Improving moderator responsiveness in online peer support through automated triage

Abstract

**Background:** Online peer-support forums require oversight to ensure they remain safe and therapeutic. As online communities grow, they place a greater burden on their human moderators, which increases the likelihood that people at risk may be overlooked. This paper evaluates the potential for machine learning to assist online peer support by directing moderators’ attention where it is most needed.

**Objective:** To evaluate the accuracy of an automated triage system, and the extent to which it influences moderator behaviour.

**Methods:** A machine learning classifier was trained to prioritize forum messages as green, amber, red or crisis depending on how urgently they require attention from a moderator. This was then launched as a set of widgets injected into a popular online peer-support forum hosted by ReachOut.com. The accuracy of the system was evaluated using a hold-out test set of manually prioritised messages. The impact on moderator behaviour was measured as response ratio and response latency, i.e. the proportion of messages that receive at least one reply from a moderator, and how long it took for these replies to be made. These measures were compared across three periods: prior to launch, after an informal launch, and after a formal launch accompanied by training.

**Results:** The algorithm achieved 84% f-measure in identifying content that required a moderator response. Between pre-launch and post-training periods, response ratios increased by 0.9, 4.4, and 10.5 percentage points for messages labelled as crisis, red and green respectively, but decreased by 5.0 percentage points for amber messages. Logistic regression indicated that the triage system was a significant contributor to response ratios for green, amber and red messages, but not for crisis messages. Between the same periods, response latency was significantly reduced (p’s<0.001) by factors of 80%, 80%, 77% and 12% for crisis, red, amber and green messages respectively. Regression analysis indicated that the triage system made a significant and unique contribution to reducing the time taken to respond to green, amber and red messages, but not to crisis messages, after accounting for moderator and community activity.

**Conclusions:** The triage system was generally accurate, and moderators were largely in agreement with how messages were prioritized. It had a modest effect on response ratios, primarily because moderators were already more likely to respond to high priority content prior to the introduction of triage. However, it significantly and substantially reduced the time taken for moderators to respond to prioritized content. Further evaluations are needed to assess the impact of mistakes made by triage algorithm, and how changes to moderator responsiveness impact the wellbeing of forum members.

**Keywords:** Social support; Triage; Classification; Natural language processing; Field studies

Introduction

When facing tough times, often the best people to turn to are those who have been through similar challenges, who can provide empathy and support that is grounded in personal experience [1]. Asynchronous text-based forums are a common method for
facilitating such peer support online, and have been shown to reduce symptoms of distress [2] and improve one’s sense of empowerment [3]. Their online nature allows individuals to access help at any time, from any location, with minimal cost and effort [4]. They can often be accessed anonymously, to mitigate the fear of stigma that can be a barrier to help-seeking, particularly among the young [5].

While online communities have much to offer, there exist potential pitfalls. For instance, they often lack the involvement of mental health professionals; community members interactions may be influenced negatively by an individual’s current mental health status [6]; social difficulties may be exacerbated online, due to missing social cues [7] or illness related disinhibition or disorganisation [8]. Further, without appropriate oversight risky and unsafe behaviours can emerge unchecked, such as the normalization of self-harm [9] and there is some evidence that online peer-support can be misused as a method of avoidance [10].

The involvement of mental health professionals and paraprofessionals (i.e. trained volunteers) likely improves the safety and therapeutic value of online peer support [11,12,13]. Users of these online communities appear to be amenable to oversight; for example Kummervold et al. [14] obtained almost unanimous feedback that mental health professionals should actively participate in online conversations and/or provide passive safety monitoring. Outside of peer-support, increased moderation of online communities in general has been shown to improve intention to participate [15] and the quality of contributions [16].

The key barriers to greater oversight of online peer support are cost and scalability. As these communities grow, they place a greater burden on their human caretakers, which increases the likelihood that people in need may be overlooked. To address these barriers, we describe an automated triage system that aims to guide human moderators to the people whose messages most urgently require their attention. We evaluate the accuracy with which this system identifies urgent content, and the extent to which it influences moderator behaviour. The evaluation of accuracy was conducted using a dataset of manually-prioritized forum messages. The evaluation of behaviour change was conducted as a quasi-experimental time-series analysis that tracked moderator behaviour for several years prior to, and one year following the introduction of the triage system.

