Classification of mental health forum posts

Glen Pink † **Will Radford** ‡ **Ben Hachey** †
†ə-lab, School of Information Technologies
University of Sydney
NSW 2006, Australia
‡Hugo Australia
58-62 Kippax Street
NSW 2010, Australia

{glen.pink,ben.hachey}@sydney.edu.au
wradford@hugo.ai

Abstract

We detail our approach to the CLPsych 2016 triage of mental health forum posts shared task. We experiment with a number of features in a logistic regression classification approach. Our baseline approach with lexical features from a post and previous posts in the reply chain gives our best performance of 0.33, which is roughly the median for the task.

1 Introduction

The CLPsych 2016 shared task requires the triage of forum posts from the ReachOut.com forums, a support forum for youth mental health issues. The triage task centres on directing forum moderators to posts which required the most immediate attention (Calvo et al., 2016). For this task, a set of posts from the forum are each annotated with one of the labels *crisis*, *red*, *amber* or *green*, which indicate decreasing degrees of urgency of moderator addition. All unlabelled posts are made available for systems.

This task follows other studies of social media discourse as it relates to clinical psychology (Thompson et al., 2014; Schwartz et al., 2014; Coppersmith et al., 2015; Schrading et al., 2015). Analysis of ReachOut.com posts is interesting as posts are made by young individuals who have originally come to the forum seeking some kind of help, but over time may participate in several different capacities. Typically most users will initially need support, but this need may substantially increase or decrease over time; users may also support each other or use the forums for activity unrelated to mental health.

Our approach to this task was primarily focussed on implementing a straightforward baseline and experimenting with a few ideas derived from experience looking at the data in detail. While the data itself is definitely sequenced, we choose not to model this as a sequence problem, primarily because we expect the meaningful sequences to be fairly short: typically users either create new posts that are generally relevant to the original post in a thread, or reply to a specific post.

We further motivate this local post comparison by considering the annotation flowchart distributed with the data. Many labelling decisions are affected by whether the user's state is considered to be the same, or if their condition has gotten worse. Key to this task is capturing change in author language, and identifying how this reflects a change in their state-of-mind and change of condition.

We implement a feature set based on basic post features and author history and thread context, using the sequence of replies that lead to a post as the context for that post. We experiment with a number of additional features, but our baseline approach provides our best result of 0.33, which puts our performance at the median overall.

2 Features

We make use of post lexical features, author history and thread history for classification.

2.1 Preprocessing

Prior to extracting features, we perform some basic preprocessing on post text. We unescape HTML entities, remove images and replace emoticons with the

Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 180–182,
San Diego, California, June 16, 2016.

name of the emoticon to simplify processing. We remove blockquotes entirely, as we want extracted features to be from the content of the current post. We tokenise using the NLTK TweetTokenizer, as we expect the web forum text to be fairly casual and similiar to the Twitter domain for the purposes of tokenisation.

2.2 Lexical features

We extract unigrams and bigrams as post features, and continue to use this feature space for the below contexts.

2.3 Reply chain features

Instead of using the sequence of posts in a thread as context, we make use of the chain of replies to a post as the context for that post. We make use of two posts in that context: the most recent post before the current post that has the same author as the current post, and the most recent post to the current post. We retrieve unigrams and bigrams for these posts. We then extract three different types of features: the intersection of unigrams and bigrams with the current post; those that occur in the current post but not the previous post; and those that occur in the previous post but not the current. Note that there are separate feature spaces for author posts and non-author posts.

2.4 Unused features

We experimented with a number of features which did not improve results. These include use of n-gramfeatures from the first post in thread of the post; use of lemmas instead of words; cosine similarity between post bag-of-words; and thread type. We manually identify these thread types for threads which have a substantially different structure to others, such as the Turning Negatives Into Positives and TwittRO. We identify 1 post as `game`, 2 as `media` (e.g. image threads), 5 as `semi-structured` and 5 as `short` (e.g. TwittRO).

3 Data and training

The released training corpus contains 65,024 posts, 947 of which are annotated with triage labels. For development, we split this into a train set of 797 posts and a development set of 250 posts. We use a scikit-learn logistic regression classifier, using a grid search over a regularization hyperparameters

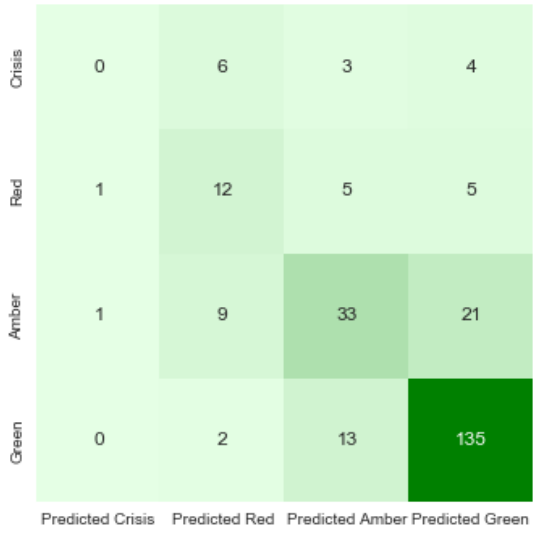

Figure 1: Confusion matrix on the development data.

