Early Detection of Mental Health Conditions from Social Media

Final Project Report: 17-737 Artificial Intelligence for Social Good

Alina Rath {arath}, Ashutosh Baghel {abaghel}, Yue Zhang {yuezhan5}
Instructor: Professor Fei Fang, Teaching Assistant: Ryan Shi
Carnegie Mellon University, Pittsburgh, PA

https://github.com/ashutoshbaghel/emnlp17-depression

Abstract

Mental Health is a growing global concern, and depression affects more than 264 million people worldwide [WHO Report, 2020]. Depressed individuals commonly turn to social media and other online forums for support before medical professionals. Many prior works have attempted predicting mental health disorders through machine learning models on social media posts of users. However, these works have used extensive feature engineering which make it less generalizable. In this paper, we expand upon the more recent works that employ neural networks without any feature engineering to make predictions if a user is depressed. We evaluate multiple model architectures and experiment with newer advances in Natural Language Processing using Neural Networks. We conclude on which methods work best at scale when tested on large datasets of Reddit user posts.

1. Background

Ensuring healthy lives and promoting well-being is one of the United Nations sustainable development goals. By 2030, the United Nations targets to reduce by one third premature mortality from non-communicable diseases through prevention and treatment and promote mental health.[17] Contrary to other health conditions, mental health problems are overlooked and under-addressed. People often neglect to detect early signs of mental health issues. Depression is an example of one such mental health disorder and more than 264 million people of all ages suffer from depression as per the WHO Report, 2020. In today's society, social media has become an important means to communicate and spread information. As of 2015, over 3.2 billion people (which is half the world's population) were using the internet.[18] Online forums can help reduce the stigma associated with mental health disorders and can help people find support groups. However, some of these people may still require professional medical attention. In this paper, we aim to analyze Reddit data for early detection of depression and warning signs, which could help the identification of individuals suffering from mental illness before they require help from healthcare professionals.

2. Related Works

A lot of work has been done around the mental health of social media users by analyzing their language usage on these platforms (Resnik et al., 2013 [1]; De Choudhury et al., 2013 [2]; Coppersmith et al., 2014a [3], 2014b [4]; Mitchell et al., 2015 [6]; Tsugawa et al., 2015 [7]; Coppersmith et al. [5], 2015a; Althoff et al., 2016 [8]; Mowery et al., 2016 [9]; Benton et al., 2017b [10]; Yates et al., 2017 [11]). These works have utilized data from platforms such as Twitter, Reddit and Facebook.

Many machine learning and Natural Language Processing techniques have been used previously. Prior researches have explored SVM (De Choudhury et al., 2013 [2]; Tsugawa et al., 2015 [7]), Logistic Regression (De Choudhury et al., 2014 [12]; Preotiuc-Pietro et al., 2015 [13]), Random Forest (Reece et al., 2016 [14]; Cacheda et al., 2019 [15]), and Naïve Bayes (Nadeem et al., 2016 [16]) and achieved effective performance. However, these works rely heavily on domain-dependent and application-specific features and use extensive feature engineering. These make the task less generalizable across platforms and does not scale well.

Some more recent works have used Deep Learning techniques for contextual representation and understanding of the texts. They do not require any feature selection and are highly generalizable. Yates et. al [11] were able to outperform previous works using Convolutional Neural Nets as feature extractors on all posts by a user before classifying them as depressed or not. Since, better techniques have emerged in Deep Learning and Natural Language Processing. Emergence of contextualized embeddings such as BERT, the rise of attention and more recently transformers, all seem to have surpassed older techniques at various tasks. We hope to experiment with these techniques and analyze if any of them can improve the accuracy of the prediction task. We hope that such models may get used for early detection of mental-health issues.

Table 1: Number of Records in the Data

	Diagnosed Users	Control Users
Training	3,070	35,753
Validation	3,070	35,746
Testing	3,070	35,775

Table 2: Augmented Data Set

	Diagnosed Users	Control Users
Training	20,000	20,000
Validation	3,070	35,746
Testing	3,070	35,775

3. Data

3.1 Dataset Used

Collection of labeled data for modeling mental disorder signs from social media posts is not a trivial task, since it requires identifying the social media users' associated mental health information. Prior studies mainly use data from two sources: self-declared mental health studies or surveys, and annotated social media posts with mental health keywords. The self-declared mental health studies or surveys typically require long and controlled data collection processes, which are more suitable for well-funded research and clinical studies. The annotated post data are generated manually from identifying posts where users discuss their experience with mental disorders.