These evaluations were made within the context of ReachOut.com, an Australian web-based youth mental health service that aims to intervene early in the onset of mental health problems in young people. ReachOut.com is well known and popular among its target population. Approximately 1 in 3 young people in Australia are aware of the site [17], and in 2015 the website received about 1.58 million Australian visitors [18]. 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 [19]. ReachOut.com’s online community includes a lively peer-support forum for 14-25-year-olds and ReachOut.com employs community managers, and recruits and trains youth volunteers collectively referred to as the Mod Squad. These volunteers monitor posts and respond as needed with encouragement, compassion and referrals to relevant resources. This paper investigates how to ensure the moderation provided by these professionals and volunteers remains scalable.

**Methods**

This section describes the triage system that was deployed, and the method by which it was evaluated.
The triage system

The triage system automatically prioritizes each new forum message as belonging to one of the following four categories:

- **Green** indicates a message can be safely left for the community to address, without requiring intervention from a moderator. Most forum messages are expected to fall into this category.
- **Amber** indicates a message that is important, but not urgent. It is appropriate for the moderator to wait and see if the community will respond to it. If no response is forthcoming, then a moderator should intervene.
- **Red** indicates a message that should be responded to as soon as possible, likely because the author is in distress, or the message content may be triggering to others.
- **Crisis** indicates the author or someone they know is at risk of harm. A moderator should respond as soon as possible and enact ReachOut.com's existing escalation protocol if appropriate.

The triage system is embedded into the forum, as a sidebar that provides a “to-do list” of crisis, red and amber messages that have not received a response from a moderator. Additional widgets are also embedded below each forum message, to display the priority assigned to it and whether it requires attention. Further details (including a screenshot) can be found in Multimedia Appendix 1.

The underlying algorithm that assigns these priorities relies on supervised machine learning, meaning that it learns automatically from examples of manually prioritized forum messages. The advantage of this approach—as opposed to manually specified rules—is that it can easily be adapted to new prioritisation schemes or new online communities simply by feeding it new examples of prioritized messages. The algorithm can also easily be maintained by learning from any corrections it is given by moderators. Further details of the algorithm—and the features it relies on—can be found in Multimedia Appendix 2.

**Evaluation of accuracy**

The evaluation of accuracy was conducted using a test set of manually prioritized messages that was withheld from any training or tuning of the algorithm. Both training and testing data were sourced from the Computational Linguistics and Clinical Psychology (CLPsych) 2016 shared task [20], which provided 1,227 messages (947 for training and 280 for testing) that were extracted from ReachOut.com forums and manually labelled (i.e. given one of the four priority labels described above) by three independent annotators. Reliability analysis indicated that these annotators reached a Fleiss’s Kappa of 0.706 and pairwise Cohen’s Kappa scores ranging between 0.674 and 0.761, indicating that though the task is somewhat subjective, a reasonable level of inter-rater agreement was achieved.

The primary evaluation metric used was f-measure, or the harmonic mean of recall (i.e. sensitivity) and precision (i.e. positive predictive value). As this is a multiclass classification problem where rare classes (e.g. red, crisis) are of greater interest. Scores were macro-averaged across the classes, after excluding the majority (i.e. green) class. We also report the algorithm’s performance in separating out content that is flagged for moderators’ attention (i.e. amber, red or crisis) from content that can safely be left for the community to address (i.e. green), and in separating urgent content (i.e. red, crisis) from content that can safely wait (i.e. green, amber). In both cases we report f-measure of the minority (i.e. flagged or urgent) class. Interested readers are directed to [20], for more detailed information about the dataset and evaluation metrics.
Evaluation of the impact on moderator response behaviour

A quasi-experimental time-series analysis was used to compare moderator behaviour before and after the introduction of the triage system. The primary measures of interest were response ratio and response latency: i.e. the proportion of messages that received at least one reply from a moderator, and the time elapsed between a message being created and receiving its first reply from a moderator. The analysis considers only replies made by moderators, to messages authored by peers (i.e. ordinary community members). All messages made by other types of forum members (e.g. trainee moderators and other affiliates) are ignored, as are messages made by moderators unless in response to a message from a peer.

