Text Analysis and Automatic Triage of Posts in a Mental Health Forum

Ehsaneddin Asgari[1] and **Soroush Nasiriany**[2] and **Mohammad R.K. Mofrad**[1]
Departments of Bioengineering[1] and Electrical Engineering and Computer Science[2]
University of California, Berkeley
Berkeley, CA 94720, USA
asgari@ischool.berkeley.edu, snasiriany@berkeley.edu, mofrad@berkeley.edu

Abstract

We present an approach for automatic triage of message posts in ReachOut.com mental health forum, which was a shared task in the 2016 Computational Linguistics and Clinical Psychology (CLPsych). This effort is aimed at providing the trained moderators of ReachOut.com with a systematic triage of forum posts, enabling them to more efficiently support the young users aged 14-25 communicating with each other about their issues. We use different features and classifiers to predict the users' mental health states, marked as green, amber, red, and crisis. Our results show that random forests have significant success over our baseline mutli-class SVM classifier. In addition, we perform feature importance analysis to characterize key features in identification of the critical posts.

1 Introduction

Mental health issues profoundly impact the well-being of those afflicted and the safety of society as a whole (Üstün et al., 2004). Major effort is still needed to identify and aid those who are suffering from mental illness but doing so in a case by case basis is not practical and expensive (Mark et al., 2005). These limitations inspired us to develop an automated mechanism that can robustly classify the mental state of a person. The abundance of publicly available data allows us to access each person's record of comments and message posts online in an effor to predict and evaluate their mental health.

1.1 Shared Task Description

The CLPsych 2016 Task accumulates a selection of 65,514 posts from ReachOut.com, dedicated to providing a means for members aged 14-25 to express their thoughts in an anonymous environment. These posts have all been selected from the years 2012 through 2015. Of these posts, 947 have been carefully analyzed, and each assigned a label: green (the user shows no sign of mental health issues), amber (the user's posts should be reviewed further to identify any issues), red (there is a very high likelihood that the user has mental health issues), and crisis (the user needs immediate attention). These 947 posts-label pairs represent our train data. We then use the train data to produce a model that assigns a label to any generic post. A separate selection of 241 posts are dedicated as the test data, to be used to evaluate the accuracy of the model.

2 Methods

Our approach for automatic triage of posts in the mental health forum, much like any other classification pipeline, is composed of three phases: feature extraction, selection of learning algorithm, and validation and parameter tuning in a cross validation framework.

2.1 Feature extraction

Feature extraction is one of the key steps in any machine learning task, which can significantly influence the performance of learning algorithms (Bengio et al., 2013). In the feature extraction phase we extracted the following information from the given XML files of forum posts: author, the authors rank-

Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 153–157,
San Diego, California, June 16, 2016.

ing in the forum, time of submission and editing, number of likes and views, the body of the post, the subject, the thread associated to the post, and changeability of the text. For the representation of textual data (subject and body) we use both tf-idf and the word embedding representation of the data (Mikolov et al., 2013b; Mikolov et al., 2013a; Zhang et al., 2011). Skip-gram word embedding which is trained in the course of language modeling is shown to capture syntactic and semantic regularities in the data (Mikolov et al., 2013c; Mikolov et al., 2013a). For the purpose of training the word embeddings we use skip-gram neural networks (Mikolov et al., 2013a) on the collection of all the textual data (subject/text) of 65,514 posts provided in the shared task. In our word embedding training, we use the word2vec implementation of skip-gram (Mikolov et al., 2013b). We set the dimension of word vectors to 100, and the window size to 10 and we sub-sample the frequent words by the ratio $\frac{1}{10^3}$. Subsequently, to encode a body/subject of a post we use tf-idf weighted sum of word-vectors in that post (Le and Mikolov, 2014). The features are summarized in Table 1. To ensure being inclusive in finding important features, stop words are not removed.

