CLPsych2019 Shared Task: Predicting Users' Suicide Risk Levels from Their Reddit Posts on Multiple Forums

Victor Ruiz[a*], Lingyun Shi[a*], Jorge Guerra[a,b], Wei Quan[a,c], Neal Ryan[d], Candice Biernesser[d], David Brent[d], and Fuchiang Tsui[a-c]

[a]*Tsui Laboratory, Children's Hospital of Philadelphia, Philadelphia, PA, USA,*
[b]*Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, USA,*
[c]*Drexel University, Philadelphia, PA, USA,*
[d]*Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA*
**: Authors contributed equally*

Abstract

We aimed to predict an individual suicide risk level from longitudinal posts on Reddit discussion forums. Through participating in a shared task competition hosted by CLPsych2019, we received two annotated datasets: a training dataset with 496 users (31,553 posts) and a test dataset with 125 users (9610 posts). We submitted results from our three best-performing machine-learning models: SVM, Naïve Bayes, and an ensemble model. Each model provided a user's suicide risk level in four categories, i.e., no risk, low risk, moderate risk, and severe risk. Among the three models, the ensemble model had the best macro-averaged F1 score 0.379 when tested on the holdout test dataset. The NB model had the best performance in two additional binary-classification tasks, i.e., no risk vs. flagged risk (any risk level other than no risk) with F1 score 0.836 and no or low risk vs. urgent risk (moderate or severe risk) with F1 score 0.736. We conclude that the NB model may serve as a tool for identifying users with flagged or urgent suicide risk based on longitudinal posts on Reddit discussion forums.

Keywords: suicide, Reddit, machine learning, predictive modeling

I. Introduction

Suicide poses a challenge to our society. It is the 10th leading cause of death in the United States for all ages, and most importantly it is the second leading cause of death for 64 millions of youths between the ages of 10 and 24.(NIMH, 2018)(Howden and Meyer, 2011) Meanwhile, the use of social media among the young population is getting more poupular.

Social media websites such as Reddit discussion forums serve as a common platform for people to express their thoughts, and many people feel more comfortable discussing or sharing their mental state including suicidal thoughts on social media than they are in person. Moreover, people who can get access to the internet may not have adequate resources for mental health care. In contrast to the electronic health records that recorded the interactions between patients and clinical care providers, on-line social media posts illustrate conversations between a user and an online audience mostly comprised of non-clinicians. In March 2019, Reddit was estimated to have 542 million monthly visitors and 234 million unique users, 53.9% of which with bases in the United States.(Wikipedia,) There is a need to study potential suicide risks based on social media posts as a part of public health surveillance.(De Choudhury et al., 2017)

Current state-of-art approaches for mental health condition prediction leveraged machine learning (ML) and natural language processing (NLP). Common ML algorithms include support vector machines, Naïve Bayes, etc. NLP techniques include part of speech, bag-of-words modeling, word embeddings, etc. The performance of those models measured by micro-averaged F1 score ranged between 0.4 and 0.76,(Calvo et al., 2017) and by macro-averaged F1 score ranged between 0.5 and 0.84.(Shing et al., 2018) A macro-averaged score computes the metric independently for each risk level (class) and then takes the average across all levels regardless of the number of samples in each risk-level group, whereas micro-average treats each post equally regardless of class. Thus, a macro-averaged score carries more per-post weight for those risk levels (categories) with fewer posts.

In this study, we hypothesized that we can develop advanced data-driven predictive models that can predict individual suicide risk level from longitudinal posts on Reddit discussion forums.

Our study has three key contributions. First, we developed 10 feature domains based on NLP and feature engineering, described in Section II.2, including clinical findings and semantic role labeling (those were not commonly included in previous shared tasks competition for social media data(Shing et al., 2018)) for the prediction of suicide risk from Reddit posts. Second, we developed several state-of-the-art machine learning models including deep neural network models for the prediction task. Third, we developed a modeling strategy for improving prediction accuracy.

II. Methods

This section describes study datasets, text preprocessing, feature engineering, predictive modeling, and evaluation metrics.