These measures were compared across three distinct periods: pre-launch, post-launch, and post-training. The pre-launch period captured moderator behaviour prior to the introduction of the triage system - from 19th of July 2012 to 5th of August 2016 (i.e. 1478 days). The post-launch period captured the interim in which the triage system was available, but not accompanied by any guidance about how it should be used or how often it should be consulted. It extended from 5th of August 2016 to 27th of November 2016 (i.e. 114 days). The post-training period captured moderator behaviour when the triage system was fully integrated into their workflow, having been launched with a detailed training session for all moderators. It extended from 27th of November 2016 until 16th of October 2017 (i.e. 323 days).

During these periods, all priorities (i.e. green, amber, red or crisis labels) were assigned automatically. During the post-launch and post-training periods, they were assigned immediately as each message was created, and immediately revealed to users of the triage system. For the pre-launch period, they were assigned retroactively, simply by running the triage algorithm over the previously collected messages. These priority labels were not revealed to moderators until after the launch of the triage system, and thus were unable to influence their behaviour during the pre-launch period.

The volume of content that required moderation varied over time, and could potentially have a strong effect on moderator response behaviour. To account for this variability, we constructed histograms for each message that captured activity levels during 5 hour-long periods starting 2 hours prior to the message being created, the hour that the message was posted and ending 3 hours afterwards. These histograms record the number of messages created and the number of unique authors, separately for peers and moderators. Together these four histograms measure the load placed on moderators (i.e. the business of the forum) and the number of active moderators available to share that load.

For further details about how moderators were identified, how replies were tracked for messages, and how the activity histograms were constructed, please refer to Multimedia Appendix 3.

Statistical analysis

The triage algorithm was trained and evaluated using the scikit-learn Python Library, and the evaluation of moderator behaviour was conducted using IBM SPSS 25 statistical software for Mac. To assess whether the presence of the triage system and moderators’ training with it (i.e. period) were significant factors for moderator response ratio, direct logistic regression was performed separately for each priority level (green, amber, red, crisis), while also accounting for moderator and peer activity at and around the time the message was posted. The impact of the triage system and moderator’s training on moderator response latency was similarly evaluated using separate linear regression models for each priority level. Kolmogorov-Smirnov values and visual inspection indicated that response latency had a non-normal distribution, which was corrected
using log transformation. Where significant overall tests were reported, follow-up pairwise group differences were examined. Finally, to reduce the likelihood of type-I error, adjusted alpha levels were applied to account for multiple comparisons (p=0.004).

**Ethical approval**

This research was approved by the Human Research Ethics Committee at the University of Sydney as protocol 2016/064.

**Results**

**Algorithm accuracy**

Table 1 shows the performance of the triage algorithm using the dataset and official metrics provided by the CLPsych 2016 shared task [20]. The system outperformed previous task participants in macro-averaged f-measure, ranked 3rd in separating flagged (i.e. crisis, red or amber) messages from un-flagged (i.e. green) messages, and 4th in separating urgent messages (i.e. crisis or red) from non-urgent (i.e. amber or green) ones.

<table>
<thead>
<tr>
<th></th>
<th>Macro-averaged F1</th>
<th>Flagged F1</th>
<th>Urgent F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim et al. [21]</td>
<td>0.42</td>
<td>0.85</td>
<td>0.62</td>
</tr>
<tr>
<td>Malmasi et al. [22]</td>
<td>0.42</td>
<td><strong>0.87</strong></td>
<td>0.64</td>
</tr>
<tr>
<td>Brew [23]</td>
<td>0.42</td>
<td>0.78</td>
<td><strong>0.69</strong></td>
</tr>
<tr>
<td>Deployed to ReachOut.com</td>
<td><strong>0.65</strong></td>
<td>0.84</td>
<td>0.60</td>
</tr>
</tbody>
</table>

**Volume and severity of messages**

Table 2 provides a summary of the number of messages posted by peers within each evaluation period, and how the algorithm automatically prioritized them. As expected, within each period (i.e. each column) the majority of messages were green, and there were progressively fewer messages as priority increased.

