Depression and Self-Harm Risk Assessment in Online Forums

Andrew Yates[†*] **Arman Cohan**[‡*] **Nazli Goharian**[‡]

[†]Max Planck Institute for Informatics,
Saarland Informatics Campus Saarbruecken, Germany

[‡]Information Retrieval Lab, Department of Computer Science,
Georgetown University, Washington DC, USA

ayates@mpi-inf.mpg.de
{arman,nazli}@ir.cs.georgetown.edu

Abstract

Users suffering from mental health conditions often turn to online resources for support, including specialized online support communities or general communities such as Twitter and Reddit. In this work, we present a framework for supporting and studying users in both types of communities. We propose methods for identifying posts in support communities that may indicate a risk of self-harm, and demonstrate that our approach outperforms strong previously proposed methods for identifying such posts. Self-harm is closely related to depression, which makes identifying depressed users on general forums a crucial related task. We introduce a large-scale general forum dataset consisting of users with self-reported depression diagnoses matched with control users. We show how our method can be applied to effectively identify depressed users from their use of language alone. We demonstrate that our method outperforms strong baselines on this general forum dataset.

1 Introduction

Mental health remains a major challenge in public health care. Depression is one of the most common mental disorders and 350 million people are estimated to suffer from depression worldwide (WHO, 2010). In 2014 an estimated 7% of all U.S. adults had experienced at least one major depressive disorder (2015). Suicide and self-harm are major related concerns in public mental health. Suicide is one of the leading causes of death (CDC, 2015), and each suicide case has major consequences on the physical and emotional well-being of families and on societies in general. Therefore identifying individuals at risk of self-harm and providing support to prevent it remains an important problem (Ferrari et al., 2014).

Social media is often used by people with mental health problems to express their mental issues and seek support. This makes social media a significant resource for studying language related to depression, suicide, and self-harm, as well as understanding the authors' reasons for making such posts, and identifying individuals at risk of harm (Coppersmith et al., 2014a). Depression and suicide are closely related given that depression is the psychiatric diagnosis most commonly associated with suicide. Research has demonstrated that forums are powerful platforms for self-disclosure and social support seeking around mental health concerns (De Choudhury and De, 2014; Manikonda and De Choudhury, 2017). Such support forums are often staffed by moderators who are mental health experts, trained volunteers, or more experienced users whose role is to identify forum posts suggesting that a user is at risk of self-harm and to provide support.

Studies have shown that self expression and social support are beneficial in improving the individual's state of the mind (Turner et al., 1983; Choudhury and Kiciman, 2017) and thus such communities and interventions are important in suicide prevention. However, there are often thousands of user posts published in such support forums daily, making it difficult to manually identify individuals at risk of self-harm. Additionally, users in acute distress need prompt attention, and any delay in responding to these users could have adverse consequences. Therefore, identifying individuals at risk of self-harm in such support forums is an important challenge. Identifying signs of depression in general social media, on the other hand, is also a difficult task that has appli-

[*] The first two authors contributed equally to this work.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2968–2978
Copenhagen, Denmark, September 7–11, 2017. ©2017 Association for Computational Linguistics

cations for both better understanding the relationship between mental health and language use and for monitoring a specific user's state (e.g., in the context of monitoring a user's response to clinical care). In this work we propose and evaluate a framework for performing self-harm risk assessment and for identifying depression in online forums.

We present a general neural network architecture for combining posts into a representation of a user's activity that is used to classify the user. To address the challenge of depression risk assessment over the general forums, we introduce a large-scale novel Reddit dataset that is substantially larger than the existing data and has a much more realistic number of control users. The dataset contains over 9,000 users with self-reported depression diagnoses matched with over 107,000 control users. We apply our approach to *(1)* identify the users with depression on a general forum like Reddit, and to *(2)* estimate the risk of self-harm indicated by posts in a more specific mental-health support forum. Our methods perform significantly better on both datasets than strong existing methods, demonstrating that our approach can be used both to identify depressed users and to estimate the risk of self-harm posed by individual posts.