2.2 Automatic Triage

The Random Forest (RF) classifier (Breiman, 2001) is employed to predict the users mental health states (green, red, amber, and crisis) from the posts in the ReachOut forum. A random forest is an ensemble method based on use of multiple decision trees (Breiman, 2001). Random forest classifiers have several advantages, including estimation of important features in the classification, efficiency when a large proportion of the data is missing, and efficiency when dealing with a large number of features (Cutler et al., 2012); therefore random forests fit our problem very well. The validation step is conducted over 947 labeled instances, in a 10xFold cross validation process. Different parameters of random forests, including the number of trees, the measure of split quality, the number of features in splits, and the maximum depth are tuned using cross-validation. In this work, we use Scikit implementation of Random Forests (Pedregosa et al., 2011).

Our results on the training set show that incorporation of unlabeled data in the training using label propagation by means of nearest-neighbor search does not increase the classification accuracy. Therefore, the unlabeled data is not incorporated in the training.

For the comparison phase, we consider multi-class Support Vector Machine classifier (SVM) with radial basis function kernel as a baseline method (Cortes and Vapnik, 1995; Weston and Watkins, 1998).

3 Results

Our results show that random forests have significant success over SVM classifiers. The 4-ways classification accuracies are summarized in Table 3. The evaluations on the test set for the random forest approach are summarized in Table 3.

3.1 Important Features

Random Forests can easily provide us with the most relevant features in the classification (Cutler et al., 2012; Breiman, 2001). Random Forest consists of a number of decision trees. In the training procedure, it can be calculated how much a feature decreases the weighted impurity in a tree. The impurity decrease for each feature can be averaged and normalized over all trees of the ensemble and the features can be ranked according to this measure (Breiman et al., 1984; Breiman, 2001). We extracted the most discriminative features in the automatic triage of the posts using mean decrease impurity for the best Random Forest we obtained in the cross-validation (Breiman et al., 1984).

Our results shows that from the top 100 features, $\frac{88}{100}$ were related to the frequency of particular words in the body of the post, $\frac{4}{100}$ were related to the posting/editing time (00:00 to 23:00) and the day in the month (1^{st} to 31^{th}), $\frac{4}{100}$ were indication of the author and author ranking, $\frac{2}{100}$ were related to the frequency of words in the subject, $\frac{1}{100}$ was the number of views, and $\frac{1}{100}$ was the number of likes a post gets.

The top 50 discriminative features, their importance, and their average values for each class are provided in Table 3.1. We have also presented the inverse document frequency (IDF) to identify how

Features Extracted from ReachOut forum posts		
Feature	Description	Length
Author	One hot representation of unique authors in 65755 posts.	1605
Ranking of the author	One hot representation of the author category.	25
Submission time	Separated numerical representations of year, day, month, and the hour that a post is submitted to the forum.	4
Edit time	Separated numerical representations of year, day, month, and the hour that a post is edited in the forum.	4
Likes	The number of likes a post gets.	1
Views	The number of times a post is viewed by the forum users.	1
Body	Tf-idf representation of the text in the body of the post.	55758
Subject	Tf-idf representation of the text in the subject of the post.	3690
Embedded-Body	Embedding representation of the text in the body of the post.	100
Embedded-Subject	Embedding representation of the text in the subject of the post.	100
Thread	One hot representation of the thread of the post.	3910
Read only	If the post is readonly.	1

Table 1: List of features that have been used in the automatic triage of ReachOut forum posts

Classifiers / Features	Random Forest Classifier	SVM Classifier
Tf-idf features	$71.28\% \pm 2.9\%$	$42.2\% \pm 3.1\%$
Embedding features	$71.26\% \pm 4.0\%$	$42.2\% \pm 4.0\%$

Table 2: The average 4-ways classification accuracies in 10xFold cross-validation for the random forest and support vector machine classifiers tuned for the best parameters on two different sets of features. Embedding features refer to use of embeddings for the body and the subject instead of tf-idf representations.

