CLPsych 2016 Shared Task: Triaging content in online peer-support forums

David N. Milne and **Glen Pink** and **Ben Hachey** and **Rafael A. Calvo**
University of Sydney
NSW 2006, Australia
{david.milne,glen.pink,ben.hachey,rafael.calvo}@sydney.edu.au

Abstract

This paper introduces a new shared task for the text mining community. It aims to directly support the moderators of a youth mental health forum by asking participants to automatically triage posts into one of four severity labels: *green*, *amber*, *red* or *crisis*. The task attracted 60 submissions from 15 different teams, the best of whom achieve scores well above baselines. Their approaches and results provide valuable insights to enable moderators of peer support forums to react quickly to the most urgent, concerning content.

1 Introduction

When facing tough times, the best support often comes from someone who has been through similar experiences (Pfeiffer et al., 2011). Forums are a simple way to facilitate such peer-support online, but when they involve vulnerable people and sensitive subject matter they require careful cultivation. There is growing evidence that online peer-support without professional input has limited effectiveness (Kaplan et al., 2011), and Kummervold et al. (2002) obtained almost unanimous feedback from forum users that professionals should actively participate or offer a safety net of passive monitoring.

The need for human moderation raises concerns of cost and scalability. This provides opportunity for text mining and NLP to augment human moderators by allowing them to focus on the individuals and posts that most urgently require their attention. For example, affect detection could locate emotionally charged posts (Calvo and D'Mello, 2010), and Yin et al. (2009) could identify malicious users. For the domain of mental health, De Choudhury et al. (2013) could prioritize clinically depressed individuals, and O'Dea et al. (2015) could help moderators respond quickly to suicidal ideation.

There has recently been a great deal of research that mines social media texts for mental health, but most have been isolated investigations. This paper introduces a new shared task for researchers to collaborate on and concretely compare what does and does not work. It releases a dataset of forum posts that have been manually annotated with how urgently they require a moderator's attention.

To our knowledge, the only other shared task involving social media and mental health is Coppersmith et al. (2015), who aim to detect depression and PTSD on Twitter. Other shared tasks have used data that is easier to de-identify: Pestian et al. (2012) focus on emotion detection within anonymized suicide notes, while Pradhan et al. (2014) and their predecessors focus on making clinical records easier to digest and understand.

The remainder of the paper is structured as follows. The next section describes ReachOut: an online community of Australian youth that provides both data and motivation. Section 3 describes the dataset extracted from these forums and the annotation process. Section 4 summarizes the methods and common themes of participating teams, and Section 5 contains their results. Our use of public yet sensitive data raises complex ethics issues that are addressed in Section 6. The final section describes some of the opportunities and challenges that remain unexplored and invites readers to participate.

Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 118–127,
San Diego, California, June 16, 2016.

2 The ReachOut forums

ReachOut.com is an Australian non-profit established in 1996 to support young people. It offers on-line resources about everyday topics like family, school and friendships, as well as more difficult issues such as alcohol and drug addition, gender identity, sexuality, and mental health concerns. About 1 in 3 young people in Australia are aware of the site (Metcalf and Blake, 2013), and it received about 1.8 million visitors in 2014 (Millen, 2014). In a survey conducted in 2013, approximately 77% of visitors reported experiencing high or very high levels of psychological distress, which indicates that the site is reaching people in need (Metcalf and Blake, 2013). 46% of these distressed visitors reported feeling more likely to access (for the first time) professional support after their visit.

Much of this success is due to the strong on-line community that has developed around ReachOut, thanks to a lively peer-support forum. This offers a safe, supportive environment for 14-25 year-olds to anonymously share their personal experiences.

Maintaining this environment and ensuring it remains a positive place to be requires a great deal of effort. ReachOut employs several senior moderators full-time, and also recruits and trains new young people each year as volunteer peer moderators. Collectively, this *Mod Squad* listens out for anything that might require attention, responding when needed with encouragement, compassion and links to relevant resources. In extreme cases they will occasionally redact content that is overly distressing or triggering, or where the author has jeopardized their own safety and anonymity. There is an escalation process to follow when forum members might be at risk of harm. Not all of the moderators' actions are so dire however; often they step in to congratulate someone for making progress, or simply to keep conversation flowing and build rapport.

