A Domain Independent Social Media Depression Detection Model

Sven Marnauzs
Department of Mathematics
Boise State University
svenmarnauzs@u.boisestate.edu

Jugal Kalita
Department of Computer Science
UC Colorado Springs
jkalita@uccs.edu

Abstract

Due to negative social stigmas surrounding mental health disorders, a large portion of the affected population are reluctant to seek help. In consequence, many of those in need remain undiagnosed and uneducated about their condition. Language contains information on the author's mental state and demographic, and has been leveraged by NLP models that learn to predict each one independently or jointly. With the advent of social media, we have access to an unprecedented amount of natural language data. However, these models require large labeled data sets that can be very expensive and time consuming to create. In addition, these data sets are typically derived from a single domain such as Twitter, Facebook, Reddit, Instagram, etc. As such, they are unable to generalize well to other mediums outside of their domain. Focusing on the diagnosis of depression disorder, the present contribution aims to create a domain independent depression detection model that uses a novel combination of deep learning, NLP, and machine learning techniques while using datasets much smaller than those found in the literature.

Introduction

The World Health Organization (WHO) reports that mental health disorders such as depression and anxiety are among the largest contributors to global disability. These types of disorders are detrimental to every aspect of an affected individuals health. Depression alone is estimated to affect more than 300 million people worldwide. Yet, a large portion of this population remains undiagnosed and reluctant to seek help due to the negative social stigma surrounding such actions or time and financial limitations.

There has been a large push to use recent advancements in Natural Language Processing (NLP) that can leverage social media data as a depression detection system (De Choudhury et al. 2013), (Coppersmith, Dredze, and Harman 2014), (Coppersmith et al. 2015). There has also been work on using multitask learning to predict mental disorders (depression, PTSD, bipolar, etc.) and gender simultaneously in an attempt to be able to use smaller datasets (Hovy, Mitchell, and Benton 2017). More recently, there has been work to negate the affects of data imbalance in datasets created for the depression detection task (Cong et al. 2018), (Gerych, Agu, and Rundensteiner 2019).

However, previous studies have either only used data from one type of domain or have data from multiple domains but no way of fine tuning that data for new data sets. To the best of our knowledge, there is not any work being done to create a general depression detection system that can be fine tuned to data outside of its domain. For example, a depression detection classifier based on Facebook data may not classify well on data taken from Twitter. In general, end-to-end deep learning approaches are not feasible on small datasets.

In the present contribution, we propose a general depression detection model. This model will take in labeled natural language data provided by the user, feed it to various depression classifiers from different domains, and then use predictions from those classifiers as features in a machine learning model. Our hypothesis is that the machine learning model will be able to determine which classifiers are able to best diagnose depression in the new domain.

Related Work

Predicting Depression via Social Media

Using the CES-D (Center for Epidemiologic Studies Depression Scale) questionnaire, De Choudhury et al. (2013) determined the depression level of 1,583 crowdsourced subjects. At the end of the session, they asked each subject if they would like to share their Twitter username. From those that agreed, they created a database of 544 twitter user names and their corresponding tweets, and then labeled each user as depressed or not depressed based on their answers to the CES-D questionnaire. The final database contained 171 users classified as depressed and 305 as not depressed. For those classified as depressed, they only kept tweets from one year prior to their onset or diagnosis date in an attempt to create a model that can predict future episodes of depression. They found that their SVM based model with 188 features was able to classify depressed users with about 70 percent accuracy. While their results were promising, their collection and feature extraction methods were very time consuming and expensive.

Multitask Learning for Mental Health Conditions with Limited Data

It has been shown that there can be a considerable performance jump when transitioning from Single Task Learning (STL) to MTL (Caruana 1993). Hovy, Mitchell, and Benton (2017) were the first to use deep-learning in the task of mental disorder detection. In addition, they used a combination of automatic data collection from twitter and hand-annotation to create their labeled database. While their methods may not be as concrete as De Choudhury's, they were certainly less time consuming. They developed neural MTL models trained on twitter data for 8 mental condition prediction tasks and 2 auxiliary prediction tasks (neurotypicality, and gender). They found that their most complex MTL models (ones that simultaneously trained the most tasks) performed significantly better than independent STL models when they had smaller amounts of data on certain conditions such as bipolar disorder and PTSD. They theorize that when they force the model to predict conditions which have large amounts of data available, they significantly improve the prediction accuracy of other similar conditions with small amounts of available data.