II.1 Datasets

We received two datasets from the CLPsych2019(Zirikly et al., 2019): 1) a training dataset and 2) a test dataset. Both datasets comprised annotated posts on the Reddit discussion forum and its sub-discussion forms, also known

Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pages 162–167,
Minneapolis, Minnesota, June 6, 2019. ©2019 Association for Computational Linguistics

as *subreddits*. The training dataset study period is between 2005 and 2015, comprising 31,553 posts from 496 Reddit users with the cohort definition: a user had at least one post on the *SuicideWatch* subreddit; users who posted on the SuicideWatch may not be of risk to suicide. The data elements in the training dataset included a user id, a subreddit name, a post title and body from the user's posts in any subreddit, and post timestamp in a unified time zone. The CLPsych2019 organization provided the gold standard for the training dataset.(Shing et al., 2018; Zirikly et al., 2019) Following the same cohort definition, the test dataset comprised 9,610 posts from 125 Reddit users. We received the training and test datasets one month and five days before the competition deadline, respectively.

The study is approved under the Children's Hospital of Philadelphia IRB.

II.2 Natural Language processing and Feature Engineering

II.2.1. Text preprocessing

We performed a series of preprocessing pipeline including sentence splitting, tokenization, removal of stop words, part of speech tagging, and lemmatization.(Posada et al., 2017)

II.2.2 Feature domains from users' posts

Similar to the work by Shing et. al.(Shing et al., 2018), we developed the following feature domains:

Clinical findings: A social media post may contain clinical findings such as depression, schizophrenia, cancer, etc. We utilized the clinical Text Analysis and Knowledge Extraction System (cTakes)(Savova et al., 2010) developed by the Mayo clinic, to extract clinical findings from each post. cTAKES extracts each finding with a Concept Unique Identifier (CUI) represented in the standard Unified Medical Language System (UMLS) developed by the National Library of Medicine (NLM). We also flagged suicide attempt related CUIs (SA CUIs) using a pre-defined CUI list from our previous suicide attempt study with electronic health records (EHR).(Tsui et al., 2019)

Social determinants of health (SDOH): We classified each sentence into one or more of the 11 social categories that we previously developed.(Quan et al., 2019; Liu et al., 2019) The 11 categories included: 1) social environment, 2) education, 3) occupation, 4) housing, 5) economic, 6) health care, 7) interaction with legal system, 8) social support circumstances and social network, 9) transportation, 10) spirituality and 11) other (e.g., exposure to disaster, war, other hostilities, and access to weapons, etc.).

Emotion and health-disorder association: We identified posts' lemmas that matched terms in the Word-Emotion Association Lexicon developed by Mohammad et. al.(Mohammad and Turney, 2013), as well as a lexicon compiled from terms available in the list of psychological disorders(,). We identified words in a post associated with emotion categories, e.g., joy, sadness, fear, etc.

Readability score: Readability score provides a gauge for the level of understanding of a document. We used spaCy library to calculate 7 readability scores for each post: (1) automated readability index, (2) Coleman-Liau index, (3) Dale-Chall index, (4) Flesch-Kincaid grade level, (5) Flesch-Kincaid reading ease index, (6) forecast index and (7) smog index.

Semantic role labeling (SRL): SRL is a linguistic process that identifies semantic roles, e.g., subject, object and verb, of a sentence. We used two latest state-of-the-art statistical SRL models: Bidirectional Long Short-Term Memory (BiLSTM) model(He et al., 2017) and the Embeddings from Language Models (ELMo)(Peters et al., 2018), which provides deep contextualized words representations, to identify the semantic role labels and predicate-argument structure from each sentence in a user's post. The identified predicate-argument information indicates detailed semantic structure and roles, i.e., "who" did "what" to "whom" at "where" and "when". Table 1 shows an example. SRL plays a critical role for revealing self-referential thinking.

Table 1. Semantics analysis of a sample sentence from a Reddit forum. The right column in the table demonstrates the identified argument labels (subject and object labels), predicate and negation labels from the sentence on the left column after applying SRL process; the *arg0* tag, the *arg1* tag, and *argm-negation* tag represent the subject "I", the object "the loneliness and pain", and the sentence negation, respectively.

Sentence in a post	Predicate-argument structure
"I can't handle the loneliness and pain anymore."	"arg0": " I", "argm-mod": " ca", "argm-negation": " n't", "predicate": " handle", "arg1": "the loneliness and pain"

Sentiment levels: A sentiment level provides a gauge for the level of sentiment of a sentence. We used Stanford CoreNLP(Manning et al., 2014) to identify 5 sentiments: "*Very Negative*", "*Negative*", "*Neutral*", "*Positive*", "*Very Positive*" for each post. To create the features, per user, we calculated the following averages: 1) micro average: the sum of all sentiments across all the post of a user divided it by the total number of sentences across those post per that user; 2) macro average: the sum of each post level sentiment vector of a user divided by the total number of post by that use; 3) post-level vector: the sum of all sentiment vectors in a post divided by the total number of the sentence in that post.

