Analysis of Online Suicide Risk with Document Embeddings and Latent Dirichlet Allocation

Noah Jones  
Affective Computing  
MIT Media Lab  
Cambridge, USA  
nconcjones@mit.edu

Natasha Jaques  
Affective Computing  
MIT Media Lab  
Cambridge, USA  
jaquesn@mit.edu

Pat Pataranutaporn  
Fluid Interfaces  
MIT Media Lab  
Cambridge, USA  
patpat@mit.edu

Asma Ghandeharioun  
Affective Computing  
MIT Media Lab  
Cambridge, USA  
asma_g@mit.edu

Rosalind Picard  
Affective Computing  
MIT Media Lab  
Cambridge, USA  
picard@mit.edu

Abstract—Machine learning to infer suicide risk and urgency is applied to a dataset of Reddit users in which the risk and urgency labels were derived from crowdsourcing consensus. We present the results of machine learning models based on transfer learning from document embeddings trained on large external corpora, and find that they have very high F1 scores (.83 – .92) in distinguishing which users are labeled as being most at risk of committing suicide. We further show that the document embedding approach outperforms a method based on word importance, where important words were identified by domain experts. Finally, we find, using a Latent Dirichlet Allocation (LDA) topic model, that users labeled at-risk for suicide post about different topics to the rest of Reddit than non-suicidal users.

Index Terms—suicide, machine learning, automatic risk assessment, online, forum

I. INTRODUCTION

Suicide accounted for 1.4% of all deaths globally in 2016, and is the second leading cause of death worldwide among those aged 15-29 years old [1]. Predicting suicidal outcomes is a key challenge given factors such as under reporting [2], limited access to care [3], and poor risk identification [4]. Recently, Machine learning (ML) and Natural Language Processing (NLP) have emerged as tools for estimating suicide risk [5]. Researchers have identified linguistic markers for suicide from textual information such as blogs [6], poems [7], clinical notes [8], suicide notes [9], text messages [10], electronic medical records [5] and online social media data [11], [12]. A common approach is to derive linguistic features from psychological literature or emotion and mental disease-based lexical categories [13]. Researchers are also beginning to explore more complex methods, such as deep-learning-based approaches, often resulting in significant performance gains [11].

We developed a method to augment suicide risk assessment on the Reddit Forum r/SuicideWatch using transfer learning from stacked, contextualized, pooled document embeddings extracted from concatenated user posts. This model achieved the highest overall accuracy in the Shared Task at the 2019 NAACL Workshop on Computational Linguistics and Clinical Psychology (CLPsych), and the highest F1 score for detecting which users are most urgently at risk of suicide [14]. In this paper, we describe how those results were obtained, discuss strengths and limitations of the approach and conduct automated topic analysis to give additional insights.

II. DATASET

The labeled dataset we used was developed by Shing et al. [13]. It consists of an anonymized set of every available Reddit posting from 2005-2015, and an extracted set of labels for users who posted on r/SuicideWatch. Reddit data is public and users are anonymized; however, Shing et al. took an extra level of precaution by replacing Reddit ID’s with numeric identifiers. Sensitive information were de-identified using Named Entity Recognition and replaced with special tokens. Shing et al. defined four categories to consider in assessing suicide risk level (T=Thoughts of suicide, L=logistics/access, C=context, and F=feelings) based on Corbitt-Hall et al.’s [15] definitions of risk categories. Annotation labels came from 934 users on the crowdsourcing platform CrowdFlower. A second dataset was released after the Shared task, which collected expert labels. Experts had high inter-rater reliability in their agreement (Krippendorf’s $\alpha > .8$) compared to the crowdsourcers (Krippendorf’s $\alpha = .554$). Shing et al. compared the expert and crowdsourcer labels for 245 users and reported a macro F1 score of 0.5047 for consensus human prediction. Crowdsourced ratings were reported as being biased toward labels indicating more severe risk. The corpus of 919 r/SuicideWatch posts have labels at the user level with one of four different risk categories: (a) no risk, (b) low risk, (c) moderate risk, and (d) severe risk. These are also grouped into binary urgency levels for the urgency assessment metric: (a,b) are non-urgent and (c,d) are urgent. For further details, please see Shing et al.’s full descriptions of the data together with a set of challenges to perform with it [13].

