Triaging Mental Health Forum Posts

Arman Cohan, Sydney Young and **Nazli Goharian**
Information Retrieval Lab
Department of Computer Science
Georgetown University
{arman, nazli}@ir.cs.goergetown.edu, sey24@georgetown.edu

Abstract

Online mental health forums provide users with an anonymous support platform that is facilitated by moderators responsible for finding and addressing critical posts, especially those related to self-harm. Given the seriousness of these posts, it is important that the moderators are able to locate these critical posts quickly in order to respond with timely support. We approached the task of automatically triaging forum posts as a multiclass classification problem. Our model uses a supervised classifier with various features including lexical, psycholinguistic, and topic modeling features. On a dataset of mental forum posts from ReachOut.com[1], our approach identified critical cases with a F-score of over 80%, showing the effectiveness of the model. Among 16 participating teams and 60 total runs, our best run achieved macro-average F1-score of 41% for the critical categories (The best score among all the runs was 42%).

1 Introduction

Social media such as Twitter, Facebook, Tumblr and online forums provide a platform for people seeking social support around various psychological and health problems. Analysis of social media posts can reveal different characteristics about the user, including their health and well-being (Paul and Dredze, 2011). Information exchange through social media concerning various health challenges has been extensively studied (Aramaki et al., 2011; Lampos and Cristianini, 2012; Yates et al., 2014; De Choudhury and De, 2014; Parker et al., 2015; Yates et al., 2016). Prior research has also studied social media to analyze and characterize mental health problems. Coppersmith et al. (2014) provided quantifiable linguistic information about signals of mental disorders in Twitter. Schwartz et al. (2014) analyzed Facebook status updates to build a model for predicting the degree of depression among users. Topic modeling approaches have been also investigated in automatic identification of depression from social media (Resnik et al., 2015).

Apart from prior work in general linguistic analysis for identifying mental disorders, there have been some efforts to investigate self-harm communications in social media (Won et al., 2013; Jashinsky et al., 2014; Thompson et al., 2014; Gunn and Lester, 2015; Sueki, 2015). In these works, large scale analysis of Twitter posts have been performed to identify correlations of self-harm language with actual suicide rates. On the individual level, Burnap et al. (2015) used an ensemble classification approach to classify tweets into suicide related topics such as reporting of suicide, memorial and social support. De Choudhury et al. (2016) analyzed a collection of posts from Reddit to characterize the language of suicide related posts and to predict shifts from discussion of mental health content to expression of suicidal ideation.

Compared to Twitter and Facebook which are general purpose social platforms, online mental health forums are virtual communities that are more focused on mental health issues. In these forums, users provide help and support for one another along with forum moderators. An example of such forums is ReachOut.com, which is an online youth mental health service providing information, tools and support to young people aged 14-25. Similar to many other mental health support forums, ReachOut.com provides methods for communicating anonymously about mental issues and seeking help and guidance from trained moderators. There are sometimes posts that indicate signs of self-harm. These posts need to be prioritized and attended to by the moderators as soon as possible to prevent potential harm to the at-risk user.

We propose an approach to identify forum posts

[1] http://forums.au.reachout.com/

Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 143–147,
San Diego, California, June 16, 2016.

indicating signs of self-harm; furthermore, we focus on triaging the posts based on the criticality of the content. We approach this task as a multiclass classification problem. We utilize a regularized logistic regression classifier with various sets of features extracted from the post and its context in the thread. The features include lexical, psycholinguistic and topic modeling features. In CLPsych 2016 shared task, among 60 total submitted runs by all participants, our approach achieved above median results for all of our submitted runs which shows the effectiveness of our approach. Furthermore, our best run achieved the F-1 score of 0.41 for critical categories while the best score over all the runs were 0.42.

2 Identifying self-harm posts

We identify mental health forum posts that indicate signs of self-harm and also triage these posts. The posts showing no ideation of harm are labeled as green, while the other posts are labeled as amber, red, and crisis based on their criticality. We approach this task as a multiclass classification problem. We extract lexical, contextual, psycholinguistic and topic modeling features to train the classifier.

2.1 Features

Lexical features We examine several lexical features for indications of the user's mental health. The Linguistic Inquiry Word Count (LIWC) (Pennebaker et al., 2015) is a psycholinguistic lexicon that quantifies the mental state of an individual in terms of attributes. As it contains close to 100 attributes, we experiment with different subsets to identify the most relevant measures. We identify the affective attributes subset of LIWC as the most helpful features, which include positive emotion, negative emotion, anxiety, anger, sadness, and swear.