Topic modeling: Topic modeling provides an unsupervised-based learning to map each post into a predefined number of topics. We used the unsupervised learning Latent Dirichlet Allocation (LDA) to identify 10, 20 and 30 topics from all the posts.

Empathy topics: We used Empathy text analysis tool to identify 196 pre-defined topics(Fast et al., 2016) from each of the posts, e.g., death, negative emotion, sadness, etc. Each post has an empathy vector, E_i^{196x1}, where *i* represents a post, and each topic, $e_{i,j} \in Z, [0,100]$.

Doc2Vec model: We built a Doc2Vec model via distributed bag of words (DBOW) based on the training Reddit posts, and represented each Reddit post as a 300x1vector.

Aggregate Statistics (AS): We created summary statistics features that characterize users' posting habits. Table 2 summarizes the list.

Table 2. Aggregate statistics based on feature domains

Feature Domain	Statistics at the post and user levels
Clinical Finding	• Individual CUI counts from all posts • Average count of each CUI per post • Average count of each CUI per CUI-post (CUI-post refers to the post with at least one identified CUI) • Total count of distinct CUIs from all posts • Total count of SA CUIs per user • Total count of SA CUI-posts per user (SA CUI-post refers to the post with at least one identified SA CUI) • Total count of distinct SA CUIs per user
Semantic Role Labeling (SRL)	• Average count of each *arg0* and *arg1* per post • Minimum/Maximum counts of each *arg0* and *arg1* in one post • Average count of "negative"-*arg0* per post (An "negative"-*arg0* refers to the arg0 with an *argm-negation* modifier for the predicate as shown in Table 1) • Minimum/Maximum count of each "negative"-Arg0 in one post • Count of distinct *arg0* and distinct *arg1* values per user • Minimum/Maximum count of distinct *arg0* and *arg1* values in one post • Average number of part-of-speech tags (nouns, verbs, adjectives, adverbs, etc.) in the last two years
SDOH	• Total number and percentage of sentences in each social determinants of health category
Forum Posting Behavior (FPB)	• Number of total posts for the user in all subreddits • Number of total posts for the user in in the last two years • Number of weeks with posts to the *SuicideWatch* subreddit • Number of active days between the first and last posts • Average post time difference between 2am (EST) and the post time in the last two years • Average length (characters) of posts in the last two years • Days since last post to the *SuicideWatch* subreddit • proportion of the user's posts containing the word 'edit' in the last two years • Proportion of posts made between 2a and 6m EST • proportion of posts made during weekends (Saturday and Sunday) in the last two years • Maximum number of consecutive weeks in which users' made posts to *SuicideWatc*h in the last two years. • All subreddits that the user posted to in the last two years • Number of posts to *SuicideWatch* by week in the last two years (1x104 vector) • Number of posts made by users to *SuicideWatch* in the last two years.
Sentiment	• Proportions of sentiment score at post and sentence levels
Readability	• Averages of 7 readability at post and sentence levels
Emotion	• Average count of each emotion-related term across all posts
Topic modeling	• Average count of each topic across all posts

II.3 Predictive modeling and evaluation

We developed seven machine learning models: Naïve Bayes (NB), gradient boosting (GB), random forest (RF), support vector machine (SVM), and deep neural networks including augmented convolution neural networks (CNN) and long short-term memory neural networks (LSTM). Unlike conventional deep neural networks, we developed augmented deep neural networks included input not only from freetext posts (Doc2Vec) but also the user-level aggregate statistics defined in Section II.2.