Methods	Accuracy	Non-green vs . green accuracy
Random Forest & tf-idf features	79%	86%
Random Forest & embedding features	78%	86%

Table 3: The results of evaluation over 241 test data points.

much information each word has encoded within the collection of posts (Robertson, 2004). Many interesting patterns can be observed in the word usage of each class. For example, the word 'feel' significantly more often occurs in the red and crisis posts. Surprisingly, there were some stop-words among the most important features. For instance, words 'to' and 'not', on average occur in green posts $\frac{1}{2}$ of times of non-green posts. Another example is the usage of the word 'me', which occurs more frequently in non-green posts. Furthermore, the posts with more 'likes' are less likely to be non-green.

Subject: As indicated in Table 3.1 posts which have word 're' in their subjects are more likely to belong to the green class.

Time: As shown in Figure 1 and Table 3.1 the red posts on average are submitted on a day closer to the end of the month. In addition, the portion of red and crisis message posts in the interval of 5 A.M. to 7 A.M. was much higher than the green and amber posts.

4 Conclusion

In this work, we explored the automatic triage of message posts in a mental health forum. Using Random Forest classifiers we obtain a higher triage accuracy in comparison with our baseline method, i.e. a mutli-class support vector machine. Our results showed that incorporation of unlabeled data did not increase the classification accuracy of Random Forest, which could be due to the fact that Random Forests themselves are efficient enough in dealing with missing data points (Cutler et al., 2012). Furthermore, our results suggest that employing full vocabularies would be more discriminative than using sentence embedding. This could be interpreted as the importance of occurrence of particular words rather than particular concepts. In addition, taking advantage of the capability of Random Forest in the estimation of important features in classification, we explored the most relevant features contributing in the automatic triage.

Acknowledgments

Fruitful discussions with Meshkat Ahmadi, Mohsen Mahdavi, and Mohammad Soheilypour are gratefully acknowledged.