Label	Precision	Recall	F-score
macro-avg	0.42	0.41	0.42
crisis	0.00 (0/0)	0.00 (0/13)	0.00
red	0.58 (14/24)	0.61 (14/23)	0.60
amber	0.68 (40/59)	0.62 (40/64)	0.65

Table 1: Final scores for run 1 settings on development data.

over 10-fold cross validation over the train set. Results on development data in Table 1. Figure 1 shows the confusion matrix, including green classifications. We note that a large number of confusions happen between amber and green, largely due to their larger representation in the data. For the full task we use the full 947 posts for training. The test set adds an additional 731 posts.

We experimented with using a cascaded classification approach, classifying crisis v. non-crisis, red v. non-red and amber v. non-amber in sequence, however this approach did not perform well. We also experimented with treating the task as a regression task, mapping *crisis* to a value of 1.0, *red* to 0.66, *amber* to 0.33, and *green* to 0.0. The idea is that we expect there to be a gradient to post severity rather than a distinct underlying set of 4 labels, and this gradient may be better modelled via a regression approach. Our implementation has lower results than our approach using discrete labels, but we consider this to be a possible direction for future approaches to this task.

run	score	accuracy	ngvg	ngvg accuracy
1	0.33	0.78	0.73	0.85
2	0.32	0.76	0.72	0.83

Table 2: Official results. *ngvg* is *non-green vs green*.

Label	Precision	Recall	F-score
crisis	0.00 (0/0)	0.00 (0/1)	0.00
red	0.61 (11/18)	0.41 (11/27)	0.49
amber	0.50 (23/46)	0.49 (23/47)	0.49

Table 3: Run 1 per-label scores.

4 Results

We submit two runs, for both L2 (run 1, with regularisation parameter $C = 1$) and L1 (run 2, with regularisation parameter $C = 100$) regularisation. Our official results are in Table 2, with per-label breakdowns of each run in Tables 3 and 4.

While other labellings fall outside the official metric for the shared task, we are interested in the performance of a system trained on only *non-green* vs *green* as opposed to all 4 triage labels. We run this configuration with the same settings as run 1. This configuration has an F-score of 0.80 on our development data, and a score of 0.82, which above our multiple label F-score of 0.73. This may be a useful setup for a two-stage classification or an actual implementation for ReachOut.com moderators.

5 Discussion

Run 1 performs at the median, and may be an informative baseline. Interestingly, many of the features that we explored decreased or did not significantly improve performance. This is possibly due to feature sparsity: the amount of training data is relatively small, and most of these features likely are not informative. We note that L2 regularisation gives our best performance, the data set is small, and L2 keeping more features from the training data helps compensate for feature sparsity better than L1 regularisation.

Notably, both of our runs returned very few crisis labellings: both returned 1 labelling which was incorrect. This is somewhat surprising, particularly as a label F-score of 0% is particularly penalised with a macro-averaged metric, however given the lack of instances for training this is not unreasonable.

label	Precision	Recall	F-score
crisis	0.00 (0/0)	0.00 (0/1)	0.00
red	0.52 (11/21)	0.41 (11/27)	0.46
amber	0.50 (23/46)	0.49 (23/47)	0.49

Table 4: Run 2 per-label scores.

6 Conclusion

We participated in the CLPsych 2016 shared task, providing a baseline approach using a small feature set that gave a near-median performance of 0.33. We look forward to continuing to work on this task.

References

Rafael A Calvo, M Sazzad Hussain, Kjartan Nordbo, Ian Hickie, David Milne, and P Danckwerts. 2016. Augmenting Online Mental Health Support Services. *Integrating Technology in Positive Psychology Practice*, page 82.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 Shared Task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado, June 5. Association for Computational Linguistics.

Nicolas Schrading, Cecilia Ovesdotter Alm, Ray Ptucha, and Christopher Homan. 2015. An Analysis of Domestic Abuse Discourse on Reddit. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2577–2583, Lisbon, Portugal, September. 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.

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.