2 Related Work

There is a growing body of related work analyzing mental health-related discourse and language usage in social media to better discover and understand mental health related concerns (Resnik et al., 2013; De Choudhury et al., 2013; Coppersmith et al., 2014b,a; Mitchell et al., 2015; Tsugawa et al., 2015; Coppersmith et al., 2015a; Althoff et al., 2016; Mowery et al., 2016; Benton et al., 2017b). To investigate NLP methods for identifying depression and PTSD users on Twitter, a shared task (Coppersmith et al., 2015b) at the 2nd Computational Linguistics and Clinical Psychology Workshop (CLPsych 2015) was introduced where the participants evaluated their methods on a dataset of about 1800 Twitter users. Other work has used data from approximately 900 Reddit.com users to support self-reported diagnosis detection (Losada and Crestani, 2016). Previous work identifying depression and other mental health problems, including the methods participating in CLPsych 2015 (e.g. (Resnik et al., 2015; Preoţiuc-Pietro et al., 2015)) heavily rely on utilizing features such as LIWC (Pennebaker et al., 2015), topic modeling, manual lexicons, or other domain-dependent application-specific features. Aside from the effort required to design effective features, these approaches usually model the problem with respect to the selected features and ignore other indicators and signals that can improve prediction. In contrast, our model only relies on text and is not dependent on any external or domain-specific features. Previous self-reported diagnosis detection datasets contained a limited number of both control users and diagnosed users. In contrast to this, we construct a new dataset with over 9,000 depressed users matched with a realistic number of control users.

In addition to general studies addressing mental health, related work has also specifically studied suicide and self-harm through social media (Jashinsky et al., 2014; Thompson et al., 2014; Gunn and Lester, 2015; De Choudhury et al., 2016; Coppersmith et al., 2016). Recently, CLPsych 2016 (Hollingshead and Ungar, 2016) investigated approaches for detecting the self-harm risk of mental health forum posts (Milne et al., 2016). Most related work in this area uses variations of linear classifiers with some sort of feature engineering; successful methods have employed: a combination of sparse (bag-of-words) and dense (doc2vec) representation of the target forum posts (Kim et al., 2016), a stack of feature-rich Random Forest and linear Support Vector Machine (SVM) (Malmasi et al., 2016), an RBF SVM classifier utilizing similar sets of features (Brew, 2016), and various contextual and psycholinguistic features (Cohan et al., 2016, 2017). In contrast to the above works, our model does not use any general or domain specific feature engineering; it learns appropriate representations of documents by considering only their textual content.

Our proposed models consist of a shared architecture based on a CNN, a merge layer, model-specific loss functions, and an output layer (as we will describe in §4). While our model shares similarities with CNN-based models in prior work (Kalchbrenner et al., 2014; Kim, 2014; Xiao and Cho, 2016), it focuses on learning representations of user's posts and combining the post representations into an overall representation of the user's activity. In the case of self-harm risk assessment, we experiment with several loss functions to determine whether considering the ordinal nature of self-harm risk labels (i.e., green, amber, red, and crisis) can improve performance. Evaluation results suggest that the model variant using this loss function is more robust than our other variants.

3 Data

3.1 Depression dataset construction.

We created a new dataset to support the task of identifying forum users with self-reported depression diagnoses. The Reddit Self-reported Depression Diagnosis (RSDD) dataset was created by annotating users from a publicly-available Reddit dataset[1]. Users to annotate were selected by identifying all users who made a post between January 2006 and October 2016 matching a high-precision diagnosis pattern.[2] Users with fewer than 100 posts made before their diagnosis post were discarded. Each of the remaining diagnosis posts was then viewed by three layperson annotators to decide whether the user was claiming to have been diagnosed with depression; the most common false positives included hypotheticals (e.g., "if I was diagnosed with depression"), negations (e.g., "it's not like I've been diagnosed with depression"), and quotes (e.g., "my brother announced 'I was just diagnosed with depression' "). Only users with at least two positive annotations were included in the final group of diagnosed users.

A pool of potential control users was identified by selecting only those users who had *(1)* never posted in a subreddit related to mental health, and *(2)* never used a term related to depression or mental health. These restrictions minimize the likelihood that users with depression are included in the control group. In order to prevent the diagnosed users from being easily identified by the usage of specific keywords that are never used by the control users, we removed all posts by diagnosed users that met either one of the aforementioned conditions (i.e., that was posted in a mental health subreddit or included a depression term).

For each diagnosed user and potential control user, we calculated the probability that the user would post in each subreddit (while ignoring diagnosed users' posts made to mental health subreddits). Each diagnosed user was then greedily matched with the 12 control users who had the smallest Hellinger distance between the diagnosed user's and the control user's subreddit post probability distributions, excluding control users with 10% more or fewer posts than the diagnosed user. This matching approach ensures that diagnosed users are matched with control users who are interested in similar subreddits and have similar activity levels, preventing biases based on the subreddits users are involved in or based on how active the users are on Reddit. This yielded a dataset containing 9,210 diagnosed users and 107,274 control users. On average each user in the dataset has 969 posts (median 646). The mean post length is 148 tokens (median 74).