X-A-BiLSTM: a Deep Learning Approach for Depression Detection in Imbalanced Data

Deep learning approaches to the social media depression detection task are hindered by imbalanced datasets. In an attempt to lift this weight, Cong et al. (2018) use a novel combination of traditional machine learning techniques and deep-learning. To validate their proposed model, they used the RSDD dataset (Yates, Cohan, and Goharian 2017). Language data of each author in the training set was initially feed into an XGBoost model that significantly reduced the degree of imbalance in the dataset by weeding out negative samples. The positive predictions were then feed into an attention Bi-LSTM deep-learning model to output the final classes of positive and negative samples. Their proposed model surpassed several other state-of-the-art deep-learning models in the social media depression classification task.

Data

Data was collected from two social media domains. The first is the RSDD dataset from Yates et al. (2017), which is composed of public Reddit posts. The second is the CLPsych 2015 shared task Twitter dataset from Coppersmith et al. (2015). Both datasets were created via a combination of automatic retrieval and hand annotated labeling similar to the procedures of Coppersmith et al. (2014). Social media users choose to publicly post statuses of their mental health for a variety of reasons such as looking for support and sympathy from their social network. Another common reason to publicly report their diagnosis is to educate others about their condition. Nevertheless, publicly available posts on being clinically diagnosed with depression provides us with an opportunity to explore the language differences between depressed and non-depressed users. As both the RSDD dataset and the CLPsych 2015 shared task dataset use these publicly available posts of diagnosis, their ground truth values share a similar degree of reliability. This allows us to conduct a exploration of the differences and similarities in how users communicate on their respective platforms. Our intuition tells us that there should be a difference in the way that depressed and non-depressed users use language. For example, a good generalized model trained on domain A should be able to classify users on domain B with a precision close to that of a good model being tested and trained on solely on B. The acquired datasets allow us to explore the correctness of this intuition.

We also make use of age and gender labeled datasets provided by various 2017 PAN shared tasks (Potthast et al. 2017). These datasets are of the tweets from hundreds of Twitter users, where each user is given a label for gender, and then their respective age group (18-24, 25-34, 35-49, 50-XX). Since depression is slightly correlated with age and gender, our intuition tells us that incorporating this information into our model should increase its predictive accuracy.

Data preprocessing

In an attempt to remove the effect of domain specific text patterns on our model, we take all data through various preprocessing techniques. However, we still seek to capture as much unique and quirky language from the users as possible, so we make our best effort strike a balance between preprocessing and leaving the data as is. We also must take into account that our primary model is an n-gram character language model (CLM). Therefore, by reducing the character vocabulary of the text, we substantially reduce the complexity of our model.

RSDD Reddit dataset As Reddit posts have essentially no character limit, they are typically filled with special formatting such as newlines, tabs, links, etc. The first step was to remove these unnecessary gaps between sections of text. This first step greatly reduced the complexity of our higher order n-gram CLMs. Next, we substituted references to links, usernames, and subreddits with special characters or just simply removed them. It is possible that references to subreddits could potentially help a model determine the classification for a user. For example, depressed users may mention r/depression more often than other users. However, to keep our model as generalized as possible, we chose to remove such references as they are not present in other social media domains. Another decision we made was to change all text to lowercase for the sake of simplicity.

CLPsych 2015 and PAN 2017 Twitter dataset Even though Twitter has its own unique style when compared to Reddit, most of the preprocessing steps were identical. We remove links, usernames, newlines, tabs, and unnecessary whitespace. Unique to twitter is the retweet type of post. We removed all retweets as these tweets are not written by the user. We chose to keep in standard punctuation, however these may need to be removed so that we can reduce the complexity of our CLMs. At the order four CLM, we already had 10^6 four-grams, and we were not able to upscale

to a higher order due to a memory error. Due to complications, the PAN dataset was not filtered or processed in any way.