3 Data and annotation

The ReachOut Triage Shared Task dataset consists of 65,024 forum posts written between July 2012 and June 2015. The data is structured in XML and preserves all metadata such as when the post was made, who authored it, and where it fits in the navigational structure of boards, threads, replies and quotes. We discuss the ethical considerations of using such sensitive yet public data in Section 6.

The vast majority posts are left unannotated, to provide a testbed for unsupervised and semi-supervised approaches such as topic modelling, co-training and distant supervision. A subset of 1,227 posts were manually annotated by three separate judges (the first three authors of the paper) using a semaphore pattern to indicate how urgently they require a moderators attention:

- **Crisis** indicates that the author (or someone they know) is in imminent risk of being harmed, or harming themselves or others. Such posts should be prioritized above all others.
- **Red** indicates that a moderator should respond to the post as soon as possible.
- **Amber** indicates that a moderator should address the post at some point, but they need not do so immediately.
- **Green** identifies posts that do not require direct input from a moderator, and can safely be left for the wider community of peers to respond to.

The annotation task began with the judges discussing the first $\sim$200 posts and arriving at a collective decision for each, guided by an informal annotation and triage criteria provided by Reachout. At that point the judges were able to formalize their decision process into the flowchart shown in Figure 1. This illustrates some of the complexity and subjectivity involved in the task: the judges (and future algorithms) have to consider both the textual content of the post and the sentiment behind it (e.g. that a post is *red* because it describes *current distress*), and also the trajectory of how authors follow up on their own previous concerning posts (e.g. that a post is *amber* because a prior situation has not worsened, but is also not entirely resolved).

Within the annotation system, posts were always viewed in the full context of how they were found in the live forum, rather than as an independent chunk of text. Posts were annotated against the flowchart to capture both the semaphore annotation and a more detailed sub-annotation. They could also be annotated as *ambiguous* if they fell outside the logic provided by the flowchart.

