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

Noah Jones
Affective Computing
MIT Media Lab
Cambridge, USA
ncjones@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 is applied to a dataset of the suicidality of Reddit users in which the suicide risk labels were derived from knowledge of expert clinicians. 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 most at risk of committing suicide. Thus, these models could potentially provide valuable aid in triaging care for individuals most in danger. We compare the document embedding approach with one which incorporates expert domain knowledge. Word importance is assessed as a way of suggesting signs that could indicate suicide risk in online posts. Finally, we learn a Latent Dirichlet Allocation (LDA) topic model and find that suicidal users 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]. In the United States alone, the suicide rate has increased by 24% over the past 20 years, and is now among the top 10 causes of death [2]. Moreover, there has been a generational shift in suicide-related outcomes, where young adults aged 25 and below are experiencing more distress, major depression, suicidal thoughts, and making more suicide attempts [3].

These issues are compounded by problems with risk identification and poor access. Some authors have argued that relatively little progress in identifying risk factors has lagged [4]. In addition, 24 million Americans live in federally designated mental health care shortage regions [5]. Even for those that are able to receive care, many psychotherapists lack the specialized clinical training needed to adequately support these populations [6].

Automatic risk detection may be a promising strategy to deal with these challenges. Researchers such as Nock and Coppersmith [7] suggest that the science of suicide assessment and treatment can be beneficially augmented by capturing people’s experiences in-situ. In addition, many people are spending an increasing amount of time online, and in online discussion forums such as Reddit and ReachOut which provide opportunities for people to deal with mental health issues, gain support, and find connections. Because online communities have a much higher proportion of users under the age of 26 [8], analyzing online posts may be a promising way to reach the demographic at greatest risk of suicide [3].

Here we tackle the problem of predicting suicide risk from a dataset provided through the 2019 ACL Workshop on Computational Linguistics and Clinical Psychology (CLPsych), which proposed a shared task focused on suicide risk classification and screening [9]. The dataset for the task was created using online posts from the website reddit.com.

We completed the risk assessment portion of the CLPsych challenge, using various models to classify the degree of suicide risks of Reddit users. Our best model used an approach involving transfer learning from stacked, contextualized, pooled document embeddings [10] extracted from concatenated user posts. This model achieved the highest overall accuracy in the competition, and the highest F1 score on the sub-task of detecting which users are most urgently at risk of suicide.

The ability to detect which individuals are urgently at risk could provide significant clinical value in terms of focusing scarce resources on those most in need. However, this is an extremely difficult task, even for trained professionals. According to Ahmedani and colleagues (2015), 25% of suicidal patients met with a health professional one week prior to their attempt [11]. Continuous in-situ monitoring between visits can ensure more responsive treatment, and accurate risk triaging will play a large role in the effectiveness of that [11].

In this paper we present both the description of and results of our entry to the CLPsych competition, and additional results focused on the binary classification task of separating individuals posting on r/SuicideWatch into two groups: those urgently at risk and those not urgently at risk. We compare the performance of models based on transfer learning from deep neural networks trained on large external corpora, to models trained specifically for suicide prediction by incorporating expert domain knowledge. While the domain-knowledge models


dedicated mental health care shortage regions [5]. Even for those that are able to receive care, many psychotherapists lack the specialized clinical training needed to adequately support these populations [6].

Automatic risk detection may be a promising strategy to deal with these challenges. Researchers such as Nock and Coppersmith [7] suggest that the science of suicide assessment and treatment can be beneficially augmented by capturing people’s experiences in-situ. In addition, many people are spending an increasing amount of time online, and in online discussion forums such as Reddit and ReachOut which provide opportunities for people to deal with mental health issues, gain support, and find connections. Because online communities have a much higher proportion of users under the age of 26 [8], analyzing online posts may be a promising way to reach the demographic at greatest risk of suicide [3].

Here we tackle the problem of predicting suicide risk from a dataset provided through the 2019 ACL Workshop on Computational Linguistics and Clinical Psychology (CLPsych), which proposed a shared task focused on suicide risk classification and screening [9]. The dataset for the task was created using online posts from the website reddit.com.

We completed the risk assessment portion of the CLPsych challenge, using various models to classify the degree of suicide risks of Reddit users. Our best model used an approach involving transfer learning from stacked, contextualized, pooled document embeddings [10] extracted from concatenated user posts. This model achieved the highest overall accuracy in the competition, and the highest F1 score on the sub-task of detecting which users are most urgently at risk of suicide.