III. METHODS

We apply a transfer learning approach based on leveraging pre-trained document embeddings from deep neural network models trained on other language tasks beyond Reddit posts. This type of transfer learning may be particularly valuable in suicide risk prediction where the difficulty of obtaining reliable clinical labels limits the size of the data. We train several classifiers using both stacked contextualized string document embeddings (FLAIR) [16] and GloVe Embeddings [17].
While these models are powerful, clinicians and domain experts could object to the opaque, black-box nature of deep neural network features. Therefore, we also train models that combine word embeddings with domain expertise to create a set of domain-knowledge features. We study how these models are able to provide additional insights that would not be apparent when relying solely on document embeddings.

A. Preprocessing

We removed de-identification tokens. Since all posts belonging to a specific user were given the same label and were relatively short, they were concatenated and a space was added between punctuation. The concatenated documents ranged from 70 to 512 tokens.

B. Features

1) Document Embeddings: We explored the use of a new form of deep neural network word embedding model which is able to obtain a more reliable representation of longer pieces of text, and can thus create more reliable document embeddings [16]. These recently proposed FLAIR embeddings have comparable performance to methods such as ELMo embeddings and reach state-of-the-art performance on Named Entity Recognition tasks [19].

Because the model operates directly on characters and does not need to limit its vocabulary by stemming words, it is better able to represent the context of a sentence (for example, word tense in a sentence). Note that FLAIR embeddings are computed on words as they occur in the context of a sentence, so the same word might have a different embedding depending on the context. When combined with hyperparameter tuning, FLAIR embeddings have shown superior performance on tasks even above fastText embeddings and Google AutoML [20].

We use embeddings generated by the forward and backward models in addition to pretrained GloVe embeddings [17]. The word embeddings are stacked together and pooled with mean pooling. Thus, we use the average embedding of words in the document. We then applied principal component analysis to reduce the dimensionality of the embeddings from 4196 to 120 while still explaining 95% of the variance in the data.

2) Domain Knowledge Features: For our second approach, we employed expert knowledge, and included a dictionary of common Twitter search terms based on known suicide risk factors from Jashinsky et al. [18]. We filtered the terms proposed by Jashinsky and colleagues [18] retaining and clustering those terms that pertained to the four categories of risk assessed in the CLPsych dataset discussed above.

The final list of terms and their associated categories that we derived is shown in the Table III.

To employ these terms to create a suicide risk classifier, we compute the word2vec embedding of each term in the list, and store it in a feature dictionary. For each post, we create a number of features based on the cosine similarity between the word2vec embedding of the words in the post, and the embeddings of the terms and topics stored in the feature dictionary. Note that it is not possible to apply FLAIR embeddings in this context, because FLAIR embeddings depend on the context of the rest of the sentence to create the embedding.

To create the domain-knowledge features, for each category of suicidal terms (T, L, C, or F) we compute aggregate statistics about the similarity of words in the post to words in the category, including the average, median, skew, mode and max. We also compute the cosine similarity between the embeddings of each word in the post and each word in the feature dictionary, and use the median of these values to create a similarity feature for each dictionary word. This resulted in a final set of 71 features.

C. Classifiers

We experimented with classifiers including Random Forests, Logistic regression 2, and an SVM using a scikit-learn implementation module (Pedregosa et al., 2011). We performed hyper-parameter tuning using the validation set. Random Forest classifiers performed the best with both embedding and domain-knowledge features and were used in the final models.