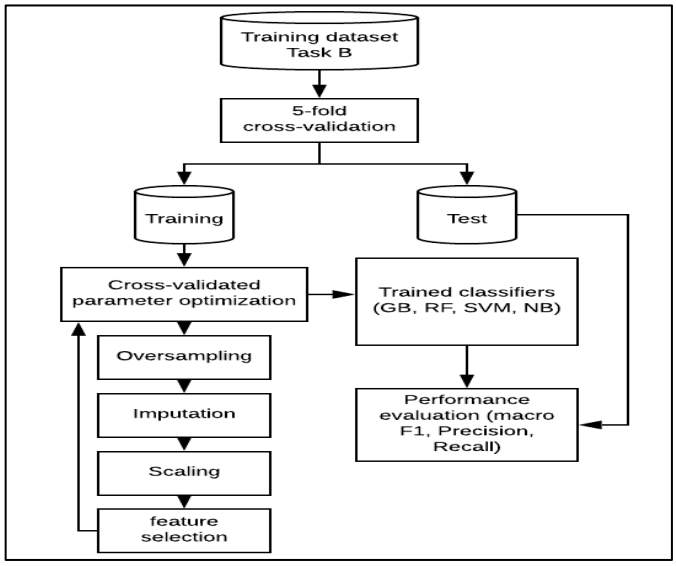

Figure 1. Experimental design and modeling process.

Figure 1 outlines our modeling and evaluation process. We trained and optimized models in a nested 5-fold cross-validation approach, and each model was optimized based on macro-averaged F1 score, which is the main measure for ranking models developed by the shared-task participants. First, we oversampled the sparsely-represented classes to alleviate the existing class imbalance.(Chawla et al., 2002) Then, we conducted imputed missing values with variable means, and either did not scale variables, or scaled values $\in R$ to [0, 1]. Then we performed a two-phase feature selection process. First, we applied a correlation-based feature-selection filter(Hall, 1999), and then conducted a forward greedy search over an increasing number of features selected based on information gain feature ranking.(Tsui et al., 2017)

For the competition, each team was limited to submit up to three models' results, we chose top two models and added an ensemble model based on our three best-performing models. We used the 5-fold average of macro-averaged F1 scores to evaluate each model. The models used to submit results to the competition were re-trained with the full training dataset following the same approach used during cross-validation.

We created an ensemble classifier from our best 3 performing models. Predictions from these models were

used to tally votes and generate the final predictions of the ensemble classifier. Since there were more risk categories (4) than the number of classifiers (3) in the ensemble, it is possible that all models produce different predictions. In this scenario, we created a rule by favoring the classes that were likely to be misclassified.

Besides macro-averaged F1, our evaluation metrics include macro-averaged accuracy, precision and recall. We further compared the performance based on binary classifications, i.e., flagged risk (low, moderate, and severe risks) vs. no risk, and urgent risk (moderate, and severe risks) vs. others.

III. Results

Table 3. Risk level distributions in two datasets.

	Training Dataset	Test Dataset
No risk	127 (25.6%)	32 (25.6%)
low risk	50 (10.08%)	13 (10.4%)
moderate risk	113 (22.78%)	28 (22.4%)
Severe risk	206 (41.53%)	52 (41.6%)
Number Subreddits covered	3662	1593

Table 3 shows the distributions of users in 4 different risk categories in the training and test datasets. Both datasets have low counts in the low risk level and share almost the same distribution.

Table 4. Average 5-fold predictive model performance from the training dataset, measured by the macro-averaged F1 score followed by the number of variables (features) used by a model in parentheses.

	NB	GB	RF	SVM	CNN	LSTM
Marco-F1 score	0.422	0.412	0.395	**0.432**	0.367	0.147
# of variables	75	100	100	100	796	796

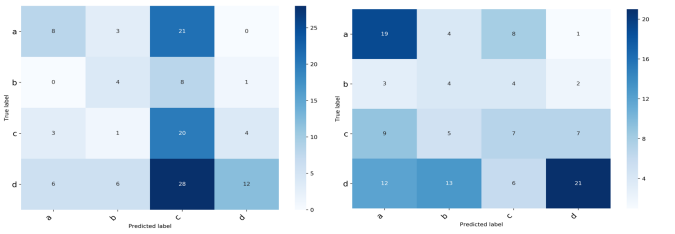

Figure 2. Confusion matrix of the NB model (left) and the SVM model (right).

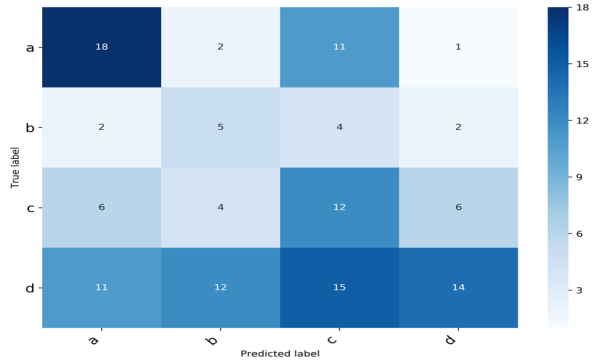

Figure 3. Confusion matrix of the ensemble model.