The ability to detect which individuals are urgently at risk could provide significant clinical value in terms of focusing scarce resources on those most in need. However, this is an extremely difficult task, even for trained professionals. According to Ahmedani and colleagues (2015), 25% of suicidal patients met with a health professional one week prior to their attempt [11]. Continuous in-situ monitoring between visits can ensure more responsive treatment, and accurate risk triaging will play a large role in the effectiveness of that [11].

In this paper we present both the description of and results of our entry to the CLPsych competition, and additional results focused on the binary classification task of separating individuals posting on r/SuicideWatch into two groups: those urgently at risk and those not urgently at risk. We compare the performance of models based on transfer learning from deep neural networks trained on large external corpora, to models trained specifically for suicide prediction by incorporating expert domain knowledge. While the domain-knowledge models
provide greater interpretability, the deep learning embedding features provide the most accurate performance.

Finally, we train a Latent Dirichlet Allocation (LDA) topic model on posts to other Reddit forums (not including r/SuicideWatch), made by a set of randomly sampled non-suicidal users, and posts by those who have previously posted on r/SuicideWatch with expert-derived labels. We compare the topics discussed by these different user groups, and find notable differences in the topics focused on by potentially suicidal users. We hope that these findings can provide beneficial insight into the online behavior of at-risk individuals, even when they are not posting on explicitly suicide-oriented channels such as r/Suicidewatch.

The contributions of this paper are as follows:

- Models which leverage recently proposed dynamic document embeddings to obtain a better understanding of language context, and are able to accurately distinguish which users posting in an online forum are most at risk of committing suicide.
- An analysis of the trade-offs between leveraging transfer learning from powerful deep learning models trained on large datasets, and using more interpretable features incorporating domain-specific expertise.
- An analysis of text-based signs of suicide risk, including differences in the types of posts made by the general Reddit population versus by users identified as at risk of suicide.

II. Description

Much of the prior work regarding mental health estimation in social media has explored non-clinician or approximate methods for creating severity labels. These labels, although good for approximation, lack ground truth clinical assessment. Against this backdrop, the CLPysch dataset used for this paper includes annotation labels from clinical experts with high interrater reliability in their agreement (Krippendorf’s α > .8) [9]. More specifically, the consensus labels originally generated using a crowdsourcing platform called CrowdFlower were cross-examined using an expert dataset of 250 ratings that had high interrater reliability.

The corpus of 982 posts was tagged at the user level. Posts belonged to one of four different risk categories (a, b, c, d), which are described in Table I. To create a simplified binary classification task, we focus on the ‘urgent’ task in the CLPysch competition, which grouped labels into two categories: urgent (c, d) vs. non-urgent (a, b). A model that can accurately classify urgent vs. non-urgent users would allow for effectively triaging those individuals most in need of help.

III. Methods

Deep neural networks can be used to compress high-dimensional data into meaningful vector representations in a lower-dimensional embedding space. This embedding space is valuable because the distance between two embedding vectors can represent the semantic dissimilarity between two data points [12]. Distance can easily be assessed with cosine similarity or Euclidean distance. When such networks are trained in an unsupervised manner on very large text datasets (e.g. [13]), they are able to learn a robust and generalizable representation of language that can be readily transferred to provide performance enhancements on language tasks where data is more scarce [14]. This type of transfer learning is particularly valuable in the suicide risk prediction setting, where the difficulty of obtaining reliable clinical labels limits the size of the dataset.

Thus, we apply a transfer learning approach based on leveraging pre-trained document embeddings obtained from deep neural network models trained on other language tasks. We train several classifiers using both stacked contextualized string document embeddings (FLAIR) [10] and GloVe Embeddings [15].

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 semantic similarity vectors. We study how these models are able to provide additional insights that would not be apparent when relying solely on document embeddings.

A. 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 [10]. These recently proposed FLAIR embeddings have been able to reach state-of-the-art performance on Named Entity Recognition tasks. The embeddings are generated by passing sentences as sequences of characters into a character-level language model to obtain word-level embeddings [10]. 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, the tense of the words in the 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 [16].

We used the embeddings generated by the forward and backward models in addition to pretrained GloVe embeddings [15]. The word embeddings are stacked together and pooled with mean pooling. Thus, we use the average embedding of words in the document. Afterwards, we applied a principal component analysis (PCA) to reduce the dimensionality of the embeddings from 4196 to 120 while still explaining 95% of the variance in the data.