Model Architecture

Baseline Model Social media is rampant with internet-slang, misspelled words, niche text patterns, and obscure references. As a simple solution to capture all of these nuances, we employ n-gram character-level language models with k-smoothing to score an aggregate of x tweets at a time, where the scores indicate whether a user is depressed or not. Our approach follows closely to the MIQ team approach in CLPsych 2015 shared task competition; see Coppersmith et al. (2015) for more details. Through this approach we examine how likely a sequence of characters is to be generated by a depressed or non-depressed user. We begin by building an n-gram character-level model for each condition based on the training subset of our data. For each user in the test set, we score an aggregate of x tweets based on its character-level n-grams. Let T_x be the aggregate of x tweets for a given user, D the CLM for depression, and D' the CLM for control. Our scoring function f is thus

$$f(T_x) = \frac{\sum_{T_x} \log p(c_D) - \log p(c_{D'})}{|T_x|}$$

where $p(c_D)$ is the probability of an n-gram character sequence appearing in model D, and $p(c_{D'})$ is the probability of an n-gram character sequence appearing in model D'. To compute the final score for each individual user, we average the scores in a sliding window of five x tweet aggregates at a time. Once a score is obtained for each window, the median of the set of window scores is used as the final score for the user. Our model was essentially identical between training and testing on the RSDD and CLPsych datasets. However, for the RSDD data, we did not use a sliding window. The predictive accuracy of our CLMs will serve as our cross-domain baseline.

Improved model To improve upon the performance of our baseline cross-domain model, we propose a model that uses a combination of conventional machine learning, deep learning, and multi-task learning to create an automatic domain independent social media depression detector. We use a deep-learning transformer model from Google's Tensor2Tensor library to pre-train age, gender, and depression classifiers (Vaswani et al. 2018). To test our cross-domain model, we use the RSDD dataset to train our depression classifier, and then test the model on our Twitter data. We use the depression, gender, and age labels as semi-supervised features in a Random Forest Classifier (RFC) model, where the labels are produces by feeding in the Twitter data to each pre-trained classifier. We show a generalized diagram of our model in Figure (1), where $C : [C_1, C_2, ..., C_n]$ is the set of classifiers, n is the number of classifiers, a is the number of authors in the local database, $O : [C_{1,i}, C_{2,i}, ..., C_{n,i}]$ is a $n \times a$ dimensional matrix that holds the outputs of each classifier for author i where $i \, \epsilon [1, 2, ..., a]$, and y is an array that contains a label for each author (depressed or not depressed).

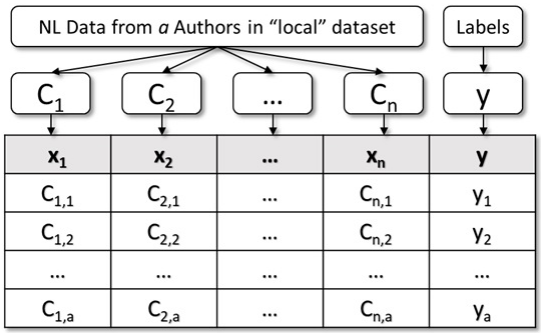

Figure 1: Semi-supervised ML Model Flowchart

As can be seen in the above diagram, natural language data from n Twitter authors is processed by the pre-trained classifiers in C, and the outputs O will be used as values for each author's feature vector. This dataset of feature vectors and corresponding y labels will then be split into training and testing datasets used to train the RFC model.

Experiments

RSDD For the RSDD data, we experimented with one, two, and three gram CLMs. We also experimented with many different combinations of window and aggregate post size. We found that we obtained the best results when compiling 20 ($x = 20$) posts and omitting the window approach entirely. Filtered and unfiltered data were also trained and tested on. Filtered data out-preformed unfiltered data in every case.
We also trained a transformer model on the RSDD data, but we did not conduct an in depth evaluation of this model. This model was only used for generating semi-supervised depression labels of Twitter users in our CLPsych 2015 dataset.

CLPsych 2015 shared task Due to a smaller dataset size, we were able to expand to a four order CLM in addition to the lower orders. We also experimented with different combinations of window and aggregate post size, and found that aggregating 20 posts together with a window size of 5 gave us the best results. As with the RSDD experimentation, filtered data out-preformed unfiltered in every case.
We train a transformer model on the CLPsych Twitter data to compare its predictive accuracy against our cross-domain model and our CLM model. The transformer model gave a binary label to each post of a user. For the final score of each user, we used the median of an array of 20 post sliding window scores, where each window was scored as the ratio of depressed posts to non-depressed posts. This experiment was done to verify that the results of the transformer model trained on a single domain were as good as or better than our baseline model results.