To further quantify the emotions associated with a forum post, we use DepecheMood (Staiano and Guerini, 2014), which is a lexicon of 37k entries. In this lexicon, each expressions is assigned a relevance probability to each of the following 8 dimensions of emotions: fear, amusement, anger, annoy, apathy, happiness, inspiration and sadness. The final emotion distribution of each post is computed by sum of the probabilities for individual terms in the post divided by the total number of terms in the post. In addtion to the probabilities associated with each of the emotions, we also consider the dominant emotion as a separate feature. To distinguish between the subjective and objective forum posts, we utilize the MPQA subjectivity lexicon (Wilson et al., 2005). Each term in the lexicon has a prior polarity value of "positive", "negative", or "neutral". We assign +1 to positive, -1 to negative, and 0 to neutral. The final subjectivity feature of a post is the sum of all individual subjectivity values divided by the total number of terms in the post.

Inspection of the forum posts reveals that in many cases the critical posts consist of a lengthy post body which does not indicate any signals of self-harm. However, the author changes the tone eventually and ends the posts with a sentence that indicates signs of potential self-harm. Therefore, to also account for the final mental state of the user, we consider features extracted from the last sentence separately. Specifically, we extract subjectivity and LIWC affective features of the last sentence. To account for variations of the mental state of the user throughout the post, we also consider the variance of sentence level emotions as a separate feature.

Contextual Features During analysis, it became evident that to understand some of the posts completely, one needs to also consider the rest of conversation in the corresponding thread. Thus, we also extracted features that would provide context for the post. We consider the author's prior posts in the thread, as well as the surrounding (previous and next) posts by other users. We also considered the subject of the thread as a separate feature.

Textual Statistics We examine two types of textual statistics for each post. We categorize each thread based on the number of posts (n) in the thread: $n \leq 5$, $5 < n \leq 10$, $10 < n \leq 20$, $20 < n \leq 50$, $50 < n$. We also consider the frequency of certain seed words within the post that would signal the most serious posts. The seed word list contain "want to die", "harm[ing] myself", and "suicid[e/al]".

Topic modeling Topic modeling has been previously shown to be effective for identification of mental health problems (Resnik et al., 2015). Therefore, we utilize topic models for mapping each post to a set of predefined number of topics. We use LDA to extract the topics associated with each post. We infer the topics by training the LDA model on the entire ReachOut forum dataset.

Run	Features	Boost
1	body, author's posts, subject, emotion, thread length, LIWC (affective, female) , and seed terms.	C +.2
2	body, author's posts, emotion, thread length, LIWC (affective, female, negate), and seed terms	C +.3 R +.2 A +.1
3	body, author's posts, emotion, thread length, LIWC (affective, female, negate), seed terms, and last sentence	C +.3 R +.2

Table 1: The feature sets for each of the runs and the boosting values of Crisis (C), Red (R) and Amber (A) categories.

2.2 Classification

We experimented with several classification algorithms including SVMs with linear and rbf kernels, Random Forests, Adaboost and Logistic Regression. We also experimented with ensemble of these classifiers. Logistic regression with L1 regularization provided the best results based on 4 fold cross validation on the training set. We noticed that the classifier's recall for critical categories was quite low especially in cases of "crisis". This is expected given the low number of training posts in the critical categories. To improve the recall, we boost the prediction probabilities of the classifier for the critical categories by a constant value. We conducted a full grid search on the boosting values for each categories and based on the results on the training set, we selected two of the boosting settings for the final runs.

3 Experimental setup

The data provided by the CLPsych 2016 Shared Task consists of forum posts from Reachout.com, a mental health forum for individuals between 14-25 years old. The data contains 1,188 annotated posts with triage labels. 947 of these posts were provided for training, while 241 posts were withheld for testing. The class breakdown of the 947 training labeled posts is 39 crisis, 110 red, 249 amber, and 549 green.

The official evaluation metric for the shared task is macro-averages of F1-scores for the crisis, red, and amber categories. We also report macro-average of F1-scores and accuracy for the non-green versus the green class labels. We use stratified 4-fold cross-validation on the training dataset of 947 posts. The baseline is a classifier with unigram bag-of-words features from the body of the posts.

