A Transformer Based Model for Detecting Depression Among Reddit Users

Michael Knight,[1] Alex Korman,[2] Jonah Buchanan[3]
Vanderbilt University[1 2 3]
michael.a.knight@vanderbilt.edu,[1] alex.korman@vanderbilt.edu,[2] jonah.p.buchanan@vanderbilt.edu[3]

Abstract

As social media has evolved, it has created a space for communities to form where individuals can share their experiences and offer advice to others. With the spike in mental health issues related to the COVID-19 virus, these communities have become increasingly popular and more important in providing support to those that need it. We focus on identifying depressed users on social media by performing linguistic analysis on their historical posting data. The sequential nature of these posts offers insight into build ups of helpless feelings. Our model uses a LSTM to classify users as depressed or not depressed by examining a user's social media history. Our results suggest strong potential for the ability to detect at-risk users based off their social media history but little increase in efficacy with the augmentation of the historical data. We examine the potential lifesaving effects of this model and possible privacy ramifications.

Introduction

With the recent stay at home orders due to the spread of COVID-19, up to 40% of adults have reported struggling with mental health or substance abuse (Czeisler et al., 2020). Specifically, symptoms of depression and anxiety have been reported approximately three times as much as this time last year, disproportionately affecting young adults (Czeisler et al., 2020). This uptick in depression cases in particularly concerning when considering the potential loss of life to suicide. Globally, about 16,000,000 suicide attempts take place each year leading to 800,000 deaths (WHO, 2014). Even before COVID-19, there has been a concerning uptick in suicide cases, increasing by 24% in the last twenty years (Curtin et al., 2016). This has led to suicide becoming a top ten cause of death in the United States (Curtin et al., 2016). Individuals who suffer from depression often keep it hidden, with 60% of those who died by suicide denying having suicidal ideations (McHugh et al., 2019).

Reddit is a social media service that provides different specialized communities in the form of "subreddits". It averages over 52,000,000 daily users across the globe (Patel 2020). With the specialized communities and longer post formats, Reddit creates a space that allows for greater discussion and experience sharing that other forms of social media, e.g. Twitter.

Reddit users have founded a wide range of subreddits in order to address mental health issues, such as /r/depression, /r/depressed, /r/mentalhealth, /r/suicide, and so on. These have grown rapidly and quickly become a resource for people across the globe to share their struggles with mental health and for others to offer advice. Over 763,000 users subscribe just to /r/depressed and /r/depression.

Recently, researches have begun applying traditional linguistic methods in an attempt to identify these users who are at risk for depression and suicide. Recent works attempting to analyze posts purely off the linguistic characters have fallen short (Coppersmith et al., 2018). Other researchers have shown greater success by accounting for the sequential nature of these posts to model how these feelings often build up over time.

The LSTM's used in these previous models have showed some flaws still. Researchers have attempted new models that account for the inherent variance in the time between posts (LSTM assumes identical intervals). This method was implemented in the STATENet model, showing substantial improvement over traditional methods.

Contributions

We propose our model, which takes into account the emotional state of an individual over time and the linguistic features of their posts, as a new way to classify individuals on Reddit as depressed or not depressed. Our model uses a transformer combined with a dense layer to learn and differentiate the differences in language between depressed and non-depressed users. In the test environment, our model has shown high levels of accuracy (>90%). This level of

accuracy is comparable with other models in this domain. We later discuss the practicality and ethical implications of this model.

In practice, this model would work within a human-in-the-loop system to help identify at risk individuals. With the current status of the model and considering the failures of past risk detection systems, it is important that the actions based off it are non-intrusive.

Related Work

In the past, mental health professionals have used primarily questionnaires or interviews in order to evaluate an individual's suicidal ideation or depression levels. Tests such as the Depression Anxiety Stress Scale (Coppersmith et at., 2018) or the Adult Suicidal Ideation Questionnaire (Fu et al., 2007) have been the go-to for mental health practitioners studying suicide and depression. These methods, while valuable for gathering lab data, are rather poor for identifying suicidal individuals in practice. Often times, individuals who have suicidal ideation are either unable or unmotivated to access mental health resources leading to them not getting the appropriate help that they need (Zachrisson et al., 2006). In fact, up to 80% of individuals suffering from suicidal ideation have not undergone psychiatric treatment, signaling that there is a vast number of individuals who could benefit from these systems (WHO, 2014). The intrusive manner of these surveys oftentimes can increase the depressive symptoms of the individuals taking them (Harris and Goh, 2016). For these reasons, researchers are looking for less intrusive methods that can passively detect suicidal ideation and alert humans that can provide resources to those that are suffering.