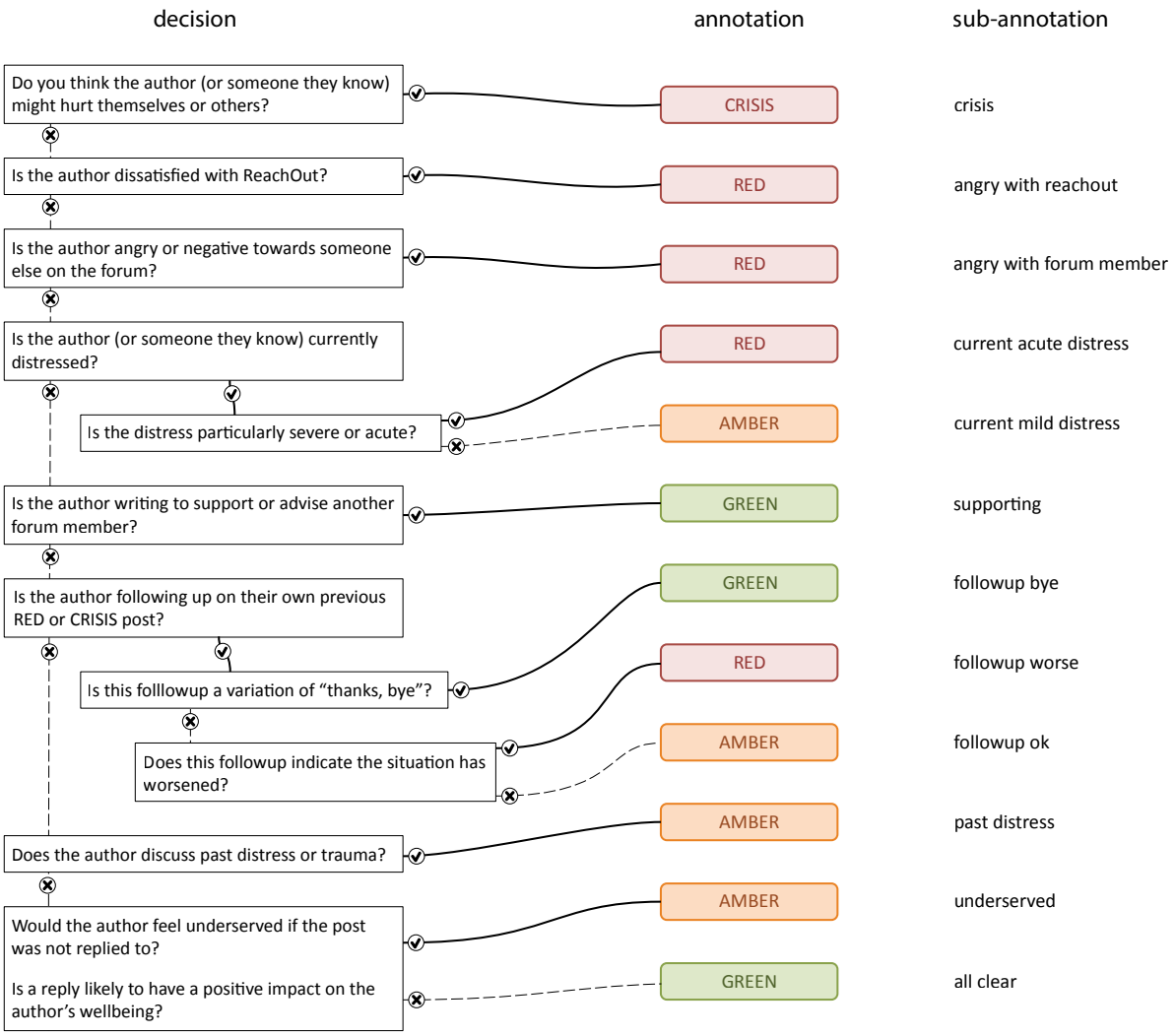

Figure 1: The triage annotation decision tree

After settling on this decision tree, the judges annotated each of the remaining posts independently. Inter-annotator agreement was then measured over these posts, excluding 22 that had been labelled as *ambiguous* by at least one of the judges. Over 977 cases (and four possible labels), the three judges achieved a Fleiss's Kappa of 0.706 and pairwise Cohen's Kappa scores ranging between 0.674 and 0.761. Viera and Garrett (2005) would interpret this as substantial agreement. Finally the judges met in person to resolve any remaining disagreements and ambiguous cases and arrive at a consensus.

Table 1 shows the final distribution of labels across the annotated portion of the dataset, and how it was split into 947 posts for training and 280 for testing. The posts were not stratified or distributed randomly, but were instead split on a particular date (the 28th of May 2015). Consequently the distribution of labels across the two sets is not entirely

	train	%	test	%
crisis	39	4	1	0
red	110	12	27	11
amber	249	26	47	19
green	549	58	166	69
total	947		280	

Table 1: Distribution of labels across training and testing data.

even, which makes the task somewhat more challenging and realistic. It also ensures that features can be extracted from the behaviour leading up to each post without accidentally contaminating training data with testing data.

4 Shared task submissions

Teams were given roughly 4 weeks from being provided with the training data to submitting test results. Teams were permitted to submit a maximum of 5 runs. We received 60 submissions from the 15 teams participating in the task. In this section we look at the various approaches to the task, and what techniques were and were not successful. First we briefly describe the top-performing approaches, and then summarise techniques used across systems.

4.1 Top systems

The top three systems achieved similar performance via very different approaches.

Kim et al. (2016) base their approach on SGD classifiers with a small feature space, varying several different aspects of implementation. There features consist of only TF-IDF weighted unigrams, and post embeddings using sent2vec (Le and Mikolov, 2014). Their best run was an ensemble of three classifiers which, in contrast to other teams, were trained on the 12 sub-annotation labels (e.g. *current acute distress*) as opposed to the 4 coarse labels. They find that this substantially increases *red* recall and *amber* precision, this suggests a better assignment of labels around the *red*/*amber* boundary. They incorporate a classifier which make sentence-level predictions, summing the distributions across sentences to select the label for a post.

Malmasi et al. (2016) implement a meta-classifier approach. Base linear SVM classifiers are constructed from a larger feature space than the other top-performing systems, they generate these base classifiers for both the target posts as well as preceding and following posts. These base classifiers are in turn used to train a meta-classifier, which is extended to a Random Forest of meta-classifiers. They find that Random Forests outperform SVMs at the meta-classifier level, but there is some performance variation between classifiers which they expect is due to the randomness inherent in training Random Forests. Despite the lower result, their RBF-kernel SVM meta-classifier still performs well, suggesting robustness of this approach.