The Reddit Self-reported Depression Diagnosis (RSDD) dataset differs from prior work creating self-reported diagnoses datasets in several ways: it is an order of magnitude larger, posts were annotated to confirm that they contained claims of a diagnosis, and a realistic number of control users were matched with each diagnosed user. The lists of terms related to mental health, subreddits related to mental health, high-precision depression diagnosis patterns, and further information are available[3]. We note that this dataset has some (inevitable) caveats: *(i)* the method only captures a subpopulation of depressed people (i.e. those with self-reported diagnosis), *(ii)* Reddit users may not be a representative sample of the population as a whole, and *(iii)* there is no way to verify whether the users with self-reported diagnoses are truthful.

3.2 Self-harm assessment.

For self-harm risk assessment we use data from mental health forum posts from ReachOut.com, which is a successful Australian support forum for young people. In addition to providing peer-support, ReachOut moderators and trained volunteers monitor and participate in the forum discussions. The NAACL 2016 Computational Linguistics and Clinical Psychology Workshop (Hollingshead and Ungar, 2016) released a Triage dataset containing 65,024 forum posts from ReachOut, with annotations for 1,227 posts indicating the author's risk of self-harm (Milne et al., 2016). The annotations consist of one of four labels: green (indicating no action is required from ReachOut's moderators), amber (non-urgent attention is required), red (urgent attention is required), and crisis (a risk that requires immediate attention).

[1] https://files.pushshift.io/reddit/

[2] e.g., "I was just diagnosed with depression."

[3] http://ir.cs.georgetown.edu/data/reddit_depression/

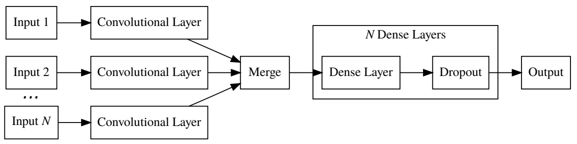

Figure 1: The general neural network architecture shared among our user and post classification models. Each input (e.g., each of a user's posts) is processed by a convolutional network and merged to create a vector representation of the user's activity. This vector representation is passed through one or more dense layers followed by an output layer that performs classification. The type of input received, merge operation, and output layer vary with the specific model.

3.3 Ethical concerns.

Social media data are often sensitive, and even more so when the data are related to mental health. Privacy concerns and the risk to the individuals in the data should always be considered (Hovy and Spruit, 2016; Šuster et al., 2017; Benton et al., 2017a). We note that the risks associated with the data used in this work are minimal. This assessment is supported by previous work on the ReachOut dataset (Milne et al., 2016), on Twitter data (Coppersmith et al., 2015b), and on other Reddit data (Losada and Crestani, 2016). The RSDD dataset contains only publicly available Reddit posts. Annotators were shown only anonymized posts and agreed to make no attempts to deanonymize or contact them. The RSDD dataset will only be made available to researchers who agree to follow ethical guidelines, which include requirements not to contact or attempt to deanonymize any of the users. Additionally, for the ReachOut forum data that was explicitly related to mental health, the forum's rules require the users to stay anonymous; moderators actively redact any user identifying information.

4 Methodology

We describe a general neural network architecture for performing text classification over multiple input texts. We propose models based on this architecture for performing two tasks in the social media and mental health domains that we call *self-harm risk classification* and *detecting depression*. The task of self-harm risk classification is estimating a user's current self-harm risk given the user's post on a mental health support forum and the previous posts in the thread. The task of detecting depressions in users is identifying Reddit users with self-reported depression diagnoses given the users' post histories (excluding posts containing mental health keywords or posted in subreddits related to mental health).

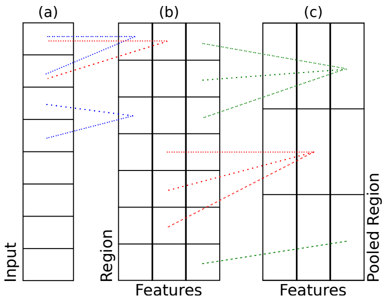

Figure 2: The convolutional network component of our architecture. A convolutional layer takes a series of terms as input *(a)* and applies l filters to a k-term sliding window to derive feature values for each window or region *(b)*; $k = 2$ and $l = 3$ shown here. A max pooling layer considers non-overlapping region sequences of length n *(b)* and keeps the highest feature value for the sequence *(c)*; $n = 3$ shown here.

While both tasks are focused on predicting a user's mental health status, they differ in both the type of classification performed (i.e., estimating severity on a four point scale vs. boolean classification) and in the amount of data available. Our general architecture is based on a two step process: *(1)* identifying relevant features in each input text, and *(2)* combining the features observed in the model's inputs to classify the user.