Cross domain testing baseline The main objective of the current contribution is to approve upon a baseline test on how well a model trained on one social media domain will fair when given data from a different domain. We experimented with using the one, two, and three RSDD CLMs

to score Twitter users from the CLPsych 2015 shared task dataset. As this kind of cross domain testing has not been explored before, we use the result as a baseline test, and seek to improve upon it.

Improved cross domain model To improve upon the baseline, we use a RFC with features extracted from feeding Twitter data into our pre-trained transformer model classifiers. Each pre-trained classifier outputs a label for each post of a user. We then use the sliding window approach (size 20) to produce the final labels for gender, age, and depression for each user in the Twitter dataset. This dataset was split into training and testing, where the training set was used to train the RFC model, and the testing set was used for validation and baseline comparison tests.

Results

Baseline model results

Figure (1) shows the best results of training and testing a CLM on the RSDD dataset. We found that taking the median of 20 post aggregate scores that were scored by a second order CLM trained and tested on filtered text gave us the best results. The AUC of our model is 0.79.

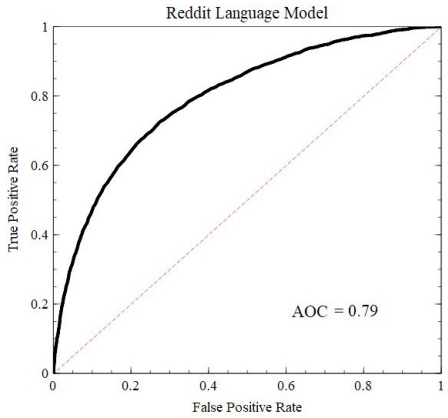

Figure 2: The ROC curve for a second order CLM trained and test on filtered Reddit data.

Figure (2) shows the best results of training and testing a CLM on the CLPsych dataset. We found that taking the median of window scores of size five gave us the best results, where each window took the mean of 20 post aggregate scores that were scored by a second order CLM trained and tested on filtered text. We found the average precision was 0.65, and the AOC was 0.80 for our model.

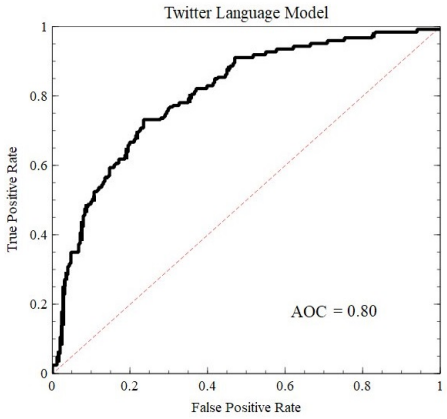

Figure 3: The ROC curve for a order four CLM trained and tested on filtered Twitter data.

Figure (3) shows how well our order three Reddit CLM preformed on classifying twitter users. The parameters of both models were identical to the models which gave us the best results in their respective domains. The AUC and average precision of the Reddit based CLM was 0.67 and 0.55 respectively. For comparison, the AUC and average precision of the Twitter based CLM was 0.78 and 0.62 respectively.

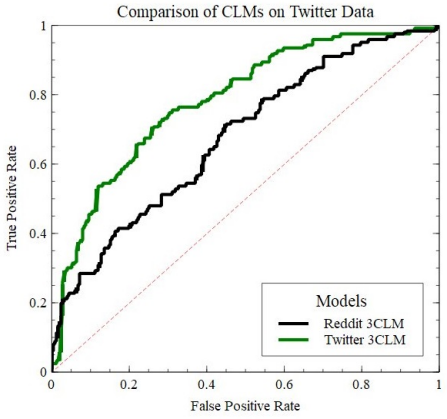

Figure 4: A comparison of ROC curves for order three CLMs trained on filtered Twitter data and Reddit data, and then tested only on Twitter data.

Figure (4) shows a comparison between a $n = 3$ CLM trained and tested on Twitter data, and a $n = 3$ CLM trained on Reddit but tested on Twitter. For the Twitter CLM, we obtained an AUC score of 0.78 with and average precision of 0.63. The Reddit CLM gave us an AUC score of 0.63 with an average precision of 0.55. We were unable to upscale to a higher-order CLM for the RSDD data due to various complications. As such, we thought it fair to compare the 3-gram RSDD CLM to the 3-gram CLPsych CLM instead of the better performing 4-gram CLM.