Brew (2016) experiment with leveraging unlabelled data, but their baseline RBF-kernel SVM achieves a better score than any of their more elaborate approaches. Features used were TF-IDF weighted unigrams and bigrams, author type, post kudos, and whether a post is the first in its thread. They provide analysis in their system description paper, one observation is that the official metric may give unstable results which happen to overly benefit their implementation in this instance. Accuracy results in Section 5 may support this, as the accuracy of this system is slightly below the other top systems, but even across unofficial metrics this is still one of the top-performing approaches.

4.2 General approaches

Systems generally used a logistic regression or SVM classifier, or an ensemble of these classifiers. Most systems learned coarse-level labels only and used a relatively straightforward learning setup.

Successful approaches use several different types of features: as well as features derived from post content, we find systems include features derived from post metadata and larger forum structure.

4.2.1 Post content features

Systems extract typical features from post subjects and body text. Most systems preprocess the text to handle HTML entities, and extract unigram and bigram features, potentially using lemmatised tokens. Better performing systems weight these n-grams with TD-IDF (Kim et al., 2016; Brew, 2016), or incorporate embeddings. Top-performing (Malmasi et al., 2016) make use of further n-gram features, adding character n-grams, token skip-grams and POS n-grams to the above.

Lexicons, particularly the Linguistic Inquiry and Word Count (Pennebaker et al., 2015) lexicon, are used as measure of emotion (Cohan et al., 2016) and sentiment (Malmasi et al., 2016). Cohan et al. (2016) additionally leverage DepecheMood (Staiano and Guerini, 2014) to identify emotions associated with a post, and the MPQA subjectivity lexicon (Wilson et al., 2005) to distinguish between objective and subjective posts. In particular, Cohan et

al. (2016) apply these lexicons to the final sentence in a post, in an effort to capture the final mental state of the user, particularly where it relates to self-harm in lengthy posts that do not otherwise indicate self-harm. Zirikly et al. (2016) use the NRC Word-Emotion Association Lexicon (Mohammad and Turney, 2013) for emotion-word features.

Cohan et al. (2016) generate LDA topics of each post. Amiri et al. (2016) generate 30 topics over the full ReachOut.com corpus as well as the `reddit.com/r/Depression` subreddit. They then similarly use post topics as features.

Some approaches incorporate sentiment techniques into classification. Zirikly et al. (2016) label sentences with sentiment, using CoreNLP (Manning et al., 2014), and use counts of each sentiment as features, Shickel and Rashidi (2016) make use of sentiment labelling in a similar way. Almeida et al. (2016) add sentiment dictionaries.

Semi-structural features are included by Zirikly et al. (2016). One simple feature is the count of user names mentioned in posts. Other features capture repeated syntax, such as a popular thread that asks users to systematically turn negatives into positives. Deeper syntactic features are included by Malmasi et al. (2016), as are Brown cluster features. Wang et al. (2016) make explicit use of emoticons.

4.2.2 Post metadata features

Participants made only limited use of post metadata. Author ranking (the role of the author on the site) and kudos were the most used elements; followed by times posts were created and edited, the thread and board they belong to, and the number of times they are viewed.

4.2.3 Forum structure features

Most systems made little use of forum structure and hierarchy other than using thread ID as a simple feature. Malmasi et al. (2016) make use of posts before and after the target post; Pink et al. (2016) use post reply chains as a source of features; and a number of systems generate features from posts in context or aggregated features, such as the number of posts in a thread (Cohan et al., 2016). Brew (2016) add a feature which indicates whether a post is the first in a thread, which may be a useful straightforward feature, given how the data was annotated.

Most systems do not consider unlabelled posts. As mentioned, Cohan et al. (2016) and Amiri et al. (2016) build LDA models over the data. Zirikly et al. (2016) experiment with a semi-supervised SVM.

5 Results

In this section we only consider the best run for all teams. Readers are encouraged to refer to the individual system description papers for results of all runs.

5.1 Metric

The official metric for the shared task is macro-averaged F-score, because it gives more weight to the infrequent yet more critical labels than a micro-average. Identifying a metric that appropriately targets downstream requirements is difficult, particularly as desired recall is different across labels: a lower precision may acceptable for a higher recall *crisis* labelling, but not for *amber*. Brew (2016) provide some analysis of the stability of this metric. We note that ordering results by accuracy produces a fairly similar ordering.