4.1 Shared Architecture

Our proposed models share a common architecture that takes one or more posts as input, processes the posts using a convolutional layer to identify features present in sliding windows of text, merges the features identified into a vector representation of the user's activity, and uses a series of dense layers to perform classification on the merged vector representation. The type of merging performed and the output layers are properties of the model variant, which we describe in detail in the following section. Convolutional networks have commonly been applied to the task of text classification, such as by Kim (2014). We use categorical cross-entropy as a loss function with both methods, but also experiment with other loss functions when performing severity classification.

First, the model takes one or more posts as input and processes each post with a convolutional network containing a convolutional layer and a pooling layer. This process is illustrated with a max pooling layer in Figure 2. The convolutional layer applies filters to a sliding window of k terms *(a)* and outputs a feature value for each sliding window region and each filter *(b)*. The same filters are applied to each window; each filter can be viewed as a feature detector and the overall process can be conceptualized as looking for windows of terms that contain specific features. The features are not specified a priori through feature engineering, but instead are learned automatically when the model is trained. After identifying the features present in each region (i.e., sliding window), a max pooling layer considers non-overlapping regions of length n and keeps the highest feature value for each region *(c)*. This step eliminates the regions (i.e., sliding windows) that do not contain useful features, which reduces the size of the convolutional network's output. The same convolutional network is applied to each input post, meaning that the model learns to look for the same set of features in each.

After each input post has been processed by a convolutional network, the output of each convolutional network is merged to create a representation of the user's activity across all input posts. This representation is processed by one or more dense layers (i.e., fully connected layers) with dropout (Srivastava et al., 2014) before being processed by a final output layer to perform classification. The type of output layer is dependent on the model variant. Our shared model architecture is illustrated in Figure 1. The architecture's hyperparameters (e.g., the sliding window size k, the number of convolutional filters used, and type of pooling) also vary among models and are described in the supplemental material. Both the convolutional and dense layers use ReLU activations (Nair and Hinton, 2010) in all model variants.

4.2 Models

4.2.1 Depression detection

Our model for depression detection takes a user's posts as input and processes each post with a convolutional network. Each convolutional network performs average pooling to produce its output. That is, the model considers non-overlapping sequences of n posts and keeps the average feature value across all sequences. These post representations are then merged with a second convolutional layer to create a user representation; we found this approach led to more stable performance than using a second average pooling or max pooling layer. The user representation created by the merge step is then passed to one or more dense layers before being passed to a dense output layer with a softmax activation function to perform classification. The number of dense layers used is a hyperparameter described in §5. Categorical cross-entropy is used as the model's loss function.

4.2.2 Self-harm risk assessment

Our model for self-harm risk classification takes two inputs: the target post being classified and the prior posts (if any) in the target post's thread. The prior posts provide context and are thus useful for estimating the risk of self-harm present in the target post. The two inputs are both processed by a convolutional network as in user-level classification, but in this case the convolutional network's outputs correspond to a representation of the target post and to a representation of the target post's context (i.e., the prior posts in the thread). Given that these two outputs represent different aspects, they are merged by concatenating them together. This merged representation is then passed to one or more dense layers and to an output layer; the type of output layer depends on the loss function used. There are four self-harm risk assessment model variants in total:

Categorical Cross Ent. uses an output layer with a softmax activation function, and categorical cross-entropy as its loss function. This mirrors the output layer and loss function used in the user level classification model.

MSE uses an output layer with a linear activation function, and mean squared error as its loss function. The model's output is thus a single value; to perform classification, this output value is rounded to the nearest integer in the interval $[0, t-1]$, where t is the number of target classes.

The final two loss functions perform metric learning rather than performing classification directly. They learn representations of a user's activity and of the four self-harm risk severity labels; classification is performed by comparing the euclidean distance between a representation of a user's activity (produced by the final layer) and each of the four severity label representations.

Method		Convolution			Dense Layers	Dropout	Class Balance
		Size	Filters	Pool Len.			
Reddit	Cat. Cross Ent.	3	25	all (avg)	1 w/ 50 dims	0.0	Sampled
ReachOut	Cat. Cross Ent.	3	150	3 (max)	2 w/ 250 dims	0.3	Weighted
	MSE	3	100	3 (max)	2 w/ 250 dims	0.5	Sampled
	Class Metric	3	100	3 (max)	2 w/ 150 dims	0.3	Sampled

Table 1: The hyperparameters used by each model.

Class Metric: Let d be the size of the output layer and X be the layer's d-dimensional output. *Class Metric* learns a d-dimensional representation of each class C_i such that $||X - C_i||_2$ is minimized for the correct class i; this is accomplished with the loss function: $L_{i,p,n} = max(0, ||X_i - C_p||_2 - ||X_i - C_n||_2 + \alpha)$ where C_p is the correct (i.e., positive) class for X_i, C_n is a randomly chosen incorrect (i.e., negative) class, and α is a constant to enforce a minimum margin between classes. Classification is performed by computing the similarity between X_i and each class C_j.