2) Semantic Similarity Vectors: For our second approach, we capitalized on expert knowledge to create a set of features based on semantic similarity to known terms related to suicide risk. Jashinsky and colleagues (2016) [17] developed a dictionary of common Twitter search terms based on known suicide risk factors such as family violence and prior suicide
TABLE I
LEVELS OF SUICIDE RISK IN THE CLPsych DATASET

<table>
<thead>
<tr>
<th>Post label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>No Risk: I don’t see evidence that this person is at risk for suicide</td>
</tr>
<tr>
<td>b</td>
<td>Low Risk: There may be some factors here that could suggest risk, but I don’t really think this person is at much of a risk of suicide</td>
</tr>
<tr>
<td>c</td>
<td>Moderate Risk: I see indications that there could be a genuine risk of this person making a suicide attempt</td>
</tr>
<tr>
<td>d</td>
<td>Severe Risk: I believe this person is at high risk of attempting suicide in the near future</td>
</tr>
</tbody>
</table>

Attempts. Experts then filtered this list by determining whether these terms were linked to posts related to genuine suicide risk.

Expert knowledge was also used in creating the CLPsych Reddit dataset used in this paper [9]. Both experts and crowd-sourced workers on the platform CrowdFlower were instructed to assess suicide risk based on four families of risk factors: thoughts of suicide, logistics (methods/access), context, and feelings. We combined these two sources of knowledge by filtering the terms proposed by Jashinsky and colleagues (2016) [17] and retaining only those terms that pertained to the four categories of risk assessed in the CLPsych dataset. The final list of terms and their associated categories that we derived is shown in the Table IV.

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-expertise features, for each topic or category of suicidal term (thought, methods/access, context, feelings), we compute aggregate statistics about the similarity of words in the post to words in the topic, 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.

B. Classifiers

The features described above were used as input to a variety of different classifiers, assessing performance on a validation set (the test set for the CLPsych task was not provided until the final week to ensure the data could not be used for tuning the models). We experimented with classifiers including Random Forests, Logistic regression with both a Limited Memory Broyden–Fletcher–Goldfarb–Shannon (LBFG) and Stochastic Gradient Descent (SGD) optimizer, and an SVM using a scikit-learn implementation module (Pedregosa et al., 2011). We performed hyper-parameter tuning using the validation set. The final classifier used for the CLPsych task was a multinomial logistic regression with the LGFB optimizer. For the binary classification task, Random Forests performed best with both embedding and domain-specific features.

C. Word importance

A strength of the domain-engineered classifier is that it can provide more interpretable results. Specifically, we can assess the importance of each of the words proposed by Jashinsky and colleagues [17] in predicting suicide risk on Reddit. While a common approach is to assess feature importance using a metric like information gain, some authors have criticized this measure as biased [18]. Therefore, we adopt the approach of Parr and colleagues (2018) in assessing word importance using permutation importance [19]. This is done by training a classifier and assessing how much the prediction error increases when the values of a specific feature are scrambled.

1) Examples of word importance in black box classifiers:

A generalization of the permutation importance approach can actually be applied to both black-box and domain-engineered classifiers. We can take a sample post in our dataset, and iteratively remove each word from the post and re-compute the predicted suicide risk. As an estimation of word importance, we assume that words that lead to a higher change in the predicted risk when removed have greater importance to the prediction. We use this approach to assess the importance of words in the context of the larger post with both our black-box embedding classifiers, and domain-engineered classifiers, and compare the results. A similar approach was taken by Felbo et al. [14].

The downside of this approach is it is highly computationally expensive, making it difficult to assess overall word importance across many posts. Further, even if we created such an aggregated statistic, it would only be a simplification of what the FLAIR model is actually using to make a decision. However, this technique is still useful for gaining insight into how the decision functions learned by both classifiers differ.

D. Latent Dirichlet Allocation (LDA)

To gain insight into the types of discussions taking place on Reddit and how these related to suicide risk, we used Latent Dirichlet Allocation (LDA) to create a topic model of the dataset [20]. 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. LDA has proven extremely successful for text modeling in part because it assumes that the topic distribution has a sparse Dirichlet prior, meaning that it assumes each document covers only a small number of topics, and each topic is related to a small set of important words.