Our macro-averaged F1 scores from the training dataset ranged from 0.147 to 0.43. Table 4 summarizes the performance of the 6 models. SVM and NB had best F1 scores. Based on these results, we applied three models to the test dataset: SVM, NB, and an ensemble model built from the top three models: NB, SVM, and GB. The rule for breaking the tie in the ensemble model was to set the order of the preference: B (highest), C, A, then D (lowest).

Figures 2-3 show the confusion matrices of the 3 models, and Table 5 summarizes the performance of the three models submitted to the competition. These models' macro-averaged F1 scores on the holdout test dataset ranged from 0.338 to 0.379. The ensemble model had the best macro-F1 score 0.379, which was ranked 3rd among the particpating teams for this shared task competition.(Zirikly et al., 2019)

Table 5. Model performance from the test dataset. Level A-D represent no risk, low risk, moderate risk, and severe risk, respectively.

	NB	SVM	Ensemble
Marco-F1 score (4 risk levels)	0.338	0.370	**0.379**
Accuracy	0.352	**0.408**	0.392
F1 score (Flagged vs. no risk)	**0.836**	0.789	0.818
F1 score (Urgent vs. non-Urgent)	**0.736**	0.603	0.648
Level-A Precision/Recall/F1	0.471/0.250/0.327	0.442/0.594/0.507	0.486/0.562/**0.522**
Level-B Precision/Recall/F1	0.286/0.308/**0.296**	0.154/0.308/0.205	0.217/0.385/0.278
Level-C Precision/Recall/F1	0.260/0.714/**0.381**	0.280/0.250/0.264	0.286/0.429/0.343
Level-D Precision/Recall/F1	0.706/0.231/0.348	0.677/0.404/**0.506**	0.609/0.269/0.373

Table 6. Top 10 features from the feature space

Rank	Domain	Feature Description
1	SRL	Max. count of arg1 with value '*I*' in one post
2	SRL	Max. count of arg1 with value '*me*' in one post
3	FPB	Number of posts to SuicideWatch in the last two years
4	FPB	Number of weeks with any SuicideWatch posts in the last two years
5	SRL	Max. count of arg1 with value '*myself*' in one post
6	Empathy	Max. value of negative emotion in a post
7	SRL	Average count of arg1 with value '*I*'
8	Emotion	Average count of '*disgust*'-related terms across all posts
9	Empathy	Max. value of '*death*' topic across all posts
10	FPB	Max. number of SuicideWatch posts in any week in the last two years

The NB model had the best performance in two additional binary-classification tasks, i.e., no risk vs. flagged risk (any risk level other than no risk) with F1 score 0.836 and no or low risk vs. urgent risk (moderate or severe risk) with F1 score 0.736.

We started modeling from a total of 7,603 features from 10 feature domains in Section II.2, and Table 6 lists top 10 features from the whole training dataset ranked in the order of information gain. Among the top 100 features, there were 35 clinical finding features, 25 Empathy features, 17 SRL features, 14 user post-pattern features from forum posting behavior (FPB), 6 Readability features, and

3 Emotion features. Among 17 SRL features, 6 of them were related to self-referencing.

IV. Discussion and Limitations

In this study, we developed a wealth of structured features from longitudinal freetext posts, built 6 state-of-the-art machine learning models, and tested 3 models in a test dataset from the CLPsych2019 organizers. We demonstrated that data-driven machine learning models identified users with risk of suicide based on their Reddit posts. The SVM model had best macro-averaged F1 score for classifying 4 categories of suicide risk, which could be attributed by its hyperspace parameters and nonlinearity; the NB model had accurate macro-averaged F1 scores for classifying binary groups: flagged vs. no risk, and urgent risk vs. non-urgent risk groups. The NB performance may be attributed by its simple assumption and a relatively smaller number (75) of variables compared with others.