Class Metric (Ordinal) extends *Class Metric* to enforce a margin between ordinal classes as a function of the distance between classes. Given a ranked list of classes such that more similar classes have closer rankings, that is $\forall i \; sim(C_i, C_{i\pm1}) > sim(C_i, C_{i\pm2})$, we incorporate the class distance into the margin such that more distant incorrect class labels must be further away from the correct class label in the metric space. The loss function becomes $L_{i,p,n} = max(0, ||X_i - C_p||_2 - ||X_i - C_n||_2 + \alpha|p - n|)$ where $|p - n|$ causes the margin to scale with the distance between classes p and n.

5 Experiments

In this section we describe the model hyperparameters used and present our results on the depression detection and self-harm risk assessment tasks. To facilitate reproducibility we provide our code and will provide the Reddit depression dataset to researchers who sign a data usage agreement[4].

5.1 Experimental setup.

The hyperparameters used with our models are shown in Table 1. The severity risk assessment models' hyperparameters were chosen using 10-fold cross validation on the 947 ReachOut training posts, with 15% of each fold used as validation data. The depression identification model's hyperparameters were chosen using the Reddit validation set. The depression identification model's second convolutional layer (i.e., the layer used to merge post representations) used filters of length 15, a stride of length 15, and the same number of filters as the first convolutional layer. All models were trained using stochastic gradient descent with the Adam optimizer (Kingma and Ba, 2014). The hyperparameters that varied across models are shown in Table 1. The convolution size, number of convolutional filters, pooling type, pooling length, and number of dense layers was similar across all post models. Class balancing was performed with *Categorical Cross Ent.* by weighting classes inversely proportional to their frequencies, whereas sampling an equal number of instances for each class worked best with the other methods.

Addressing limited data. The post classification models' input consists of skip-thought vectors (Kiros et al., 2015); each vector used is a 7200-dimensional representation of a sentence. Thus, the convolutional windows used for post classification are over sentences rather than over terms. This input representation was chosen to mitigate the effects of the ReachOut dataset's relatively small size. The skip-thought vectors were generated from the the ReachOut forum dataset by sequentially splitting the posts in the training set into sentences, tokenizing them, and training skip-thoughts using Kiros et al.'s implementation with the default parameters. Sentence boundary detection was performed using the Punkt sentence tokenizer (Kiss and Strunk, 2006) available in NLTK (Bird et al., 2009). These 2400-dimensional forum post skip-thought vectors were concatenated with the 4800-dimensional book corpus skip-thought vectors available from Kiros et al.. Experiments on the training set indicated that using only the ReachOut skip-thought vectors slightly decreased

[4] http://ir.cs.georgetown.edu/data/reddit_depression/

performance, while using only the book corpus skip-thought vectors substantially decreased performance. As input the post models received the last 20 sentences in each target post and the last 20 sentences in the thread prior to the target post; any prior sentences are ignored.

5.2 Depression detection.

The data used for depression detection was described in §3. We compare our model against several baselines using MNB and SVM classifiers (Wang and Manning, 2012). Specifically, we consider two sets of features for the classifiers. The first set of features is the post content itself represented as sparse bag of words features (*BoW baselines*). The second set of features (*feature-rich baselines*) comprises a large set of features including bag of words features encoded as sparse weighted vectors, external psycholinguistic features captured by LIWC[5] (2015), and emotion lexicon features (Staiano and Guerini, 2014). Since our problem is identifying depression among users, psycholinguistic signals and emotional attributes in the text are potentially important features for the task. These features (as described in §2) have been also previously used by successful methods in the Twitter self-reported diagnosis detection task (Coppersmith et al., 2015b). Thus, we argue that these are strong baselines for our self-reported diagnosis detection task. We apply count based and tf-idf based feature weighting for bag of words features. We perform standard preprocessing by removing stopwords and lowercasing the input text.[6]

The data is split into training, validation, and testing datasets each containing approximately 3,000 diagnosed users and their matched control users. The validation set is used for tuning development and hyperparameter tuning of our models and the baselines. The reported results are on the test set. The depression detection models' input consisted of raw terms encoded as one-hot vectors. We used an input layer to learn 50-dimensional representation of the terms. For each target user, the CNN received up to 400 posts containing up to 100 terms; experiments on the validation data indicated that increasing the maximum number of posts did not significantly improve performance.