4 Results and discussion

Macro	NG		NG vs. G	
Average	F1	Acc	F1	Acc
Run 1	0.38	0.78	**0.82**	**0.88**
Run 2	0.33	0.75	0.80	0.86
Run 3	**0.41**	**0.80**	0.81	0.81

Table 2: Official results of the submitted runs on the test set. NG: Non-Green, G: Green, F1: F1-Score, Acc: Accuracy

Run	Crisis(1)			Red(27)			Amber(47)			Avg
	P	R	F1	P	R	F1	P	R	F1	
1	0	0	0	**62.50**	55.56	58.82	50.00	63.83	56.07	38.30
2	0	0	0	50.00	51.85	50.91	45.45	53.19	49.02	33.31
3	0	0	0	59.26	**59.26**	**59.26**	**58.93**	**70.21**	**64.08**	**41.11**

Table 3: Breakdown of Precision (P), Recall (R) and F1-Score (F1) on test set by category. The number in front of each category shows the number of gold occurrences in each category.

We evaluated different settings of features and classifiers discussed in Section 2; we then selected the settings that resulted in the best non-green macro-average F1-score as our final submitted runs (Table 1). The official results of our submitted runs are presented in Table 2. The breakdown of the results by category is presented in Table 3. Our third run achieved the highest results with 0.41 non-green average F-score (The best performance among all participants was 0.42). We were not able to identify the only instance of the crisis category correctly, hence the F-score of 0 for crisis. The detailed results of each run on the training set based on 4-fold stratified cross-validation is shown in Table 4 and the breakdown by category is illustrated in Table 5. Interestingly, while the three of the runs show comparable results on the training set (above 47%), on the test set, variation is larger. The third run, which added the context of the last sentence of the post, had the highest performance. Contrary to our expectations, the second run, which had performed the best with the training dataset showed the lowest performance with the unseen test data. This could be due to the drift caused by boosting the amber category, as also reflected in lower F-score in this category.

4.1 Feature analysis

Table 6 displays the impact of various extracted features compared with the baseline model. Overall, most of the features had a positive impact on the model's performance. The features whose addition resulted in the highest score increase are the contextual features of all the author's posts in the thread, posts not by the author in the thread, and the affective attributes and polarity of the last sentence of

	NG		NG vs. G	
	F1	Acc	F1	Acc
Baseline	36.71	86.67	75.21	81.62
Run 1	47.47	89.02	85.30	88.17
Run 2	47.67	88.38	86.12	88.60
Run 3	47.12	88.21	85.60	88.28

Table 4: Results on the training set (stratified 4-fold cross-validation). NG: Non-Green, G: Green. F1: F1-Score, Acc: Accuracy

	Crisis			Red			Amber		
	P	R	F1	P	R	F1	P	R	F1
Baseline	53.85	17.95	26.92	37.31	22.73	28.25	58.04	52.21	54.97
Run 1	33.33	20.53	25.40	52.00	47.27	49.52	68.75	66.27	67.48
Run 2	32.26	25.64	28.57	45.45	50.00	47.62	70.04	63.86	66.81
Run 3	30.30	25.64	27.78	47.06	50.91	48.91	68.78	61.04	64.68

Table 5: Results breakdown by category (training set).

Macro Average	NG		NG v. G	
	F1	Acc	F1	Acc
Body (Baseline)	36.71	86.67	75.21	81.62
Body+all LIWC	33.23	86.20	78.34	82.89
Body+thread length	34.19	86.39	75.57	81.84
Body+subject	36.47	86.94	77.57	83.21
Body+subjectivity	36.56	86.55	76.19	82.05
Body+LIWC female	36.84	86.62	75.00	81.41
Body+affective	36.88	86.52	76.71	82.37
Body+LIWC negate	37.19	86.73	75.81	81.94
Body+emotion	37.01	86.69	74.79	81.20
Body+time	37.04	86.66	75.61	81.94
Body+seeds	37.07	86.73	75.61	81.94
Body+topics	37.61	86.84	75.85	82.05
Body+last sentence	37.62	86.79	76.69	82.15
Body+surrounding posts	40.30	87.86	83.00	86.38
Body+author's posts	41.13	88.21	82.65	86.17

Table 6: Feature analysis by adding individual features to the body. NG: Non-Green; G: Green; F1: F1-Score, Acc: Accuracy

Feature Combination	NG		NG v. G	
	F1	Acc	F1	Acc
Body (Baseline)	36.71	86.67	75.21	81.62
+ affective	36.88	86.52	76.71	82.37
+ thread length	38.31	86.45	76.84	82.37
+ emotion	38.52	86.55	76.52	82.05
+ author's posts	44.81	89.05	84.5	87.65
+ LIWC female	44.93	88.81	84.77	87.86
+ LIWC negate	45.37	88.77	84.5	87.65
+ seeds	46.39	88.88	85.19	88.17

Table 7: Feature analysis for combined features. NG: Non-Green; G: Green; F1: F1-Score, Acc: Accuracy

the post. The linguistic features and textual statistics both improved and detracted from the performance of the classifier.