As machine learning methods have evolved, researches have begun to apply them to social media in order to passively detect suicidal ideation and depression. Only recently was real time monitory of social media determined to be a possible method of detecting depression (Jashinsky et al., 2014). And even more recently it was determined that machine learning algorithms are efficient in differentiating between those who are suicidal and those who are not (Braith-waite et al., 2016). This indicates to us that it would be possible to apply similar classification methods to detect depression. Classifying users with suicidal ideation or depression has been shown to have success with both binary (Sawhney et al., 2020) or multiple levels (Zirikly et al., 2019) of depression. Additionally, researches have had success classifying individuals with CNNs, LSTMs, RNNs, and LSTMs augmented with attention mechanisms (Zirikly et al., 2019).

While these models have shown success to some extent, researchers have hoped to improve them by using contextual methods. Instead of just analyzing one post, they use the previous posts by the user to see if they can provide emotional context to the most recent post (Zirikly et al., 2019). Though, in these traditional contextual methods it is often assumed by the model that the posts have equally spaced time periods between them. Sawhney et al. (2020) have used time aware LSTMs in order to use the historical posts to model the emotional state of the user and provide greater context to their most recent posts. This significantly improve performance compared to its competitors.

Methods

Notation

We consider modeling depression as a binary classification task. We assess the presence of depressive thoughts and experiences through a reddit user's historical posts and comments. The idea is that certain language, post patterns, or content can infer affect. A reddit post is denoted by $r_i \in R = \{r_1, r_2, \ldots, r_N\}$, authored by user $u_j \in U = \{u_1, u_2, \ldots, u_N\}$. Each user, u_j, has a history, $H_{i,j} = \{(h_1^i, t_1^i), (h_2^i, t_2^i), \ldots, (h_D^i, t_D^i)\}$, consisting of tweets h_k^i, that occur sequentially at time t_k^i, where $t_1^i < t_2^i < \ldots < t_D^i$. t_D^i is the time at which the user first posts in a r/depression channel(for our purposes these are treated as uniform intervals of time), which we use to indicate depression. The problem is formulated as a classification task to predict label y_i(the depressed state of the user), for the reddit history $H_{i,j}$.

Figure 1: A User's Reddit Submission History

Embedding

Multiple studies have been conducted demonstrating users' linguistic social media patterns can inform their mental state (De Choudhury et al., 2013). Word embeddings are largely

accepted as the standard for encoding text in social media (Shen et al., 2017; Sawhney et al., 2020). We implemented the base BERT (Bidirectional Encoder Representations from Transformers) from the Huggingface library to encode each reddit post. Where

$$R'_i = BertBase(r_i)$$

produces a [CLS] vector $R'_i \in R^{768}$, the embedded representation of the entire post.

Sequential Modeling

We take advantage of the emotional context provided by users' reddit history to model their developing mental state (Abdul-Mageed and Ungar, 2017). We employ an LSTM to learn the user's history over a sequence of posts. In doing so, the LSTM produces a singular representation for the entire embedded history, $H'_{i,j}$. A sigmoid activation function is applied to the representation, and then is fed to a dense layer with a dropout. We used Adam to optimize for the binary cross entropy during back propagation. (Fig. 2)

