracy of the MTL approach is not yet ready to be used in isolation in the clinical setting. However, our experiments suggest this is a promising direction moving forward. There are strong gains to be made in using multitask learning to aid clinicians in their evaluations, and with further partnerships between the clinical and machine learning community, we foresee improved suicide prevention efforts.

In the future, we plan to explore the possibility of hierarchical models, encoding the fact that certain tasks inform others more than vice versa.

Acknowledgements

The authors would like to thank Kristy Hollingshead Seitz, Glen Coppersmith, and H. Andrew Schwartz, as well as the organizers and funders of the Johns Hopkins Jelinek Summer School 2016, where large parts of this work were conducted. We are also grateful for the invaluable feedback on MTL from Yoav Goldberg, Stephan Gouws, Ed Greffenstette, Karl Moritz Hermann, and Anders Søgaard.

References

Jalal S Alowibdi, Ugo A Buy, and Philip Yu. 2013. Empirical evaluation of profile characteristics for gender classification on twitter. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, volume 1, pages 365–369. IEEE.

Robert N. Anderson. 2001. *Deaths: leading causes for 1999*. Centers for Disease Control and Prevention, National Center for Health Statistics.

Léon Bottou. 2012. Stochastic gradient tricks. *Neural Networks, Tricks of the Trade, Reloaded*, pages 430–445.

Rich Caruana, Shumeet Baluja, Tom Mitchell, et al. 1996. Using the future to" sort out" the present: Rankprop and multitask learning for medical risk evaluation. *Advances in neural information processing systems*, pages 959–965.

Rich Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48. Citeseer.

Rich Caruana. 1996. Algorithms and applications for multitask learning. In *ICML*, pages 87–95. Citeseer.

Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender Inference of Twitter Users in Non-English Contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Wash*, pages 18–21.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015a. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015b. Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality. pages 31–39. Association for Computational Linguistics.

Glen Coppersmith, Ryan Leary, Eric Whyne, and Tony Wood. 2015c. Quantifying suicidal ideation via language usage on social media. In *Joint Statistics Meetings Proceedings, Statistical Computing Section, JSM*.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Proceedings of the third workshop on computational lingusitics and clinical psychology. pages 106–117. Association for Computational Linguistics.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Yoav Goldberg. 2015. A primer on neural network models for natural language processing. *arXiv preprint arXiv:1510.00726*.

Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric analysis of bloggers' age and gender. In *Third International AAAI Conference on Weblogs and Social Media*.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of ACL*, pages 752–762.

Xiaolei Huang, Xin Li, Tianli Liu, David Chiu, Tingshao Zhu, and Lei Zhang. 2015. Topic model for identifying suicidal ideation in chinese microblog. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 553–562.

Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzell. 2016. Learning to diagnose with lstm recurrent neural networks. In *Proceedings of ICLR*.

Wendy Liu and Derek Ruths. 2013. What's in a name? Using first names as features for gender inference in Twitter. In *Analyzing Microtext: 2013 AAAI Spring Symposium*.

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

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 11–20, Denver, Colorado, June 5. Association for Computational Linguistics.

Dong Nguyen, Noah A Smith, and Carolyn P Rosé. 2011. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 115–123. Association for Computational Linguistics.

Dong Nguyen, Dolf Trieschnigg, A. Seza Dogruöz, Rilana Gravel, Mariet Theune, Theo Meder, and Franciska De Jong. 2014. Predicting Author Gender and Age from Tweets: Sociolinguistic Theories and Crowd Wisdom. In *Proceedings of COLING 2014*.

Greg Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, David J Stillwell, Michal Kosinski, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*.

Ted Pedersen. 2015. Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality. pages 46–53. Association for Computational Linguistics.

Barbara Plank and Dirk Hovy. 2015. Personality traits on twitter—or—how to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98.

Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, Hansen Andrew Schwartz, and Lyle H Ungar. 2015. The role of personality, age and gender in tweeting about mental illnesses. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, NAACL.

Daniel Preotiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015a. An analysis of the user occupational class through twitter content. In *ACL*.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015b. Studying user income through language, behaviour and affect in social media. *PloS one*, 10(9):e0138717.

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 763–772. Association for Computational Linguistics.

Alexander M Rush, David Sontag, Michael Collins, and Tommi Jaakkola. 2010. On dual decomposition and linear programming relaxations for natural language processing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1–11. Association for Computational Linguistics.

Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 78–86. Association for Computational Linguistics.

Hansen Andrew Schwartz, Johannes C Eichstaedt, Lukasz Dziurzynski, Margaret L Kern, Eduardo Blanco, Michal Kosinski, David Stillwell, Martin EP Seligman, and Lyle H Ungar. 2013a. Toward personality insights from language exploration in social media. In *AAAI Spring Symposium: Analyzing Microtext*.

Hansen Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013b. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9).

Andrew H. Schwartz, Johannes Eichstaedt, L. Margaret Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality. pages 118–125. Association for Computational Linguistics.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *The 54th Annual Meeting of the Association for Computational Linguistics*, page 231.

Charles Sutton, Andrew McCallum, and Khashayar Rohanimanesh. 2007. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *Journal of Machine Learning Research*, 8(Mar):693–723.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of EMNLP*, pages 1815–1827.

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring user political preferences from streaming communications. In *Proceedings of the 52nd annual meeting of the ACL*, pages 186–196.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring latent user properties from texts published in social media (demo). In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI)*, Austin, TX, January.