5.2 Official results

The official scores are listed Table 2. It additionally reports scores gained by treating *crisis*, *red* and *amber* as a single *flagged* label against *green*, and by treating *crisis* and *red* as a single *urgent* label against *amber* and *green*. The participants' top systems are compared against to a straightforward baseline system that uses unigrams and bigrams as features, and a default scikit-learn (Pedregosa et al., 2011) logistic regression classifier.

Results are close across different approaches: three teams tie for first place, the next two teams are behind by only a few instances. The median is 0.34.

We note that the *crisis* label only occurs once in the test data and none of the systems successfully detected it. This has a large impact on the official macro-average metric; for example, if we disregard this label from Kim et al. (2016), the score would be 0.63. Fortunately all systems suffer the same disadvantage so the relative comparisons remain fair, although it is possible that systems optimised for *crisis* labels may have been slightly disadvantaged. We expect that a more sophisticated evaluation metric is required to handle this minimally represented class:

team	official	acc	flagged	flagged acc	urgent	urgent acc
Kim et al. (2016)	**0.42**	**0.85**	0.85	**0.91**	0.62	0.91
Malmasi et al. (2016)	**0.42**	0.83	**0.87**	**0.91**	0.64	**0.93**
Brew (2016)	**0.42**	0.79	0.78	0.85	**0.69**	**0.93**
Cohan et al. (2016)	0.41	0.80	0.81	0.87	0.67	0.92
Desmet et al. (2016)	0.40	0.80	0.80	0.87	0.62	0.92
Opitz (2016)	0.37	0.79	0.76	0.85	0.50	0.91
Zirikly et al. (2016)	0.36	0.77	0.78	0.85	0.60	0.90
Rey-Villamizar et al. (2016)	0.34	0.77	0.79	0.86	0.51	0.89
Pink et al. (2016)	0.33	0.78	0.73	0.85	0.48	0.90
Asgaria et al. (2016)	0.32	0.79	0.76	0.86	0.40	0.90
Amiri et al. (2016)	0.31	0.73	0.68	0.80	0.50	0.90
Wang et al. (2016)	0.30	0.73	0.76	0.83	0.48	0.89
Almeida et al. (2016)	0.29	0.74	0.68	0.82	0.51	0.88
Shickel and Rashidi (2016)	0.23	0.76	0.79	0.86	0.09	0.83
Franco-Penya and Sanchez (2016)	0.13	0.42	0.58	0.60	0.36	0.75
baseline	0.31	0.78	0.75	0.86	0.38	0.89

Table 2: Official results for the CLPsych 2016 shared task. *official* is crisis, red and amber macro-averaged F-score, *acc* is accuracy, *flagged* is crisis + red + amber, *urgent* is crisis + red (against amber + green). Top results are bolded.

team	crisis	red	amber	green
Kim	0.00	**0.65**	0.61	**0.94**
Malmasi	0.00	0.58	**0.69**	0.93
Brew	0.00	**0.65**	0.61	0.88
Cohan	0.00	0.59	0.64	0.90
Desmet	0.00	0.57	0.63	0.90
Opitz	0.00	0.48	0.62	0.89
Zirikly	0.00	0.51	0.58	0.89
Rey-Villamizar	0.00	0.43	0.58	0.90
Pink	0.00	0.49	0.49	0.89
Asgaria	0.00	0.41	0.56	0.90
Amiri	0.00	0.44	0.48	0.85
Wang	0.00	0.36	0.55	0.87
Almeida	0.00	0.40	0.48	0.87
Shickel	0.00	0.10	0.59	0.90
Franco-Penya	0.00	0.16	0.24	0.62
baseline	0.00	0.39	0.53	0.90

Table 3: Per-label F-scores for each run in Table 2.

if any system had correctly labelled the one *crisis* instance with reasonable precision, it would likely drastically outperform other systems.

For the best run of each team, we evaluate on each label and include the results in Table 3. Generally, systems perform well on *green*, and a substantial portion of performance is reliant on the *red*/*amber* decision. The is reflected in the *flagged* result in Table 2, sorting by this column would result in a substantially different ordering.

Many of the top-performing approaches are notably different, however there are some interesting comparisons that can be made. Both Kim et al. (2016) and Brew (2016) are successful with only a small feature space. The latter system demonstrates that it is useful to consider not just the usual n-grams, but also custom features such as author type, kudos, and whether a post is first in the thread. It is interesting that the top teams achieved similar results. A larger exploration of the feature space may help identify those that are most useful.