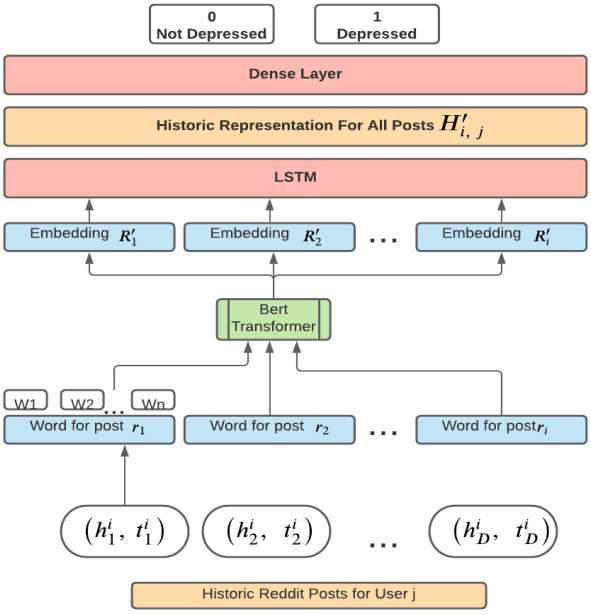

Figure 2: Model Architecture

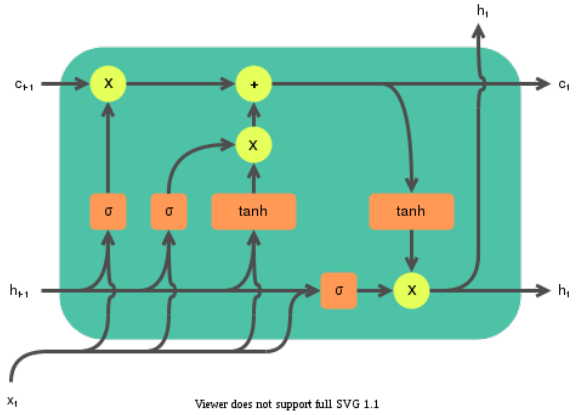

Figure 3: LSTM Cell by Guillaume Chevalier CC BY 2.0

Experiments

Data Collection

The reddit posts used for the model were collected from two different sources. A subset of the users provided by Dr. Derr, those who had posted in /r/depressed, were selected as a starting point. These were users who posted or commented on /r/depressed from June to October of 2020.

Additionally, using the Python Reddit API Wrapper (PRAW), users who had most recently posted in /r/depression were selected. Once these users were selected, all users who had less than 50% of their posts outside of depression adjacent subreddits were eliminated. This was done in order to only capture users who used the full scope of Reddit as a social media site and not just the depressive communities.

After selecting the users classified as depressed, we used PRAW to find the first time a user posted or commented in a depression subreddit. This post was considered their most recent submission. We then collected up to thirty posts or comments proceeding their most recent submission.

In order to collect users who, we classified as not depressed, we used PRAW to gather several thousand users who had most recently posted on Reddit. We then gathered the users' thirty most recent posts or comments.

Dataset

Once all data was collected it was cleaned in order to make it more uniform for the model. All posts and comments were cleaned of URL's, emojis, and non-supported characters.

Additionally, all comments or posts with less than ten words were eliminated. We then eliminated users with less than ten posts or comments remaining.

The final dataset consisted of about 5000 users. Users were annotated as follows:

Method	Model	Precision	Recall	F1	Accuracy
Baseline	2 Layer LSTM	91%	91%	91%	94%
	1 Layer LSTM	**92%**	**93%**	**93%**	**95%**
Historical	2 Layer LSTM	88%	83%	85%	92%
	1 Layer LSTM	89%	85%	87%	92%
Recent Post	2 Layer LSTM	91%	91%	91%	95%
	1 Layer LSTM	91%	92%	91%	95%

Table 1: Mean result of each model after being run ten times. Results were average after multiple runs. **Bold** indicates top performing method.8

- **Depressed:** These users had posted in depression adjacent subreddits.
- **Not Depressed:** These users had not posted in depression adjacent subreddits and were interacting with Reddit in other communities.

Of which about 81% were classified as depressed and 19% were classified as non-depressed. The average word count of the posts and comments was 233.

Experimental Settings

- **Baseline Method:** We trained and evaluated our model using both the historical context of the posts and the depressive posts (most recent post) for depressed users, tDi. This method allowed us to use a post we believed would be rich in information because it was created in a depressive context, and likely used emotive terminology. This is a strong representation of how the model would most likely be used in praxis on an actual social media account, using historical representations to flag a depressed post.

- **Historical Method:** This method trains and tests the model using purely the historical posts. There are no posts from depression adjacent subreddits included in the training of this model. This means we are hoping to pick up on possible linguistic characteristics of user posts outside of depression subreddits that may indicate that they are depressed.