Transformer/RFC model results

We explore the results of our "improved" models against the baseline models. We used the best results from each model type instead

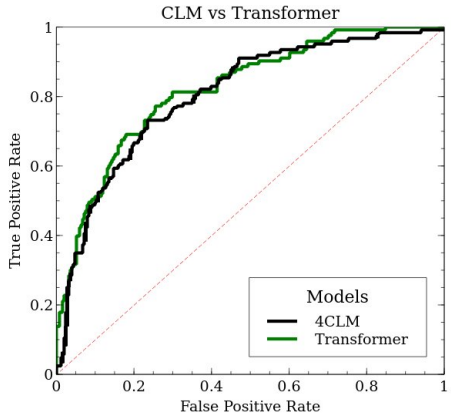

Figure 5: A comparison between the transformer model and $n = 4$ CLM trained and tested on the CLPsych 2015 Twitter data.

Figure (5) shows the results of the $n = 4$ CLM and our transformer model trained and tested on Twitter data. This single domain experiment was conducted to verify that our transformer model could perform as well as our baseline model when trained on a single domain. The 4-gram CLM had an AUC of 0.80 with an average precision of 0.65. Our transformer model had an AUC of 0.83 with an average precision of 0.72

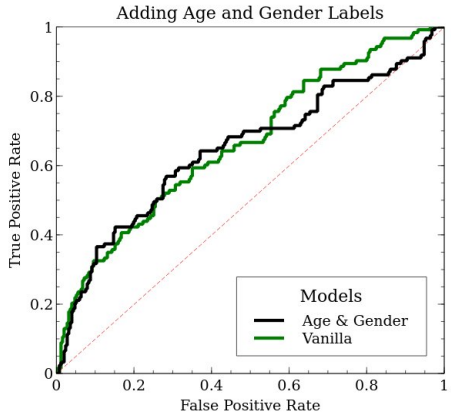

Figure 6: A comparison between our RFC model with semi-supervised labels and a "vanilla" (only depression labels) transformer model when tested on the CLPsych data. Both models were never trained on any of the CLPsych data.

We explore the effect of adding age and gender information to our model in Figure (6). Both models had an AUC of 0.65. Our vanilla model gave us an average precision of 0.53, while our RFC model had an average precision of 0.52.

Discussion

The performance of our CLM models on Reddit and Twitter data were surprisingly good (Figure (1) and (2)). We expect that as we are able to increase the order of our CLMs we will achieve better performance. As can be seen in Figure (3), our Reddit base CLM model does not preform as well as the Twitter based one. To have a fair comparison between models, both models used third order CLMs. We attribute the poor performance to the large differences in how Twitter and Reddit operate as social media sites. One possible explanation is that CLMs are not be able to pick up the true sentiment of users posts. Another could be that we simply did not take advantage of higher order CLMs. We expect that as we are able to increase the order of both CLMs, the gap between their performance will decrease. However, we do not know if this gap will ever converge using just CLMs. Nevertheless, we will use the CLM models as our baseline performance.

The performance of Google's transformer model trained and tested on the CLPsych data set was good, but not as good as we had hoped it would be, as it only slightly outperformed our simple 4-gram CLM. However, we did not attempt to use an aggregate training or scoring method. These methods significantly improved the performance of our CLMs. Therefore, we expect that we would achieve a similar performance boost for the transformer model as well.

Both the "vanilla" transformer model and our RFC model (with age and gender labels) produced very similar results. This came to our surprise, as we intuitively theorized that incorporating this extra information into a model would increase the cross-domain generalization, thus increasing the accuracy once shown data from another domain. There are multiple explanation for why our results were not satisfactory. First, we did not pre-process the age and gender data due to time constraints and complications. We also did not do a through evaluation of how well the age and gender models were performing on ground truth age and gender labeled Twitter data. Going through these tasks and evaluations would likely improve the accuracy of our semi-supervised labels for our RFC model. There is also a possibility that the RFC is not taking advantage of the gender and age labels due to over-weighing the depression labels.