Based on the top 100 features used by the SVM model, we found that SRL, Empathy, Readability, Clinical findings, and user post patterns identified in FPB were important for classification. Most importantly, our top findings revealed that frequent self-referencing like 'I', 'me', and myself' (ranked 1, 2, 5, 7, 19) and negated self-referencing (ranked 35) posed an elevated risk as illustrated in literatures.(Burke et al., 2017; Quevedo et al., 2016)

On the other hand, LDA topic modeling, sentiment analysis, and social determinants of health did not play critical roles for classification in our experiments. We attributed its low impact due to the variety of subreddits in the cohort, which possibly makes it challenge to effectively group certain topics for classifying suicide risk levels. Our sentiment tool was based on the context of movie reviews, which may not be applicable to the suicide prediction task from Reddit posts. For social determinants of health, we built the model based on clinical data, which may be limited for social media data.

The oversampling strategy for model training improved predictive performance. Our conjecture is that oversampling enables a classifier to better tune its parameters for those rare occurrences.

The deep neural networks (CNN and LSTM) did not perform well. Both DNNs employed all the features identified in the feature engineering section. The potential explanation is that there were limited number of users in low and moderate risk levels and there were many input variables. Another factor we may consider in the future is the development of more complicated DNN structure and/or the use of multiple DNNs to catch the temporal, wide variety of feature space, and system non-linearity.

V. Conclusions

In this study, the ensemble model had best macro-averaged F1 score, and Naïve Bayes performed best for identifying users with flagged or urgent suicide risk based on longitudinal posts on Reddit discussion forums in conjunction with features from clinical findings, empathy categories, semantic role labeling, user post-patterns, readability, and emotion.

Correspondence: Fuchiang (Rich) Tsui: tsuif@chop.edu

References

Taylor A Burke, Samantha L Connolly, Jessica L Hamilton, Jonathan P Stange, Y Lyn, and Lauren B Alloy. 2017. Two Year Longitudinal Study in Adolescence. , 44(6):1145–1160.

Rafael A. Calvo, David N. Milne, M. Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*.

Nitesh V Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2017. Discovering shifts to suicide ideation from mental health content in social media. In *Proc SIGCHI Conf Hum Factor Comput Syst.*, pages 2098–2110.

Ethan Fast, Binbin Chen, and Michael Bernstein. 2016. Empath: Understanding Topic Signals in Large-Scale Text. Technical report.

Ma Hall. 1999. Correlation-based feature selection for machine learning. *Diss. The University of Waikato*.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep Semantic Role Labeling: What Works and What's Next. In

Lindsay M Howden and Julie A Meyer. 2011. 2010 Census Briefs; Age and sex composition: 2010. Technical Report May.

Haixia Liu, Lingyun Shi, Neal Ryan, David Brent, and Fuchiang Rich Tsui. 2019. Developing an annotation guideline for the classification of social determinants of health from electronic healthrRecords. Technical report.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. In *Computational Intelligence*.

NIMH. 2018. Suicide.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, and Matt Gardner. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT 2018*, pages 2227–2237.

Jose D. Posada, Amie J. Barda, Lingyun Shi, Diyang Xue, Victor Ruiz, Pei-Han Kuan, Neal D. Ryan, and Fuchiang (Rich) Tsui. 2017. Predictive Modeling for Classification of Positive Valence System Symptom Severity from Initial Psychiatric Evaluation Records. *Journal of Biomedical Informatics*.

Wei Quan, Haixia Liu, Lingyun Shi, Neal Ryan, David Brent, and Fuchiang Tsui. 2019. Classifying social determinants of health from electronic health records using deep neural networks. Technical report.

Karina Quevedo, Rowena Ng, Hannah Scott, Jodi Martin, Garry Smyda, Matt Keener, and Caroline W. Oppenheimer. 2016. The neurobiology of self-face recognition in depressed adolescents with low or high suicidality. *J Abnorm Psychol.*, 125(8):1185–1200.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, September.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36.

Fuchiang Rich Tsui, Victor Ruiz, Amie Barda, Ye Ye, Srinivasan Suresh, and Andrew Urbach. 2017. Retrospective and Prospective Evaluations of the System for Hospital Adaptive Readmission Prediction and Management (SHARP) for All-Cause 30-Day Pediatric Readmission Prediction Children ' s Hospital of Pittsburgh of UPMC , Pittsburgh , PA. In *AMIA 2017*.

Fuchiang Rich Tsui, Lingyun Shi, Victor Ruiz, Neal Ryan, Candice Biernesser, and David Brent. 2019. A large-data-driven approach for predicting suicide attempts and suicide deaths. In *The 17th World Congress of Medical and Health Informatics*, Lyon, France.

Wikipedia. Reddit wikipedia.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 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: From Keyboard to Clinic*.

Psychological disorders.