- **Recent Post Method:** This method trains and tests the model only on the most recent post of each of the users. This would be the depressed post for the depressed users. Here we hope to be able to compare if there is a benefit of adding the historical data when assessing if user is depressed.

Experimental Setup

We select hyperparameters based on the highest Macro F1 obtained on the validation set for our model. These parameters were explored empirically. They were: the number of features in hidden state $H\sim \in \{8, 64, 128, 256, 400\}$, number of LSTM layers $n \in \{1,2,5\}$, dropout $\sigma \in \{0.0, 0.1, 0.2, 0.3, 0.4\}$, learning rate $l_r \in \{0.01, 0.001, 0.0001, 0.00001\}$. The optimal parameters found to be $H\sim = 400$, $n= 1$, $\sigma = 0.2$, $l_r = 0.001$. The model is trained in 40 epochs, with a batch size of 120.

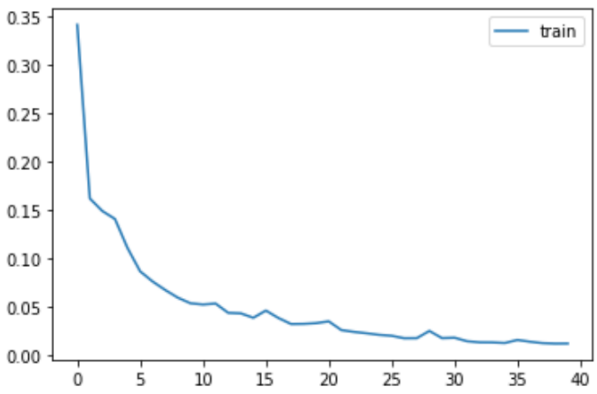

Figure 4: Training Loss for Baseline Method with 2 LSTM Layers

Results and Analysis

Performance

In Table 1 we see the results of the generated models. The statistics shown are the results obtained after running each model 10 times and averaging the results. Overall the models performed surprisingly well with the top performing model returning an accuracy of 95%.

Fig. 5 displays a confusion matrix for 1 run of the baseline method with 2 LSTM layers. Besides the statistics seen in Table 1, one important thing to note is that 66 users were labeled as not depressed when they were depressed. This is only a 1% false negative rate but for this work that is too high. Any false negative labels could lead to people who need help not receiving the resources they need. That could lead to those people potentially committing suicide. One life lost is one too many. Thus, although our model works well, it should not be used as the only mechanism to identify depression. Depression is complex and our model cannot possibly encapsulate all facets of it. In future models, it is better to be safe than sorry and marking not depressed users as depressed has much smaller impact than the reverse.

When comparing the baseline method to the historical method we do see improvements, but they are small, mainly because the accuracy of the historical method is still relatively high. The reasons for the historical method having such high accuracy could be due to linguistic cues that depressed users use in non-depression adjacent subreddit. But it is also possible, due to the ever-increasing size of Reddit, that there are communities which could be depression adjacent that we could not identify.

When comparing the baseline method to the recent post method, there is negligible increase in accuracy. This indicates to us that it is most likely most of the predictive power of this model comes from the most recent post which is the depressed post for depressed users. This is not completely out of line with peer models, considering Sawhney et al.(2020) saw marginal but signifying differences in accuracy when incorporating tweet context.

Limitations

In Table 1 we see the models developed. The first key thing to note is that both models perform extremely well, even better than STATENet which had an accuracy of 85% of detecting suicidal intent of Twitter Users (Sawhney et al., 2020). Our model is not as comprehensive as STATENet, so it is not probable that it performs better even though our work runs on Reddit rather than Twitter. It is likely that the inherent shortness of Tweets makes the STATENet problem a more challenging one. The cause of our model's high performance is most likely due to the imbalanced dataset. 81% of the data collected represented Depressed users while 19% represented Non-Depressed users. Performing a binary classification on an imbalanced dataset could lead to a high accuracy for the model which is seen in both of our top performing models. Thus, we are able to identify depression in Reddit user's, but we need more experimentation to validate the performance of our models.

Studying depression is inherently subjective and, at least with current methods, there is no catch all to represent the experiences of unique individuals. Since we pulled data from specific Reddit threads, the studied data may be susceptible to demographic and medium specific biases. More specifically, Reddit is an English-speaking platform and likely lacks the ability to provide context for users outside of the anglosphere. We recognize that depression is complex and could be significantly different for different people. Our simplification of depression to a binary classification is not meant to downplay the varying levels of the illness, but instead try to capture it all levels.