Future Work

Model exploration and fine-tuning, multi-task learning, and extra features

First, our RFC may not be taking advantage of additional information. We should explore other models that will better use this extra data. We will also attempt to fine-tune parameters of the RFC to see if it can make better use of all its features. There is also a good chance that we can make our transformer perform better across all tasks by fine-tuning parameters and learning how to train it better, as these are not trivial tasks. For example, we will train a transformer model on an aggregates of posts instead of individual ones. However, we are experiencing memory error when attempting to do so. We will likely find a way to overcome this obstacle in the near future, and when we do, we expect a large across-the-board performance increase. Text CNNs for post classification are also a topic of interest (Kim 2014). We have already begun initial experimentation using this approach.

The possibility of using multi-task learning for classifying

depressed users is very intriguing. We have proposed using a text CNN network to pre-train on the gender and age data, and then use those trained models to calculate vectors for each sample in the CLPsych dataset. We would then train another text CNN model on the CLPsych depression data while incorporating the pre-calculated vectors. We have created the groundwork for this approach, but we have yet to fully implement it. We theorize that making our text CNN model train on several tasks at once will increase the performance across all tasks. It could also improve the cross-domain compatibility of the model (i.e. perform good when given data from other domains).

Finally, the addition of features in our RFC could help improve the cross-domain performance of the model. Other features considered are personality traits, the general emotion of a users posts, and additional depression classifiers. These addional depression classifiers could each have differnet scoring methods and models. For example, we could use a combination of multi-task and single-task models using different parameters for training and scoring (i.e. different window sizes).

Obtaining relevant datasets

For our model to work as planned, we must seek out additional datasets other than RSDD and CLPsych 2015. We have already obtained several Twitter datasets related to age and gender tasks that were created for a shared task competitions at the PAN workshops held annually at CLEF. We have also obtained the DAIC-WOZ dataset which contains transcribed interviews of subjects and their ground truth scores of depression based on PHQ-8 questionnaires. Aside from these datasets, there is possibility that we will want to incorporate a dataset from a third social media domain such as Facebook. This would allow us to perform three-fold cross validation of our cross-domain model. A three-fold cross validation scheme would greatly increase the validity of our model because using just two social media domains would likely lead to over fitting our model.

Conclusion

Much work has been done to improve social media depression detection models. However, these models are trained and tested on a single social media domain. As such, they are likely unable to generalize well. We have begun to confirm our intuition by conducting a baseline test using n-gram CLMs. We found that the baseline models performed poorly when trained on Reddit data but subjected to Twitter data for testing. We sought to improve upon this baseline by using a RFC with semi-supervised features extracted by feeding Twitter data into transformer models pre-trained on depression, gender, and age labeled datasets. While our initial results are not satisfactory, there are many untraveled avenues to explore. We cannot yet conclude that our proposed "improved" model will not perform better than those in the literature when subjected to data from external domains.

References

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

Cong, Q.; Feng, Z.; Li, F.; Xiang, Y.; Rao, G.; and Tao, C. 2018. Xa-bilstm: a deep learning approach for depression detection in imbalanced data. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1624–1627. IEEE.

Coppersmith, G.; Dredze, M.; Harman, C.; Hollingshead, K.; and Mitchell, M. 2015. CLPsych 2015 Shared Task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 31–39. Denver, Colorado: Association for Computational Linguistics.

Coppersmith, G.; Dredze, M.; and Harman, C. 2014. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, 51–60.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

Gerych, W.; Agu, E.; and Rundensteiner, E. 2019. Classifying depression in imbalanced datasets using an autoencoder-based anomaly detection approach. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, 124–127.

Hovy, D.; Mitchell, M.; and Benton, A. 2017. Multitask Learning for Mental Health Conditions with Limited Social Media Data. In *EACL*.

Kim, Y. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Potthast, M.; Rangel, F.; Tschuggnall, M.; Stamatatos, E.; Rosso, P.; and Stein, B. 2017. Overview of panâĂŹ17. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 275–290. Springer.

Vaswani, A.; Bengio, S.; Brevdo, E.; Chollet, F.; Gomez, A. N.; Gouws, S.; Jones, L.; Kaiser, L.; Kalchbrenner, N.; Parmar, N.; Sepassi, R.; Shazeer, N.; and Uszkoreit, J. 2018. Tensor2tensor for neural machine translation. *CoRR* abs/1803.07416.

Yates, A.; Cohan, A.; and Goharian, N. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2958–2968. Association for Computational Linguistics.