For this study, we obtained data from the second category, where depression diagnosed users are identified for their posts containing high-precision diagnosis patterns. The Reddit Self-reported Depression Diagnosis (RSDD) dataset is made available by researchers from Georgetown University's Information Retrieval Lab. The dataset consists of three equal splits of training, validation and test data each containing 3,070 diagnosed users and control users in between 35,700 and 35,800.

This novel dataset was constructed to mimic a realistic distribution of control users versus depressed users. However, with the depressed users counting for less than 8% of the total users, this dataset is significantly imbalanced. An imbalanced dataset could hinder the performance of a classification model, because algorithms are typically designed to minimize cross entropy loss or maximize accuracy for classification tasks. When a majority vote yields a 92% accuracy rate, accuracy becomes a misleading objective. To overcome this issue, we oversampled the minority class in training data, depressed users using bootstrapping sampling with replacement. We also undersampled the control group without replacement, making sure the new training dataset's size is reasonable for memory usage. We keep the class imbalances as is in the validation and testing data, as we want to evaluate on proportions that really exist as documented below.

3.2 Data Usage Policy

Privacy is a serious concern for mental health data. For this study, Reddit posts come from a publically available channel and the user IDs are being masked and replaced with random numbers. A condition of obtaining the RSDD dataset is to sign a Data Usage Agreement satisfying the following requirements:

Researchers who wish to obtain the dataset must

1. make no attempt to contact any user in the dataset
2. make no attempt to deanonymize or learn the identity of any user in the dataset
3. make no attempt to link users in the dataset with any external information (e.g., an account on another website)
4. do not share any portion of the data, including example posts or excerpts from posts, with any other party

In addition, each of our team members completed the CITI Human Subject and Research training and gained the certification to comply with CMU's Institutional Review Board

4. Baseline Model

Our work primarily will deals with two important steps when using neural networks for natural language processing:

1. Extract meaningful feature representations (vectors) for each post made by the user.

2. Merge feature vectors from each post for the user, use a pooling operation, and run through fully connected layers to learn classification.

To establish our baseline, we have picked up two research publications that complete the above steps.

1. Kim, Y. (2014). Convolutional neural networks for sentence classification.
   EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference (https://doi.org/10.3115/v1/d14-1181)
   We have implemented this paper from scratch and achieved benchmarks results. To summarize this work, word embedding for each word is looked up in word2vec pretrained vectors. A 1-D convolution operation over word embeddings is performed followed by max-pooling. The max-pooling operation handles variable input sizes of the sentences. Finally, the output is progressed through a fully-connected layer to perform N-way classification.

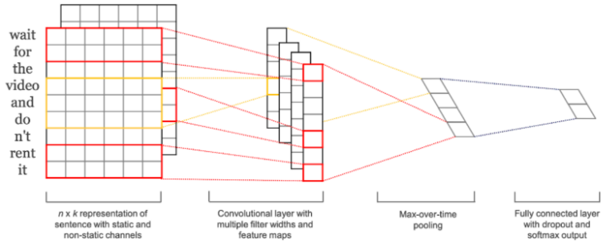
wait
for
the
video
and
do
n't
rent
it
n x k representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Figure 1: Model Architecture with 2 channels for example sentence.

2. Yates, A., Cohan, A., & Goharian, N. (2017). Depression and self-harm risk assessment in online forums. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings* (https://doi.org/10.18653/v1/d17-1322)
   In this work, vector representations are first computed similar to the work discussed above by Kim et al. Then, all such representations from each post by the user are merged to create a user representation vector. Next, these vectors are again passed through N-Dense layers with dropout before classification. This code is publicly available at: https://github.com/Georgetown-IR-Lab/emnlp17-depression

5. Architectures

In this work, we experiment with several architectural changes to the baseline model. All the experimental code is available in different branches of our code repository (https://github.com/ashutoshbaghel/emnlp17-depression).