Overall, the scarcity of crisis posts made full labelling a difficult task. However, the teams were able to achieve good scores for the *flagged* and *urgent* binary classification problems. These are promising results for supporting ReachOut's moderators.

6 Ethical considerations

In this paper we have built a shared task around publicly available data. Even though the data is already freely accessible, it needs to be treated with care and respect because it involves sensitive subject matter. The process of obtaining consent to release it to the research community was by no means straightforward. In this section we describe some of the questions and concerns that were raised in discussions with our own ethics committee, in the hope that they might be helpful to other researchers undertaking similar work. These discussions were guided by

the Australian National Statement on Ethical Conduct in Human Research; obviously each researcher should seek out the corresponding legislation relevant to them (which may differ from our own), and follow recommendations of relevant authorities.

6.1 What is the potential for harm, and how can it be minimized?

The National Statement describes a balance between benefit and risk; that any risk of harm must be offset or justified by likely benefits (either for the participants themselves or the wider community). We identified three groups of participants to whom this annotation and release of data might cause harm: to the researchers who annotated the data, to the researchers with whom the data is shared, and to the people who authored the content.

The first two groups were easily catered for, by ensuring that the researchers were aware of the potentially distressing and triggering nature of the content, and providing appropriate access to care (i.e. location-specific helplines).

The third group is of much greater concern. While these forum members have already shared their data publicly, our annotations serve to single out the most distressed and vulnerable individuals among them. Disclosing their identities could cause serious distress, and may undermine their willingness to seek help in future. Fortunately these forum members are instructed by ReachOut to keep themselves safe and anonymous, and the moderators described in Section 2 respond to and actively redact any identifying information that is inadvertently shared.

To further protect this anonymity, participating researchers were restricted from contacting contacting individuals within the dataset (i.e. via the forums), cross-referencing individuals with the dataset against any other datasets or social media accounts, or making any other attempt to identify individuals. They were also not permitted to publish any portion of the dataset (e.g. example posts) other than summary statistics, or share it with anyone else. Future users of the dataset will have the same restrictions.

6.2 Should the data be redacted?

Another possible strategy for minimising potential harm is to redact the data to remove any identifying information. This is difficult to do for public social media data, because any structure or terms that remain can be searched on and compared to reconstruct it. Counter-intuitively, the more accessible data is, the more difficult it is to share safely.

Zimmer (2010) provides a cautionary tale in which private Facebook data was shared inadvertently, despite researchers' honest efforts to protect it. The previous CLPsych shared task (Coppersmith et al., 2015) provides another example of a dataset that remains re-identifiable despite redaction. It gathered tweets from participants who self-indicated that they were suffering from depression and post-traumatic stress, and redacted them by hashing usernames and any other readily identifiable information. And yet, for many individuals there likely remains enough text to cross-reference against twitter archives. Consequently recipients of this data had to sign a privacy agreement stating they would make no attempt to re-identify them.

A safer example is Sumner et al. (2012), who shared a dataset of twitter profiles matched to self-reported ratings for the *big five* personality traits and the *dark triad* of anti-social personality traits. Here the data is more aggressively redacted by only retaining basic statistics and frequencies of terms found in the Linguistic Inquiry and Word Count (LIWC) lexicon. This obviously limits researchers to using only a narrow set of predefined features.

Another strategy would be to encode all content such that researchers could count the relative frequencies of all terms without being able to read or understand them. This allows greater freedom than Sumner et al. but is still very limiting. For example, researchers would not be able to cross-reference terms against external vocabularies or bootstrap other sources of data (Section 4.2.1), or even perform their own lemmatisation. It would also make error analysis difficult if not impossible.

In summary, it does not seem possible to render public data truly non-identifiable without greatly hindering research. Fortunately our ethics committee felt that the anonymous nature of the ReachOut forums provided good protection of privacy. Their key remaining concern was that forum members might be identifiable if they reuse user names from other forms of social media. This motivated some of the restrictions described in Section 6.1.

6.3 Should consent be obtained?

Ideally any research involving human participants should be done with their full knowledge and consent. However, this dataset involves hundreds of distinct authors, to reduce the risk that the resulting algorithms would become over-fitted to any individuals writing style. Consequently obtaining consent individually for each participant would require an impractical investment of time. Additionally, our only means of contact would be via the forum, which is a place where many participants are only active for a short period of time to ask a specific question and then move on. Consequently, a great deal of valuable data would have been lost if we required consent from each individual participant.