				Green	Amber	Red	Crisis
Rank	Feature	Importance	IDF	Average value	Average value	Average value	Average value
1	body: you	0.068	0.004	16.912 ± 24.13	3.941 ± 10.238	2.728 ± 7.077	2.432 ± 8.605
2	body: to	0.059	0.012	4.948 ± 5.739	8.964 ± 6.83	9.408 ± 7.265	9.552 ± 7.666
3	subject: re	0.053	0.03	3.904 ± 1.871	3.6 ± 1.843	3.246 ± 2.637	2.802 ± 2.112
4	$\#oflikes$	0.027	-	0.749 ± 1.104	0.353 ± 0.882	0.155 ± 0.453	0.154 ± 0.489
5	body: just	0.021	0.007	2.632 ± 6.332	6.69 ± 8.962	8.349 ± 9.697	8.702 ± 9.992
6	body: feeling	0.02	0.009	0.884 ± 3.463	2.527 ± 7.216	4.227 ± 9.606	3.188 ± 5.812
7	body: don	0.02	0.008	1.407 ± 4.523	3.998 ± 7.302	4.996 ± 7.599	9.074 ± 13.873
8	body: me	0.019	0.006	2.73 ± 6.471	7.848 ± 10.056	9.321 ± 11.432	8.264 ± 8.207
9	$\#ofviews$	0.016	-	96.016 ± 53.53	95.372 ± 50.9	92.158 ± 53.715	113.735 ± 56.293
10	body: know	0.016	0.007	1.55 ± 4.957	3.976 ± 7.806	4.863 ± 7.615	8.218 ± 11.262
11	body: want	0.015	0.008	0.548 ± 2.587	3.253 ± 7.431	3.875 ± 8.172	5.29 ± 8.699
12	body: anymore	0.013	0.013	0.063 ± 0.734	0.523 ± 2.578	2.594 ± 5.881	4.709 ± 9.327
13	body: do	0.013	0.007	1.987 ± 5.58	4.339 ± 7.226	4.741 ± 7.275	6.123 ± 8.322
14	body: and	0.011	0.009	5.629 ± 6.389	7.953 ± 7.687	10.007 ± 7.579	6.749 ± 5.905
15	body: negative	0.011	0.012	0.117 ± 1.354	1.184 ± 4.354	2.583 ± 6.404	4.446 ± 8.769
16	body: it	0.01	0.007	6.89 ± 9.562	10.607 ± 10.575	9.079 ± 9.527	7.56 ± 8.055
17	post hour (1-24)	0.01	-	9.922 ± 4.325	9.474 ± 4.135	9.118 ± 4.585	8.615 ± 4.159
18	body: my	0.01	0.007	5.137 ± 8.414	9.722 ± 10.703	10.303 ± 10.178	7.928 ± 10.775
19	body: the	0.01	0.011	4.744 ± 5.5	6.667 ± 6.064	5.95 ± 5.578	6.513 ± 6.729
20	body: for	0.01	0.008	4.418 ± 7.1	3.894 ± 5.61	3.274 ± 5.427	6.135 ± 5.89
21	body: about	0.009	0.008	1.646 ± 4.452	3.567 ± 5.711	2.11 ± 4.567	2.149 ± 4.574
22	body: so	0.009	0.008	3.387 ± 6.759	4.95 ± 7.102	7.57 ± 9.347	5.02 ± 7.942
23	body: this	0.009	0.008	2.624 ± 5.609	2.849 ± 5.489	5.302 ± 5.768	5.046 ± 6.633
24	post day (1-7)	0.009	-	15.25 ± 8.407	15.719 ± 8.625	15.3 ± 8.907	17.436 ± 8.217
25	edit day (1-7)	0.009	-	15.25 ± 8.407	15.719 ± 8.625	15.3 ± 8.907	17.436 ± 8.217
26	body: can	0.009	0.006	3.436 ± 7.302	4.297 ± 6.909	6.333 ± 7.913	12.029 ± 12.095
27	body: but	0.008	0.006	3.588 ± 6.988	7.376 ± 9.226	5.354 ± 7.634	8.245 ± 10.021
28	body: not	0.008	0.007	2.274 ± 5.459	5.037 ± 8.02	4.504 ± 7.172	3.901 ± 6.398
29	body: get	0.008	0.006	1.672 ± 4.627	3.552 ± 6.559	4.505 ± 8.02	4.35 ± 8.532
30	edit hour (1-24)	0.008	-	9.922 ± 4.325	9.474 ± 4.135	9.118 ± 4.585	8.615 ± 4.159
31	$author_x$	0.007	-	0.149 ± 0.357	0.072 ± 0.259	0.264 ± 0.443	0.308 ± 0.468
32	body: that	0.007	0.007	4.244 ± 7.687	5.513 ± 7.665	4.905 ± 7.357	3.875 ± 6.288
33	body: of	0.006	0.008	3.954 ± 5.989	4.902 ± 6.235	5.014 ± 5.904	5.425 ± 6.389
34	body: when	0.005	0.008	1.689 ± 4.25	2.998 ± 5.77	2.779 ± 5.249	2.871 ± 4.733
35	body: even	0.005	0.008	0.993 ± 3.499	1.513 ± 4.099	2.699 ± 5.337	4.37 ± 8.633
36	body: have	0.005	0.005	4.081 ± 7.854	6.196 ± 8.662	6.415 ± 8.511	5.191 ± 7.057
37	body: cant	0.005	0.013	0.033 ± 0.764	0.693 ± 4.004	1.589 ± 4.911	0.25 ± 1.091
38	body: all	0.005	0.006	1.866 ± 5.437	3.487 ± 6.37	3.691 ± 7.05	2.804 ± 6.987
39	subject: into	0.004	0.187	0.099 ± 0.511	0.391 ± 0.941	0.838 ± 1.249	0.728 ± 1.201
40	body: what	0.004	0.008	1.813 ± 4.463	2.725 ± 4.901	2.778 ± 4.744	2.577 ± 5.045
41	body: everything	0.004	0.01	0.262 ± 1.903	0.64 ± 2.8	1.726 ± 4.957	1.376 ± 3.576
42	body: $username_x$	0.004	0.016	1.096 ± 4.881	1.394 ± 5.164	0.938 ± 3.523	1.608 ± 4.565
43	body: in	0.004	0.009	3.467 ± 6.878	3.311 ± 4.559	4.241 ± 5.246	3.175 ± 4.247
44	body: feel	0.004	0.007	1.477 ± 4.989	3.145 ± 6.323	5.187 ± 8.689	3.746 ± 6.598
45	body: try	0.004	0.009	0.683 ± 3.816	1.465 ± 4.957	1.46 ± 3.793	1.902 ± 4.574
46	body: anything	0.004	0.007	0.541 ± 3	1.602 ± 4.745	2.195 ± 5.751	4.237 ± 10.067
47	body: am	0.004	0.008	1.162 ± 5.241	1.655 ± 4.619	2.523 ± 5.922	1.642 ± 4.584
48	body: at	0.004	0.007	2.033 ± 5.47	3.349 ± 6.469	3.661 ± 6.051	4.058 ± 6.735
49	body: with	0.004	0.01	2.029 ± 4.01	3.189 ± 5.024	2.679 ± 3.776	1.591 ± 2.872
50	body: safe	0.004	0.012	0.342 ± 2.802	0.163 ± 1.801	0.662 ± 3.339	2.907 ± 6.549