1. Baseline: This contains our baseline run. The model contains feature extraction using CNN with a window size of 3 words. This is done in a temporal way through all posts of the user. Average pooling is done over output of each stride in each post, to form the feature vector of the post.
   Once all posts are represented as feature vectors, Yates et. al used another convolutional layer (instead of commonly used pooling) to extract a single feature vector containing important information of all posts to represent each user as a feature vector.
   This final vector is then classified using a fully-connected feedforward network. This work only uses one word window as shown in Yates et al.

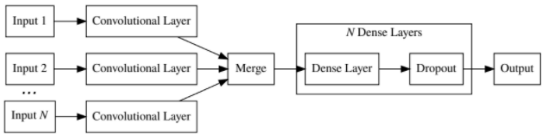

Figure 2: Architecture representing user-representation and classification.

2. Multifilter: Kim et al in 2014 had shown that instead of using just one word window, using multiple windows of varying lengths (3,4 and 5 words) results in more stable results. We extend our base code here to utilize these 3 windows of the same sizes.

3. Maxpooling: In this experiment, we replaced the second convolutional layer with a max-pooling operation. The intuition here was to find the most important post vector (which may indicate depression) instead of convolving over all posts.

4. LSTM: Lastly, we experimented with changing the model for feature extraction. As Recurrent Neural Networks are better for understanding features for words even far apart in the sentence, we used LSTM to extract the feature vector instead of CNN. The rest of the pipeline remained the same.

5. Dense-Layers: All the experiments mentioned above deal with improving feature extraction. However, we saw comparable results with all of them. The next hypothesis was to test if the classification part can be improved, that is, once the same feature vector is found, can we make better classifications. For this experiment, we started with two more dense feedforward networks. The feedforward layer dimensions that were used:

$$650 \Rightarrow 128 \Rightarrow 64 \Rightarrow 2$$

$$650 \Rightarrow 256 \Rightarrow 128 \Rightarrow 64 \Rightarrow 2$$

6. Dropout: As more dense layers come with the challenge of overfitting, we later experimented with the dropout technique, using a dropout of 0.2 after each dense layer.

$$650 \Rightarrow 128(*0.2drop) \Rightarrow 64(*0.2drop) \Rightarrow 2$$

$$650 \Rightarrow 256(*0.2drop) \Rightarrow 128(*0.2drop) \Rightarrow 64(*0.2drop) \Rightarrow 2$$

7. Attention and Transformers Model: Transformer was introduced in 2017 by Google Brain researchers to model ordered sequences of data. It is extensively used in the field of natural language processing on tasks such as machine translation and text summarization. For classifying depressed and control users, we used Transformer with a self-attention mechanism and feedforward neural network structure.
   We tried applying the algorithm to the dataset but could not fully execute the architecture given the available computational resources. Running self-attention on data of hundreds of posts per user requires high-dimensional matrix multiplications that introduce scalability issues.

8. BERT: BERT is the state-of-the-art contextualized word-embedding technique that changed how word-representations can be used. The main motivation inside BERT was to use Transformers, discussed above. We hoped to use contextualized word embeddings instead of GloVe or word2vec word embeddings to gain knowledge

Table 3: Experimental Results

Model	Class	Precision	Recall	F1
Baseline	Control	0.96	0.96	0.96
	Depressed	0.51	0.51	0.51
+Dense	Control	0.96	0.96	0.96
	Depressed*	0.52	0.51	**0.52**
++Dropout	Control	0.96	0.95	0.96
	Depressed*	0.49	0.56	**0.53**
Maxpool	Control	0.96	0.96	0.96
	Depressed	0.51	0.51	0.51
+Dense+Dropout	Control	0.96	0.93	0.95
	Depressed*	0.43	0.59	0.51
LSTM	Control	0.96	0.95	0.96
	Depressed	0.49	0.51	0.50
+Dropout	Control	0.96	0.95	0.96
	Depressed*	0.61	0.45	**0.52**
Multifilter	Control	0.96	0.95	0.96
	Depressed	0.48	0.51	0.50

from the context of their usage.

We also faced RAM depletion issues when applying BERT to the dataset. BERT uses 66 million parameters, scaling it up to hundreds of posts per user, it is a challenge to implement it with resources available to a student.

6. Experiments and Results

6.1 Evaluation Metrics

The model outputs comprise of precision, recall, and F1 score metrics for each target variable class, as well as a Receiver Operating Characteristic (ROC) curve that graphically illustrates the diagnostic ability of the classifier model. For each class,

$$Precision = TP/(TP + FP)$$
$$Recall = TP/(TP + FN)$$

F1 is the harmonic mean of precision and recall, that jointly measures the sensitivity and specificity of the classifier.

$$F\_1score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

A ROC curve is generated by plotting the true positive rate and false positive rate at various threshold settings.