Fortunately, the National Statement provides provisions for waiving the need for disclosure and consent when it is impractical to obtain it. For brevity, we will not exhaustively list all of the relevant requirements, but will instead focus on those that are particularly relevant for this research:

The first requirement is that *involvement in the research carries no more than low risk to participants*. As explained previously, the main risk here is the potential disclosure of sensitive information about the participants. Fortunately, the largely anonymous nature of ReachOut combined with the restrictions placed on researchers meant that this risk of disclosure was minimal.

Another pertinent requirement is that *there is no known or likely reason for thinking that participants would not have consented if they had been asked*. Given that the forum data is already widely shared and requires no special privileges to browse it, we argued that the participants appear to be comfortable allowing anyone to read their posts, as long as they can remain anonymous. Our focus then has been to ensure this anonymity is kept intact.

One last requirement is that *the benefits from the research justify any risks of harm associated with not seeking consent*. To our minds, this raises an obligation for the research to be more than merely an interesting text classification problem; that it must lead to something that is of direct benefit to the users and moderators of ReachOut. Consequently we are now working to build an accurate classifier from the insights gained during the shared task, and have integrated an early version of this triage system into the moderators suite of tools (Calvo et al., 2016). This system is already helping moderators respond quickly to urgent forum posts, and we hope to make it much more accurate in the near future.

7 Conclusions and future work

The CLPsych 2016 shared task was an interesting and difficult one. It asked participants to tackle the complex and somewhat subjective problem of prioritizing posts on a mental health forum, and elicited a broad array of algorithms and techniques.

The quantity and quality of participation has been excellent, and the organisers would like to thank teams for their engagement. The top performing teams performed well above the baseline, and made substantial progress on the task.

Participants were given limited time to hone their algorithms, so we hope they continue their work. There are many facets of the data still to explore, such as modelling the history and mental state of users, capturing structural and temporal data from the forum hierarchy, and further leveraging unlabelled data with semi-supervised or distantly supervised techniques. We will continue to work on and support this task and will be integrating ideas into the system used by ReachOut's moderators.

We invite interested researchers to join us on this challenging and worthwhile problem[1].

Acknowledgments

The authors thank ReachOut.com for providing the setting for the shared task; Kristy Hollingshead and Lyle Ungar for hosting and supporting the task at CLPsych; and participants for contributing to making this a successful task.

This research is funded by the Young and Well Cooperative Research Centre, an Australian-based international research centre that unites young people with researchers, practitioners, innovators and policy-makers from over 70 partner organisations. Together we explore the role of technology in young people's lives, and how it can be used to improve the mental health and wellbeing of young people aged 12-25. The Young and Well CRC is established

[1] Researchers can apply for access to the ReachOut triage dataset at `http://bit.ly/triage-dataset`

under the Australian Government's Cooperative Research Centres Program. Calvo is supported by an Australian Research Council Future Fellowship.

References

Hayda Almeida, Marc Queudot, and Marie-Jean Meurs. 2016. Automatic Triage of Mental Health Online Forum Posts: CLPsych 2016 System Description. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, San Diego, California, USA, June 16.

Hadi Amiri, Hal Daumé III, Meir Friedenberg, and Philip Resnik. 2016. The University of Maryland CLPsych 2016 Shared Task System. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, San Diego, California, USA, June 16.

Ehsaneddin Asgaria, Soroush Nasiriany, and Mohammad R.K. Mofrad. 2016. Textual Analysis and Automatic Triage of Posts in a Mental Health Forum. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, San Diego, California, USA, June 16.

Chris Brew. 2016. Classifying ReachOut posts with a radial basis function SVM. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, San Diego, California, USA, June 16.

Rafael A Calvo and Sidney D'Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37.

Rafael A Calvo, Sazzad M Hussain, David N Milne, Kjartan Nordbo, Ian Hickie, and Peter Danckwerts. 2016. Augmenting Online Mental Health Support Services. In Daniela Villani, editor, *Integrating Technology in Positive Psychology Practice*, chapter 4, pages 82–103. IGI Global.

Arman Cohan, Sydney Young, and Nazli Goharian. 2016. Triaging Mental Health Forum Posts. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, San Diego, California, USA, June 16.