Method	Precision	Recall	F1
BoW - MNB	0.44	0.31	0.36
BoW - SVM	**0.72**	0.29	0.42
Feature-rich - MNB	0.69	0.32	0.44
Feature-rich - SVM	0.71	0.31	0.44
User model - CNN	0.59	**0.45**	**0.51**

Table 2: Performance identifying depressed users on the Reddit test set. The differences between the CNN and baselines are statistically significant (McNemar's test, $p < 0.05$).

Results. The results of identifying depressed users for our model and baselines are shown in Table 2. Our proposed model outperforms the baselines by a large margin in terms of recall and F1 on the diagnosed users (increases of 41% and 16%, respectively), but performs worse in terms of precision. As described later in the analysis section, the CNN identifies language associated with negative sentiment across a user's posts.

5.3 Self-harm risk classification.

We train our methods to label the ReachOut posts and compare them against the top methods from CLPsych '16. We use the same experimental protocol as was used in CLPsych '16; our methods were trained on the 947 training posts and evaluated on the remaining 280 testing posts. We used 15% of the 947 training posts as validation data.

We report results using the same metrics used in CLPsych, which were: the macro-averaged F1 for the $amber$, red, and $crisis$ labels (*non-green* posts); the macro-averaged F1 of $green$ posts vs. $amber \cup red \cup crisis$ (*flagged* posts); and the macro-averaged F1 of $green \cup amber$ vs. $red \cup crisis$ (*urgent* posts). The *non-green* F1 was used as the official CLPsych metric with the intention of placing emphasis on classification performance for the non-green categories (i.e., those that required some response). The binary *flagged* meta-class was chosen to measure models' abilities to differentiate between posts that require attention and posts that do not, and the binary *urgent* meta-class was chosen to measure their abilities to differentiate between posts that require quick responses and posts that do not. In addition to macro-averaged F1, CLPsych also reported the accuracy for each category. We additionally report F1 macro-averaged over all classes.

[5] http://liwc.wpengine.com/

[6] During experimentation, we found tf-idf sparse feature weighting to be superior than other weighting schemes. Additional features such as LDA topics and χ^2 feature selection did not result in any further improvements.

Method	Non-green	Flagged		Urgent		All	
	F1	F1	Acc.	F1	Acc.	F1	Acc.
Baseline (Milne et al., 2016)	0.31	0.78	0.86	0.38	0.89	-	-
Kim et al. (2016)	0.42	0.85	0.91	0.62	0.91	0.55	0.85
Malmasi et al. (2016)	0.42	0.87	0.91	0.64	0.93	0.55	0.83
Brew (2016)	0.42	0.78	0.85	0.69	0.93	0.54	0.79
Cohan et al. (2016)	0.41	0.81	0.87	0.67	0.92	0.53	0.80
Categorical Cross Ent.	**0.50**	**0.89**	**0.93**	0.70	**0.94**	**0.61**	**0.89**
MSE	0.42	0.80	0.85	0.64	0.93	0.53	0.78
Class Metric	0.46	0.79	0.84	0.70	**0.94**	0.56	0.80
Class Metric (Ordinal)	0.47	0.88	**0.93**	**0.72**	0.93	0.59	0.87

Table 3: Self-harm risk assessment performance on the ReachOut test posts. F1 and accuracy are aggregated as specified by CLPsych '16. The reported results for the other methods are the official numbers from (Milne et al., 2016). The differences in performance between the following method pairs are statistically significant (McNemar's test, $p < 0.05$): *Categorical Cross Ent.* and *Class Metric*, *MSE* and *Categorical Cross Ent.*, *MSE* and *Class Metric (Ordinal)*, and *Class Metric (Ordinal)* and *Class Metric*.

Method	Non-green	Flagged		Urgent		All	
	F1	F1	Acc.	F1	Acc.	F1	Acc.
Categorical Cross Ent.	0.54	0.87	0.89	0.69	0.91	0.63	0.80
MSE	**0.87**	**0.95**	**0.96**	**0.91**	**0.98**	**0.89**	**0.93**
Class Metric	0.73	0.90	0.91	0.81	0.94	0.78	0.86
Class Metric (Ordinal)	0.85	**0.95**	**0.96**	0.89	0.97	0.88	0.92

Table 4: Self-harm risk assessment performance on the ReachOut training set using 10-fold cross validation. *Categorical Cross Ent.* performs substantially worse than on the test set, while *MSE* performs substantially better. *Class Metric (Ordinal)* continues to perform well. The difference in performance between the following method pairs are statistically significant (McNemar's test, $p < 0.05$): *Categorical Cross Ent.* and *MSE*, *Categorical Cross Ent.* and *Class Metric*, *Categorical Cross Ent.* and *Class Metric (Ordinal)*, *MSE* and *Class Metric*, and *Class Metric* and *Class Metric (Ordinal)*.