Once examining the effects of features individually, we experimented with feature combinations. Table 7 displays the building steps of our highest performing models. Feature combinations that did not result in improvements are not displayed due to space limitation. We observe that adding helpful features generally improves the results. Interestingly, while thread length alone with body decreased the non-green F1 score, when used in combination with the LIWC affective attributes, the performance improved.

Error analysis revealed that many of false negatives in critical cases include longer posts having a general positive/neutral tone. In such posts, when there is a small section indicating self-harm, the post becomes critical. However, when considering features from the entire post, the effect of that small section fades away. We tried to tackle this problem by considering affective sentence level features and expanding seed words, but it did not result in improvements. Limited training data in the critical categories hinders learning the optimal decision boundary in a high-dimensional feature space. This can be observed by looking at the performance breakdown by category (Table 5). We observe that the F-score for the crisis category is the lowest ($\sim$ 28%), then the red category ($\sim$ 48%) and finally the amber category ($\sim$ 67%). This trend among the categories is in line with the number of training examples in each category (39 crisis, 110 red and 249 amber). Since the number of features are relatively large, small number of training data limits learning the optimal decision boundary. On the other hand, when we try to reduce the feature space dimensionality, we are not capturing the characteristics that distinguish between the categories. Therefore, we argue that having more data in the critical categories would results in improvements in the absolute F-score measures.

5 Conclusions

We approached automated triaging of mental health forum posts as a multiclass classification problem by using various sets of features. The most effective features for this task proved to be the psycholinguistic, contextual and sentence level affective features. In addition, boosting the classifier predictions for the critical categories resulted in further improvements. All of our submitted runs achieved above median results among 16 participating teams and our best run, obtained non-green F-1 score of 41% (while the best overall result was 42%). The absolute measure of F1-scores for individual critical classes indicates that there is much room for future research in the analysis and classification of mental forum posts.

References

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1568–1576, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pete Burnap, Walter Colombo, and Jonathan Scourfield. 2015. Machine classification and analysis of suicide-related communication on Twitter. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 75–84. ACM.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *International AAAI Conference on Web and Social Media*, ICWSM '14. AAAI.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 34rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '16, New York, NY, USA. ACM.

John F Gunn and David Lester. 2015. Twitter postings and suicide: An analysis of the postings of a fatal suicide in the 24 hours prior to death. *Suicidologi*, 17(3).

Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. 2014. Tracking suicide risk factors through Twitter in the US. *Crisis*.

Vasileios Lampos and Nello Cristianini. 2012. Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology*, 3(4):72:1–72:22, September.

Jon Parker, Andrew Yates, Nazli Goharian, and Ophir Frieder. 2015. Health-related hypothesis generation using social media data. *Social Network Analysis and Mining*, 5(1):1–15.

Michael J Paul and Mark Dredze. 2011. You are what you tweet: Analyzing Twitter for public health. *International AAAI Conference on Web and Social Media*, 20:265–272.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. *UT Faculty/Researcher Works*.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015. Beyond LDA: Exploring supervised topic modeling for depression-related language in Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 99–107, Denver, Colorado, June 5. Association for Computational Linguistics.

H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Jacopo Staiano and Marco Guerini. 2014. Depeche mood: a lexicon for emotion analysis from crowd annotated news. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 427–433, Baltimore, Maryland, June. Association for Computational Linguistics.

Hajime Sueki. 2015. The association of suicide-related Twitter use with suicidal behaviour: a cross-sectional study of young internet users in Japan. *Journal of affective disorders*, 170:155–160.

Paul Thompson, Craig Bryan, and Chris Poulin. 2014. Predicting military and veteran suicide risk: Cultural aspects. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–6, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hong-Hee Won, Woojae Myung, Gil-Young Song, Won-Hee Lee, Jong-Won Kim, Bernard J Carroll, and Doh Kwan Kim. 2013. Predicting national suicide numbers with social media data. *PloS one*, 8(4):e61809.

Andrew Yates, Jon Parker, Nazli Goharian, and Ophir Frieder. 2014. A framework for public health surveillance. In *Language Resources and Evaluation*, LREC '14, pages 475–482.

Andrew Yates, Nazli Goharian, and Ophir Frieder. 2016. Learning the relationships between drug, symptom, and medical condition mentions in social media. In *International AAAI Conference on Web and Social Media*, ICWSM '16. AAAI.