The CLPsych dataset also included 919 posts made by the same set of users to other subreddits (forums) on Reddit (not...
TABLE II
CATEGORIES OF SUICIDAL LINGUISTIC TERMS PROPOSED BY [17] 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>

in r/SuicideWatch. To train an LDA model on Reddit, we combined these posts with 919 other randomly sampled posts from Reddit users that had never posted on r/SuicideWatch. We train the LDA model on the combined dataset, to determine a broad set of topics discussed on Reddit by both suicidal and non-suicidal users. We use the average topic coherence score [21] to determine the best number of topics to describe the data.

IV. RESULTS

A. CLPsych shared task

The best classifier trained with document embeddings on the original CLPsych task of assessing four levels of suicide risk was submitted to the CLPsych competition. The submitted classifier achieved the highest accuracy in the competition overall (60% across four risk categories), with an F1 score of .38. The relatively low F1 score was due to a high false-positive rate. As shown in the confusion matrix presented in Figure 1, the classifier frequently mis-classified labels of low and moderate risk as severe risk. Our system performed significantly better on classifying no risk vs high risk users than low and medium risk users, with the no risk group having the highest macro average f1 score. Note that in trading off precision and recall, a high false positive rate and low false negative rate is strongly preferred for this task, because it is extremely important not to fail to detect an individual who may be at risk of suicide. Our classifier obtains a recall of .942 in detecting individuals labeled as ‘d - Severe risk’.

B. Detection of urgent risk

Due to the importance of detecting urgently at-risk users, we focus our approach on the 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 III 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 is sparse.

In contrast, although the domain-engineered classifier attempts to build on expert knowledge and transfer it to the task, it does not have a good understanding of 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-engineered feature dictionary. However, the user is explaining that they do not actually wish to commit suicide, although they are in pain. The domain-engineered 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.

C. 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

<table>
<thead>
<tr>
<th>Model</th>
<th>Macro Avg. F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document Embeddings</td>
<td>.83</td>
<td>.84</td>
</tr>
<tr>
<td>Domain-engineered Semantic Similarity</td>
<td>.76</td>
<td>.78</td>
</tr>
</tbody>
</table>
signs to look for in online posts that may indicate higher suicide risk. Therefore, as described in Section III-C, we assess permutation importance of the domain-knowledge features, and present the results in Figure 3. 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 a previous finding that ibuprofen and bridge were the most important words in determining suicide risk for Crisis Text Line [22].

Interestingly, the two most important words are worthless and parents. For the former, 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 [23]. 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 [3]. Because 58% of Reddit users are under the age of 29 (compared to only 22% of adults in the U.S. population) [8], 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: Using the extension to permutation importance described in Section III-C1, we assess the importance of each word to a sample post for both the FLAIR-based and domain-engineered classifiers. As is evident in Figure 4, the domain-engineered classifier focuses heavily on words that are close in embedding space to words in the list in Table IV, 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.

D. LDA topic analysis

As described in Section III-D, we trained an LDA model on a dataset combining the posts of suicidal and non-suicidal users on subreddits other than r/SuicideWatch. The model which obtained the best average topic coherence score of \(-2.207\) had 7 topics. Table IV presents those topics, including the words most important to each topic. Salient words which were used to decide how to label each topic are bolded. Reddit is male-dominated (69% of users are male) [24], and the topics reveal a focus on video games and technology. There are also topics related to selling items via Craigslist, general advice, social relationships, human rights, and suicide. While the topics have some overlap (i.e. the word game appears in both the tech review and sales topics), they appear to be largely distinct.

Figure 5 shows which topics are discussed by users at different levels of suicide risk. Non suicide watch users have 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. Potentially suicidal users 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 be especially focused on social relationships. As suicide risk increases, the focus on discussing social relationships decreases and the focus on suicide itself increases. This could suggest that problematic relationships are a potential factor that drives users to post on r/SuicideWatch. It also aligns with our earlier findings that words related to relationships (like parents, friend, alone) are particularly important in assessing suicide risk. Previous literature on suicide has also emphasized the importance of relationships [17].

V. RELATED WORK

Machine learning (ML) and Natural Language Processing (NLP) have emerged as tools to estimate mental health [25]. Research has identified linguistic markers for suicide from textual information such as blogs [26], poems [27], clinical notes [28], and suicide notes [29]. Moreover, online social media data has been found to contain predictive information for a range of mental health conditions including depression and suicide [30], [31].
### Table IV