IV. RESULTS AND DISCUSSION

A. Detection of Urgency

Due to the importance of detecting urgently at-risk users, we focus our approach on a binary classification task of separating urgent and non-urgent users. We compare classifiers based on document embeddings derived from deep neural networks, with hand-engineered features based on domain knowledge. Table II presents the results; the document embeddings classifier achieves markedly higher performance on this task. This can be explained by the fact that the document embeddings are trained on much larger text datasets and so are able to build a more robust general understanding of language, which can be effectively transferred to the current task where labeled data are limited.

In contrast, the domain-knowledge classifier which uses a dictionary of words to attempt to build on expert knowledge and transfer it to the task, does not demonstrate good performance for language in context. As an example, consider the following post from the data set which was labeled as ‘a - No risk’: ‘I don’t really want to die, I just want the pain to stop’. This post contains words like pain and die, which are highly similar (or the same) as words like pain and killing in the domain-knowledge feature dictionary. However, the user is explaining that they do not actually wish to commit suicide, although they are in pain. The domain-knowledge classifier is not able to understand the grammatical and semantic meaning of the words in context, and so is out-performed by the deep learning features which are able to encode this information.

2We experimented with Limited Memory Broyden–Fletcher–Goldfarb–Shannon (LBFGS) and Stochastic Gradient Descent (SGD) optimizers.
TABLE I
CATEGORIES OF SUICIDAL LINGUISTIC TERMS PROPOSED BY [18] AND USED IN THE DOMAIN-SPECIFIC SEMANTIC SIMILARITY CLASSIFIER.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thought</td>
<td>thoughts, used, multiple, past, suicide, thought, killing.</td>
</tr>
<tr>
<td>Methods/Access</td>
<td>shooting, prozac, gun, suicide, went, zoloft, alcohol, pills, range, sertraline</td>
</tr>
<tr>
<td>Context</td>
<td>attempted, fight, sister, parents, abused, friend, brother, tried, suicide, dad, pain</td>
</tr>
<tr>
<td>Feelings</td>
<td>hopeless, depressed, alone, anxious, abused, empty, impulsive, worthless, sad, feel, hurt, helpless</td>
</tr>
</tbody>
</table>

TABLE II
BINARY CLASSIFICATION PERFORMANCE ON THE URGENT RISK TASK WITH STANDARD ERROR FROM 10 RANDOM INITIALIZATIONS

<table>
<thead>
<tr>
<th>Model</th>
<th>Macro Avg. F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document Embeddings</td>
<td>.79 ± .027</td>
<td>.81 ± .025</td>
</tr>
<tr>
<td>Domain-knowledge Semantic Similarity</td>
<td>.73 ± .024</td>
<td>.75 ± .023</td>
</tr>
</tbody>
</table>

B. Word importance

While the deep-learning-based models provide higher performance, it is difficult to extract decision rules that the classifier is using to determine suicide risk. This could become a problem for professionals who would like to know what signs to look for in online posts that may indicate higher suicide risk. While a common approach is to assess feature importance using a metric like information gain, some authors have criticized this measure as biased [21]. We use the approach of Parr et al. [22] to infer importance of the domain-knowledge features. This is done by training a classifier and assessing how much the prediction error increases when the values of a specific feature column of the validation set are shuffled. This is repeated for all features in the model. The results can be found in Fig. 1. We find that words about objects and substances that could actually be used to carry out a suicide (alcohol, zoloft, gun, sertraline, prozac) are some of the most important features. This mirrors previous findings [23] from the Crisis Text Line Challenge where 54 million messages were analyzed and “ibuprofen” and “bridge” appeared to be the most indicative of risk.

We find that the two most important words are worthless and parents. Feelings of worthlessness could be an important factor given work demonstrating that among 20 symptoms for depression, worthlessness was the only symptom associated with a lifetime suicide attempt [24]. The importance of parents to suicide risk is likely to be higher among individuals less than 26 years old, a demographic among whom the risk of suicide has been increasing dramatically [25]. Because 58% of Reddit users are under the age of 29 (compared to only 22% of adults in the U.S. population) [26], the Reddit dataset is likely to be representative of this population. Note that other words related to social relationships (friend, fight, alone, sister, brother, abused) also feature prominently among the most important terms to the classifier.