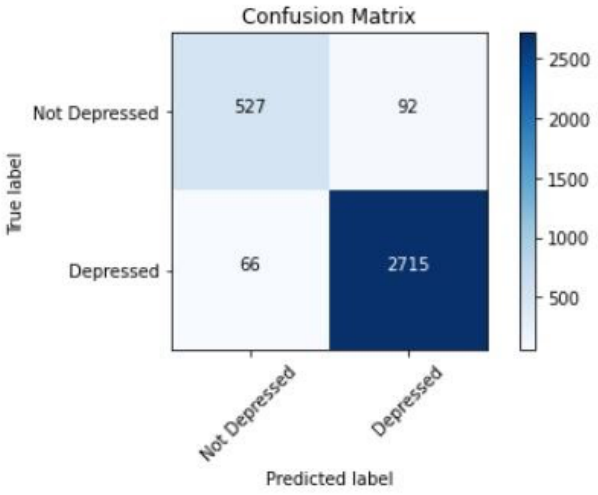

Figure 5: Confusion Matrix for Baseline Method

Practical Implications

Our model provides a way to preliminarily screen individuals that are at risk for depression in a non-intrusive nature. Clinical resources can be allocated to users who are identified as being at risk of depression. These resources can help individuals and potentially save lives. Also, the anonymity of each user is important. All personal information is stripped during our data collection. Finally, due to the limitations previously stated, our software should only be used as a preliminary screen for depression and not as a definitive measure of depression.

Discussion

Ethical Considerations

When dealing with user data, ethics is important to consider. Facebook addressed ethics when looking into suicide prevention on their site: "How can we deploy a suicide prevention system that is effective, and that protects people's privacy, i.e., that is not intrusive and respectful of people's privacy expectations?" (Gomes 2020). We addressed some issues regarding protecting user privacy, but we did not consider many things that could be incorporated in future work. One thing that we addressed is user anonymity. After we pull the data, embed the test, and divide the data into the training, validation, and test sets, there is no way to trace a user to their posts because the data was randomly divided, and the usernames were taken off of the data. Some other things we could look into regarding anonymity is to remove any emails or names in the posts.

One thing that Facebook considered that we did not was the privacy settings for a user. We just pulled all of the data from Reddit threads regardless if a user allowed it or not. In the future we could do a check of user privacy settings to ensure that we are allowed to work with their information. This could result in a much smaller dataset and thus a less effective model. But it is important to respect user privacy over efficacy.

Most importantly, we do not wish for these techniques to ever be used in an intrusive or non-private manner. Samaritan Radar provided an example to all future researchers in this domain that intrusive and non-private methods can have dire consequences and the implications of all work should be heavily weighed before implementation.

Another thing that Facebook we considered is if this process should be fully automated. They pondered the idea of using humans to review the reports generated from their software. That would require using more of their resources and money to manually check their software which has a pretty high accuracy rate. But they realize the devastating impact that a false negative could have on people. Thus, they chose to use humans to review users who had a low probability of committing suicide to ensure that their model did not produce a false positive. Human eyes could discern something that the code might misinterpret. We could consider this in our work to manually check users who are labeled as not depressed. This might require a lot of manpower so in the future we could consider giving a user a probability score that represents how likely they are to be depressed. Then we could have people go through users who are labeled between a certain percentage to double check if they are depressed or not.

The final thing that Facebook considered is target-ness vs thoroughful-ness. Target-ness looks at how they can focus resources to people who need it and thoroughly-ness looks at how many users can be looked at. In a perfect world, all users would be examined, and the appropriate people will receive the resources they need. But time and money limit the number of users we can process in a day. In order to address this challenge, they increased the number of human reviewers they used while maintaining a minimum threshold accuracy. We could consider this issue in our work manually reviewing certain users in such a way that we maximize our users and they are not just looking through countless people who are not at risk.

Conclusion

We have demonstrated a successful model for classifying users on Reddit as depressed or non-depressed based off of their historical posting content. While it provides marginal gains, we did not see the significant improvement from adding historical context that we had hoped to see. With Reddit's unique format of primarily being long discussion based likely provided almost all of the information to classify a user as depressed or not in a single post. This does not mean that there is not potential for more gains from historical context.