Results. The results on the self-harm risk assessment task for our models and for the current best-performing methods (briefly explained in §2) are shown in Table 3. We also report a baseline result which is based on a SVM classifier with bigram features. When measured by *non-green* F1, the official metric of the CLPsych '16 Triage Task, our proposed models perform up to 19% better than the best existing methods. Similarly, our models perform up to 11% better when measured with an F1 macro-averaged across all categories (i.e., *all* column) and up to 5% better with measured accuracy across all categories. *Categorical Cross Ent.* performs best in all of these cases, though the difference between the performance of *Categorical Cross Ent.* and *Class Metric* with an ordinal margin is not statistically significant.

We also evaluate the performance of our methods on the training set using 10-fold cross validation to better observe the performance differences between our model variants (Table 4). All models' perform substantially better on the training set than on the test set. This is partially explained by the fact that the models were tuned on the training set, but the large difference in some cases (e.g., the increase in the highest non-green F1 from 0.50 to 0.87) suggest there may be qualitative differences between the training and testing sets. The best-performing method on the test set, *Categorical Cross Ent.*, performs the worst on the training set. Similarly, the worst-performing method on the test set, *MSE*, performs the best on the training set. *Class Metric (Ordinal)* performs well on both the testing and training sets, however, suggesting that it is more robust than the other methods. Furthermore, there is no statistically significant difference between *Class Metric (Ordinal)* and the best-performing method on either dataset.

Top Phrases	
i went to	to scare you
my whole	to have it
sometimes i	my son was
i'm so sorry	it wasn't

Table 5: Example phrases that strongly contributed to a user's depression classification on the RSDD dataset.

5.4 Analysis

In this section we analyze the language that strongly contributed to the identification of depressed users on the Reddit dataset. Note that it is impossible to show entire Reddit posts without compromising users' anonymity; we found that even when a post is paraphrased, enough information remains that it can easily be identified using a Web search engine. For example, one Reddit post that strongly contributed to the author's classification as a depressed user contained the mention of a specific type of abuse and several comments vaguely related to this type of abuse. We attempted to paraphrase this post, but found that any paraphrase containing general language related to both the type of abuse and to the user's comments was enough to identify the user. Thus, to protect the anonymity of the users in our dataset, we do not publish posts in any form.

Rather than publishing posts, we identify key phrases in posts from users who were correctly identified as being depressed. Phrases from eight self-reported depressed users are shown in Table 5; to prevent these phrases from being used to identify users, we retain only the top phrase from each user. These phrases were identified by using the model's convolutional filter weights to identify posts in the validation dataset that are strongly contributing to the model's classification decision, and then using the convolutional filter weights to identify the phrase within each post that most strongly contributed to the post's classification (i.e., had the highest feature values).

In keeping with the design of our dataset, terms related to depression or diagnoses are not present. Instead, the model identifies phrases that often could be associated with a negative sentiment or outlook. For example, "my whole" could be part of a negative comment referring to the poster's whole life. It should be noted that the model makes classification decisions based on the occurrence of phrases across many posts by the same user. Though one can imagine how the phrases shown here could be used to convey negative sentiment, the presence of a single such phrase is not sufficient to cause the model to classify a user as depressed.

6 Conclusion

In this work we argued for the close connection between social media and mental health, and described a neural network architecture for performing self-harm risk classification and depression detection on social media posts. We described the construction of the Reddit Self-reported Depression Diagnosis (RSDD) dataset, containing over 9,000 users with self-reported depression diagnoses matched with over 107,000 similar control users; the dataset is available under a data usage agreement. We applied our classification approach to the task of identifying depressed users on this dataset and found that it substantially outperformed strong existing methods in terms of Recall and F1. While these depression detection results are encouraging, the absolute values of the metrics illustrate that this is a challenging task and worthy of further exploration. We also applied our classification approach to the task of estimating the self-harm risk posed by posts on the ReachOut.com mental health support forum, and found that it substantially outperformed strong previously-proposed methods.

Our approach and results are significant from several perspectives: they provide a strong approach to identifying posts indicating a risk of self-harm in social media; they demonstrate a means for large scale public mental health studies surrounding the state of depression; and they demonstrate the possibility of sensitive applications in the context of clinical care, where clinicians could be notified if the activities of their patients suggest they are at risk of self-harm. Furthermore, large-scale datasets such as the one presented in this paper can provide complementary information to existing data on mental health which are generally relatively smaller collections.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *arXiv preprint arXiv:1605.04462*.