1) Word importance in black-box classifiers: We also assess the importance of each word to a sample post for both the FLAIR-based and domain-knowledge classifiers using the approach of Felbo et al. [27]. We iteratively remove each word from a sample post in our dataset and re-compute the predicted suicide risk. As is evident in Fig. 2, the domain-knowledge classifier focuses heavily on words that are close in embedding space to words in the dictionary such as pills. In contrast, the predictions of the document-embeddings classifier change to at least some degree no matter which word is replaced, and more importance is placed on words that might help distinguish degree of risk, such as tempted.

![Fig. 1. Feature importance in the domain-knowledge classifier](image1.png)

![Fig. 2. Word importance comparing domain knowledge and domain agnostic classifiers](image2.png)

C. LDA topic analysis

Aside from analyzing risk at the r/SuicideWatch post level, we also attempted to understand risk at the Reddit community level. We used Latent Dirichlet Allocation (LDA) to create a topic model of the dataset [28]. In this case, LDA learns a generative model of Reddit posts, in which each post can be described by a mixture of latent topics that are discovered by the model. Each topic represents a distribution over possible words. To train an LDA model on Reddit, we combined the 919 posts from the CLPsych dataset with 919 other randomly sampled posts from Reddit users that had never posted on r/SuicideWatch. The model that obtained the best average topic coherence [29], with a score of −2.207, had 7 topics. Table III presents those topics, including the words most important to each topic. Salient words which were used to decide how to label each topic are bold.

Fig. 3 shows which topics are discussed by users at different levels of suicide risk. Non SuicideWatch users are those who
never posted on r/SuicideWatch. We see that these users, which can be conceptualized as the average Reddit user, discuss all topics fairly equally, with a particular focus on social relationships and technology.

Users posting on r/SuicideWatch and rated as higher-risk appear to focus much more strongly on discussing social relationships and seeking help for suicide than any of the other topics. This could suggest that they use Reddit mainly as a way of seeking help, rather than engaging with its other communities. Interestingly, users who have posted on r/SuicideWatch, but were deemed no risk (category ‘a’), appear to discuss most the topic of social relationships.

**V. Conclusions**

Suicide rates are increasing dramatically among those 26 years and younger [25], and because Reddit users predominantly belong to this demographic [26], we believe there is promising potential for classifiers which can automatically determine suicide risk from online posts. For example, an automated detector of risk and urgency might alert a suicide-prevention human expert to come online quickly and write a supportive response to a person who is deemed at urgent risk.

We developed models to assist in labeling suicide risk for online posts in the anonymous forum Reddit. A major limitation of this work is that the dataset risk labels were provided by crowdworkers and never proven to be associated with actual suicidal behaviors (unlike the work of Adamou et al. [30]). Future work should develop datasets where the labels have clinical validation.

The suicide risk models we presented accurately detected labels for any level of suicide risk ($F1=92$), and distinguished which users were said to be at most urgent risk with high performance ($F1=83$). While this dataset did not have ground truth clinical outcomes, which are gold standard in risk prediction [31], the model still may be utilized as a way to assess face validity, at least approximating judgments of a naive reader of the Reddit posts [14].

While we found that the best performance was obtained by leveraging representations learned by deep language models on large datasets outside of Reddit, an approach similar to the winning team of Shared task 2019 [32], this approach lacks interpretable decision rules. Therefore, we examined word importance and LDA-based topics to reveal additional insights into features related to the higher risk group. In the future, we would like to experiment with adding LDA topics as an additional input to our classifiers, and with additional ways of combining expert knowledge with transfer learning from deep neural networks trained on large datasets.

**ACKNOWLEDGMENT**

We thank Kristy Johnson for suggesting we enter the CLPsych contest and Karthik Dinakar, Judy Shen and Agata Lapedriza for helpful discussions.
REFERENCES