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 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies (NAACL HLT 2015)*, page 31.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. In *ICWSM*, page 2.

Bart Desmet, Gilles Jacobs, and Véronique Hoste. 2016. LT3 shared task submission. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, San Diego, California, USA, June 16.

Hector-Hugo Franco-Penya and Liliana Mamani Sanchez. 2016. Predicting on-line users in psychological need on Reach out forums. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, San Diego, California, USA, June 16.

Katy Kaplan, Mark S. Salzer, Phyllis Solomon, Eugene Brusilovskiy, and Pamela Cousounis. 2011. Internet peer support for individuals with psychiatric disabilities: A randomized controlled trial". *Social Science & Medicine*, 72(1):54–62.

Sunghwan Mac Kim, Yufei Wang, Stephen Wan, and Cecile Paris. 2016. Data61-CSIRO systems at the CLPsych 2016 Shared Task. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, San Diego, California, USA, June 16.

Per E Kummervold, Deede Gammon, Svein Bergvik, Jan-Are K Johnsen, Toralf Hasvold, and Jan H Rosenvinge. 2002. Social support in a wired world: use of online mental health forums in Norway. *Nordic journal of psychiatry*, 56(1):59–65.

Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Proceedings*, pages 1188–1196.

Shervin Malmasi, Marcos Zampieri, and Mark Dras. 2016. Predicting Post Severity in Mental Health Forums. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, San Diego, California, USA, June 16.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Atari Metcalf and Victoria Blake. 2013. ReachOut.com Annual User Survey Results. `http://about.au.reachout.com/`

wp-content/uploads/2015/01/ReachOut.com-Annual-User-Survey-2013.pdf. Accessed:2016-04-02.

Doug Millen. 2014. ReachOut Annual Report 2013/2014. http://about.au.reachout.com/us/annual-reports-financials. Accessed:2016-04-02.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465.

Bridianne O'Dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on Twitter. *Internet Interventions*, 2(2):183–188.

Juri Opitz. 2016. System Description. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, San Diego, California, USA, June 16.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, November.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The Development and Psychometric Properties of LIWC2015. *UT Faculty/Researcher Works*.

John P Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, Kevin B Cohen, John Hurdle, and Christopher Brew. 2012. Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights*, 5(Suppl. 1):3.

Paul N. Pfeiffer, Michele Heisler, John D. Piette, Mary A.M. Rogers, and Marcia Valenstein. 2011. Efficacy of peer support interventions for depression: a meta-analysis. *General Hospital Psychiatry*, 33(1):29–36.

Glen Pink, Will Radford, and Ben Hachey. 2016. Classification of mental health forum posts. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, San Diego, California, USA, June 16.

Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. SemEval-2014 task 7: Analysis of clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, August.

Nicolas Rey-Villamizar, Prasha Shrestha, Thamar Solorio, Farig Sadeque, Steven Bethard, and Ted Pedersen. 2016. Semi-supervised CLPsych 2016 Shared Task System Submission. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, San Diego, California, USA, June 16.

Benjamin Shickel and Parisa Rashidi. 2016. Automatic Triage of Mental Health Forum Posts for the CLPsych 2016 Shared Task. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, San Diego, California, USA, June 16.

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 (Volume 2: Short Papers)*, pages 427–433, Baltimore, Maryland, June. Association for Computational Linguistics.

Chris Sumner, Alison Byers, Rachel Boochever, and Gregory J Park. 2012. Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, volume 2, pages 386–393. IEEE.

Anthony J Viera and Joanne M Garrett. 2005. Understanding interobserver agreement: the kappa statistic. *Family Medicine*, 37(5):360–363.

Chen-Kai Wang, Hong-Jie Dai, Chih-Wei Chen, Jitendra Jonnagaddala, and Nai-Wen Chang. 2016. Combining Multiple Classifiers Using Global Ranking for ReachOut.com Post Triage. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, San Diego, California, USA, June 16.

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.

Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, 2:1–7.

Michael Zimmer. 2010. "But the data is already public": on the ethics of research in Facebook. *Ethics and information technology*, 12(4):313–325.

Ayah Zirikly, Varun Kumar, and Philip Resnik. 2016. The GW/UMD CLPsych 2016 Shared Task System. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, San Diego, California, USA, June 16.