Anxiety and Depression Association of America. 2015. Anxiety and depression, facts & statistics.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017a. Ethical research protocols for social media health research. *EACL 2017*, page 94.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017b. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Chris Brew. 2016. Classifying reachout posts with a radial basis function svm. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 138–142, San Diego, CA, USA. Association for Computational Linguistics.

CDC. 2015. Suicide fact sheet, suicide facts at a glance. *National Center for Injury Prevention and Control*.

Munmun De Choudhury and Emre Kiciman. 2017. The language of social support in social media and its effect on suicidal ideation risk. In *Proceedings of the International Conference onWeb and Social Media (ICWSM)*. AAAI.

Arman Cohan, Sydney Young, and Nazli Goharian. 2016. Triaging mental health forum posts. *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 143–147.

Arman Cohan, Sydney Young, Andrew Yates, and Nazli Goharian. 2017. Triaging content severity in online mental health forums. *Journal of the Association for Information Science and Technology*.

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

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 *CLPysch*, pages 1–10.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015b. Clpsych 2015 shared task: Depression and ptsd on twitter. *NAACL HLT 2015*, page 31.

Glen Coppersmith, Craig Harman, and Mark Dredze. 2014b. Measuring post traumatic stress disorder in twitter. In *ICWSM*.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. AAAI.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 34rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '16. ACM.

Alize J Ferrari, Rosana E Norman, Greg Freedman, Amanda J Baxter, Jane E Pirkis, Meredith G Harris, Andrew Page, Emily Carnahan, Louisa Degenhardt, Theo Vos, et al. 2014. The burden attributable to mental and substance use disorders as risk factors for suicide: findings from the global burden of disease study 2010. *PLoS One*, 9(4):e91936.

John F Gunn and David Lester. 2015. Twitter postings and suicide: An analysis of the postings of a fatal suicide in the 24 hours prior to death. *Suicidologi*, 17(3).

Kristy Hollingshead and Lyle Ungar, editors. 2016. *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics, San Diego, CA, USA.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 591–598.

Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. 2014. Tracking suicide risk factors through Twitter in the US. *Crisis*.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665. Association for Computational Linguistics.

Sunghwan Mac Kim, Yufei Wang, Stephen Wan, and Cecile Paris. 2016. Data61-csiro systems at the clpsych 2016 shared task. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 128–132, San Diego, CA, USA. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprints*, arXiv:1412.6980.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In *NIPS*, Cambridge, MA, USA. MIT Press.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Comput. Linguist.*, 32(4).

David E Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39.

Shervin Malmasi, Marcos Zampieri, and Mark Dras. 2016. Predicting post severity in mental health forums. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 133–137, San Diego, CA, USA. Association for Computational Linguistics.

L Manikonda and M De Choudhury. 2017. Modeling and understanding visual attributes of mental health disclosures in social media. In *CHI '17*.

David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo. 2016. Clpsych 2016 shared task: Triaging content in online peer-support forums. *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–127.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. *NAACL-HLT Workshop on CLPsych 2015*, page 11.

Danielle Mowery, Albert Park, Mike Conway, and Craig Bryan. 2016. Towards automatically classifying depressive symptoms from twitter data for population health. In *Proceedings of the Workshop on Computational Modeling of Peoples Opinions, Personality, and Emotions in Social Media*, pages 182–191.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. *UT Faculty/Researcher Works*.

Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H. Andrew Schwartz, and Lyle Ungar. 2015. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*.

Philip Resnik, William Armstrong, Leonardo Claudino, and Thang Nguyen. 2015. The university of maryland clpsych 2015 shared task system. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 54–60.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression. In *EMNLP*, pages 1348–1353. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Jacopo Staiano and Marco Guerini. 2014. Depechemood: a lexicon for emotion analysis from crowd-annotated news. *arXiv preprints*, arXiv:1405.1605.

Simon Šuster, Stéphan Tulkens, and Walter Daelemans. 2017. A short review of ethical challenges in clinical natural language processing. *arXiv preprint arXiv:1703.10090*.

Paul Thompson, Craig Bryan, and Chris Poulin. 2014. Predicting military and veteran suicide risk: Cultural aspects. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–6, Baltimore, Maryland, USA. Association for Computational Linguistics.

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing depression from twitter activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM.

R Jay Turner, B Gail Frankel, and Deborah M Levin. 1983. Social support: Conceptualization, measurement, and implications for mental health. *Research in Community & Mental Health*.

Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *ACL '12*.

WHO. 2010. *World Health Organization, World Health Statistics 2010*. World Health Organization.

Yijun Xiao and Kyunghyun Cho. 2016. Efficient character-level document classification by combining convolution and recurrent layers. *arXiv preprints*, arXiv:1602.00367.