<table>
<thead>
<tr>
<th>Topic</th>
<th>Terms ordered by importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suicide help</td>
<td>feel, go, like, get, want, know, think, life, time, make, even, one, my, live, try, would, people, feel like, never, fuck, tell, day, take, die, see, say, anymore, much, thing, really, end, kill, way, work, friend, everything, leave, year, noth, love, help, anything, still, depress, suicide, better</td>
</tr>
<tr>
<td>Social relationships</td>
<td>want, get, know, go, like, friend, think, help, thing, really, feel, year, people, time, one, would, live, make, say, talk, suicide, even, day, tell, try, my, need, take, start, never, work, someone, much, back, live, see, care, good, family, find, love, give, could, come, school, something</td>
</tr>
<tr>
<td>Tech review</td>
<td>game, play, use, get, look, would, buy, new, one, like, time, go, be, know, run, what, if, question, this, so, good, find, work, want, people, see, post, thank, day, do, cost, pc, make, also, team, could, allow, guy, around, way, reddit, you, link, how, player, try, any</td>
</tr>
<tr>
<td>Human rights</td>
<td>get, need, time, help, make, prison, use, homeless, like, one, would, thank, want, look, what, people, new, if, be, good, game, post, take, also, how, love, think, go, this, us, point, say, work, give, human, they, lot, nan, to, state, fund, many, join, see, find, much, still</td>
</tr>
<tr>
<td>Video games</td>
<td>charge, damage, time, have, enemy, range, like, hit, counter, use, deal, make, would, average, get, people, high, play, sub, come, need, send, you, can, this, infinity, multiply, wind, home, slow, also, long, move, first, non, thing, really, look, something, may, see, any, around, be, interest, help, tip, fire</td>
</tr>
<tr>
<td>General advice</td>
<td>gt, get, anyone, make, sd, know, need, could, say, find, use, help, if, like, do, work, look, one, you, be, ld, remember, today, how, see, give, way, god, new, so, build, pharmacy, back, think, much, come, try, start, guy, believe, go, also, my, month, seem, else, want</td>
</tr>
<tr>
<td>Services / sales</td>
<td>free, via, craigslist, craigslist via, iftt, need, help, since, first, one, state, fb, look, book, part, would, be, great, use, item, card, pick, come, gb, drive, level, want, stuff, team, case, game, price, get, please, include, if, build, know, side, play, thank experience, video, cpu, table, every, list, year, power</td>
</tr>
</tbody>
</table>

Fig. 5. Proportion of topics discussed by Reddit users by suicide risk

A common approach which has been deployed in suicide prediction is to derive linguistic features from psychological literature such as linguistic inquiry and word count, emotion features, mental disease lexicon, or depression based lexical categories [9]. Many researchers are also beginning to explore more complex methods to examine risk, such as deep-learning-based approaches, often resulting in drastic performance gains [30]. Instead of exploring first, many are looking towards methods such as hierarchical attention networks in order to improve classification [9]. Such approaches can be used to find interesting previously unrevealed features that are related to risk. In the Crisis Text Line Challenge 54 million messages were analyzed and “ibuprofen” and “bridge” appeared to be the most indicative of risk [22].

The CLPsych Shared Task encourages new methods to analyze language use as a signal for mental health. Previous winning submissions have used sentence [32] and word embeddings [33] in their final models and some authors have argued for the importance of contextual factors [33], [34]. Motivated by these considerations, we used novel dynamic word embeddings which are better at understanding word level context in posts.

### VI. Conclusions

In this paper we tackle the problem of detecting suicide risk from online posts in the anonymous forum website Reddit. The suicide risk prediction models presented here are able to accurately detect users that have any level of suicide risk (F1=.92), and distinguish which users are most at risk of suicide with similarly high performance (F1=.83). We believe this type of accurate detection could provide important clinical aid in triaging users that are most at risk of suicide.

The accuracy of our models was obtained by leveraging the robust representations learned by deep language models on large datasets. While this approach is highly performant, it lacks the ability to provide interpretable decision rules. Therefore, we present several methods for interpreting how to detect suicide risk, including analyzing the importance of specific words to a model based on incorporating expert domain knowledge. We also train an LDA topic model and assess which topics are most frequently discussed by users that have varying degrees of suicide risk. We find evidence that learned helplessness, discussion of methods for carrying out a suicide, and problematic social relationships are all important to identifying suicide risk.

### A. Future work

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. Given difficulty to distinguish between similar levels of risk, we would also like to train a hierarchical ensemble of classifiers to first distinguish whether a user is at risk of suicide or not, and then determine the level of risk within those users deemed suicidal.
ACKNOWLEDGMENT

We would like to thank Kristy Johnson for her suggestion that we enter the CLPscyh competition.

REFERENCES