Table 4: The 50 most discriminative features of posts and their mean values for each class of green, amber, red, and crisis, which are ranked according to their feature importance. For the words we have also provided their IDF.

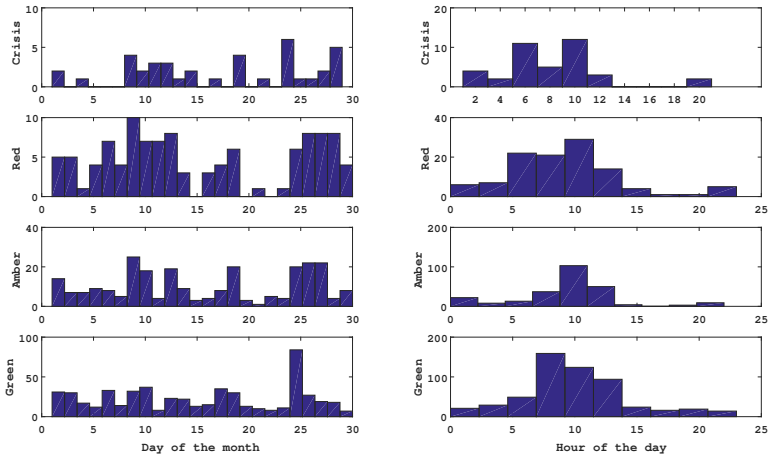

Figure 1: Histogram of message posting time distribution for each mental health state (crisis, red, amber, and green). The left plots show distribution of posts in days of the month (1-31) and the right plots show the distribution of the hours of the day.

References

Yoshua Bengio, Aaron Courville, and Pierre Vincent. 2013. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. 1984. *Classification and regression trees*. CRC press.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Adele Cutler, D Richard Cutler, and John R Stevens. 2012. Random forests. In *Ensemble Machine Learning*, pages 157–175. Springer.

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Tami L Mark, Rosanna M Coffey, Rita Vandivort-Warren, Hendrick J Harwood, et al. 2005. Us spending for mental health and substance abuse treatment, 1991-2001. *Health Affairs*, 24:W5.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830.

Stephen Robertson. 2004. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520.

TB Üstün, Joseph L Ayuso-Mateos, Somnath Chatterji, Colin Mathers, and Christopher JL Murray. 2004. Global burden of depressive disorders in the year 2000. *The British journal of psychiatry*, 184(5):386–392.

Jason Weston and Chris Watkins. 1998. Multi-class support vector machines. Technical report, Citeseer.

Wen Zhang, Taketoshi Yoshida, and Xijin Tang. 2011. A comparative study of tf* idf, lsi and multi-words for text classification. *Expert Systems with Applications*, 38(3):2758–2765.