For our study we jointly evaluate the models using F1 scores of the two classes and the area under the ROC curves.

6.2 Model Outputs

The table below details all the experimental results we generated. After each model architecture, we have included the results of adding more Dense layers and Dropout layer in the following rows.

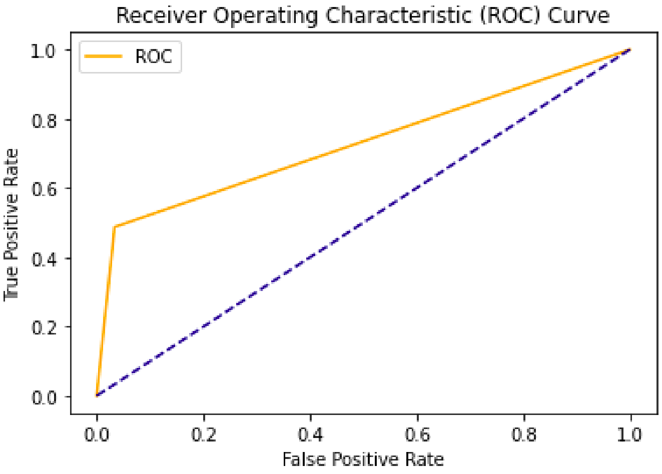

Figure 3: ROC curve of the baseline model with dense layers and dropout.

7. Results and Conclusion

In the results table above, we noted that most techniques for feature extraction (including convolution with 1 filter, with multiple filters, different pooling techniques, LSTM, etc) all perform similarly when applied to extract a user-level feature vector. The performance of the system changes minimally.

It is worth noting that adding more dense layers to the fully-connected Feedforward network boosts the classification objective. Further, adding Dropout makes the system stable. We were able to outperform the published results by Yates et al (2014) using these simple techniques. We were able to consistently reach higher F-1 scores and accuracy values as reported above.

We further explored advanced NLP systems like Transformers and BERT contextualized word-embeddings. While we were able to develop the system, these systems are not easy to scale. As this task requires feature-extraction on hundreds of posts before pooling them into one user feature-vector, we do not recommend using these heavy systems for this use-case.

8. Future Work

Future work on the topic could consider: debiasing gender stereotypes in the data and word embeddings; feature engineer with Instagram image data, labels are learned from hashtags; data augmentation and formulation of a bigger dataset, vetted by medical experts that could be used to predict stages of mental health; building a network graph based on user interactions on social media platform to analyze influence on one another for better classification; and incorporating cyberbullying analysis.

Acknowledgments

The authors thank Andrew Yates and other researchers from Georgetown University's Information Retrieval Lab for providing RSDD data. Special thanks to Professor Fei Fang from Carnegie Mellon University.

References

[1] 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.

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

[3] 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.

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

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Yates, Andrew Cohan, Arman & Goharian, Nazli. 2017. Depression and Self-Harm Risk Assessment in Online Forums. 2968-2978. 10.18653/v1/D17-1322.

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

[13] Preotiuc-Pietro D, Eichstaedt J, Park G, Sap M, Smith L, Tobolsky V, Schwartz HA, Ungar L. 2015. The role of personality, age and gender in tweeting about mental illnesses. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology. ACL; pages 21-31.

[14] Reece AG, Reagan AJ, Lix KLM, Dodds PS, Danforth CM, Langer EJ. 2016. Forecasting the Onset and Course of Mental Illness with Twitter Data. arXiv:1608.07740.

[15] Cacheda F, Fernandez D, Novoa FJ, Carneiro V. Early Detection of Depression: Social Network Analysis and Random Forest Techniques. J Med Internet Res. 2019;21(6):e12554. Published 2019 Jun 10. doi:10.2196/12554

[16] Nadeem M. 2016. Identifying Depression on Twitter. arXiv:1607. 07384.

[17] "Health - United Nations Sustainable https://www.un.org/sustainabledevelopment/health/

[18] "Internet Used by 3.2 Billion People in 2015." BBC News, BBC, 26 May 2015, www.bbc.com/news/technology-32884867.