The Reddit experience, while a challenge for this project is extremely beneficial to depressed users. By having these communities built into the framework of the website it provides a great way for depressed individuals to access resources for treating their depression and to easily connect with others that are struggling alongside them.

We believe that Reddit also provides an easy way to disseminate resources to those that need it. The social media platform often recommends posts or communities to users that they do not already subscribe to. If Reddit implemented a detection system in order to identify users who are likely to be depressed, attempting to get them help could be as simple as recommending depression adjacent subreddits or posts where people are experiencing the same struggles as the user.

We hope that these results can be used as basis for future works to further explore the possibility of early detection of depressed social media users.

With that in mind, there are many directions that future researches could improve upon this work. Primarily, a time-aware analysis of these posts has the potential to offer substantiative increases in the performance of the model. While these models seem to have enough accuracy to be implemented, it does not seem that the problem of actually using them in practice has been solved. The question of how to non-intrusively help depressed individuals identified by these models has not yet been answered. We hope that future research can find a sufficient human-in-the-loop system that will allow for the individuals identified here to properly receive the help they need.

References

Braith-waite, S; Giraud-Carrier, C; West, J; Barnes, M; and Hanson, CL. 2016. *Validating machine learning algorithms for twitter data against established measures of suicidality.* JMIR Mental Health, 3(2):e21.

Crawford, J. R., & Henry, J. D. (2003). *The Depression Anxiety Stress Scales (DASS): Normative data and latent structure in a large non-clinical sample.* British Journal of Clinical Psychology*, 42*(2), 111–131.

Coppersmith G, Leary R, Crutchley P, Fine A. *Natural Language Processing of Social Media as Screening for Suicide Risk*. Biomedical Informatics Insights. January 2018. doi:10.1177/1178222618792860

Curtin, SC, Warner, M, Hedegaard, H. *Increase in suicide in the United States, 1999-2014.* NCHS Data Brief. 2016;241:1–8.

Czeisler MÉ , Lane RI, Petrosky E, et al. *Mental Health, Substance Use, and Suicidal Ideation During the COVID-19 Pandemic — United States*, June 24–30, 2020. MMWR Morb Mortal Wkly Rep 2020;69:1049–1057.

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.

Fu KW, Liu KY, Yip PS. *Predictive validity of the Chinese version of the Adult Suicidal Ideation Questionnaire: psychometric properties and its short version.* Psychol Assess. 2007 Dec;19(4):422-9.

Gomes de Andrade, N., Pawson, D., Muriello, D. et al 2020. *Ethics and Artificial Intelligence: Suicide Prevention on Facebook*

Harris, K and Goh, MT. 2016. *Is suicide assessment harmful to participants? Findings from a randomized controlled trial*. International Journal of Mental Health Nursing, 26(2):181–190.

Jashinsky J, Burton SH, Hanson CL, West J, Giraud-Carrier C, Barnes MD, Argyle T. *Tracking suicide risk factors through Twitter in the US. Crisis*. 2014;35(1):51-9.

McHugh CM, Corderoy A, Ryan CJ, Hickie IB, Large MM. *Association between suicidal ideation and suicide: meta-analyses of odds ratios, sensitivity, specificity and positive predictive value*. BJPsych Open. 2019.

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: *Fine-grained emotion detection with gated recurrent neural networks.* In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Patel, Sahil. “Reddit Claims 52 Million Daily Users, Revealing a Key Figure for Social-Media Platforms.” *The Wall Street Journal*, Dow Jones & Company, 1 Dec. 2020.

Sawhney, R.; Gandhi, S.; Joshi, H.; Shah, R.R. 2020. *A Time-Aware Transformer Based Model for Suicide Ideation Detection on Social Media*

WHO. 2014. *Preventing suicide: A global imperative.* World Health Organization.

Zachrisson, H.D., Rödje, K. & Mykletun, A. *Utilization of health services in relation to mental health problems in adolescents: A population based survey.* BMC Public Health **6,** 34 (2006).

Zirikly, A; Resnik, P; Uzuner, O; and Hollingshead, K. 2019. *CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts.* In Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.