<table>
<thead>
<tr>
<th></th>
<th>Pre-launch</th>
<th>Post-launch</th>
<th>Post-training</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N (%</td>
<td>n (%)</td>
<td>n (%)</td>
</tr>
<tr>
<td>Green</td>
<td>50,662 (80.1%)</td>
<td>7,808 (76.7%)</td>
<td>22,008 (70.0%)</td>
</tr>
<tr>
<td>Amber</td>
<td>8,835 (14.0%)</td>
<td>1,634 (16.0%)</td>
<td>6,955 (22.1%)</td>
</tr>
<tr>
<td>Red</td>
<td>3,390 (5.4%)</td>
<td>704 (6.9%)</td>
<td>2,170 (6.9%)</td>
</tr>
<tr>
<td>Crisis</td>
<td>382 (0.6%)</td>
<td>37 (0.4%)</td>
<td>320 (1.0%)</td>
</tr>
</tbody>
</table>

*Note: Totals may be greater than 100% due to rounding.*

Figure 1 provides a more detailed timeline, with a stacked histogram of the weekly volume of messages posted by peers, split into each of the four priorities. The number of messages posted each week fluctuated strongly, but there is a general upwards trend indicating that the online community was becoming busier. At its peak, the forum received 1063 messages from peers during the week starting on the 2nd of July 2017.
Figure 1: Weekly counts of prioritized messages from ordinary forum members

Proportion of messages that received a moderator response

Table 3 provides a summary of how often messages from peers received at least one reply from a moderator. The data is organised by triage-assigned message priority, and the columns indicate whether the triage system was deployed at the time, and whether training had been conducted.

Table 3: Proportion of messages that receive a reply from a moderator

<table>
<thead>
<tr>
<th>Severity</th>
<th>Pre-launch</th>
<th>Post-launch</th>
<th>Post-training</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>(%)</td>
<td>n (%)</td>
</tr>
<tr>
<td>Green</td>
<td>17,004</td>
<td>(33.6%)</td>
<td>1,687 (21.6%)</td>
</tr>
<tr>
<td>Amber</td>
<td>5,763</td>
<td>(65.2%)</td>
<td>976 (59.7%)</td>
</tr>
<tr>
<td>Red</td>
<td>2,568</td>
<td>(75.8%)</td>
<td>461 (65.5%)</td>
</tr>
<tr>
<td>Crisis</td>
<td>308</td>
<td>(80.6%)</td>
<td>30 (81.1%)</td>
</tr>
</tbody>
</table>

Within each period (i.e. column), there is a consistent progression in which likelihood of a message receiving a reply is proportional to the priority assigned to it. The response ratios for green and red messages decrease after the introduction of triage, but recover strongly after training. Response ratios for amber messages also decrease after the introduction of triage, but do not recover after training. Response ratios for crisis messages change very little across the periods.

Logistic regression was performed to investigate whether these differences in response ratio were significant, and to what extent they were due to the introduction of the triage system. Separate logistic regression models were completed for each severity label to predict the likelihood of a message receiving a moderator response. Each model contained 21 independent variables including whether the triage system had been launched at the time of the message, whether training had been conducted, and the level of activity of moderators and community members at and around the time of the message (i.e. the histogram data described previously).

For green messages, the model was significant $\chi^2 (22, N=80,478) = 11411, \ p<0.001$, indicating that it was able to distinguish between messages that did and did not receive a moderator response. The model as a whole (i.e. including all variables) explained...
between 13.3% (Cox and Snell $R^2$) and 18.2% (Nagelkerke $R^2$) of the variance in moderator response and correctly classified 70.2% of messages. In the final model (i.e. retaining only statistically significant variables) the strongest predictor of moderator response was whether training for the triage system had been conducted, which recorded an odds ratio of 1.77. This was followed by the number of moderators online at the time of the community member message (odds ratio of 1.1). This indicated that green messages were more likely to receive a response after the triage system was introduced and moderators had been trained to use it, after controlling for all other factors in the model.

For amber messages, the model was significant $\chi^2(22, N=17,424) = 1207.2$ $p<0.001$. It explained between 6.7% (Cox and Snell $R^2$) and 9.4% (Nagelkerke $R^2$) of the variance in moderator response and correctly classified 70.3% of messages. The strongest predictor was whether triage training had been conducted, recording an odds ratio of 1.8. Again, this was followed by the number of moderators online at the time of the community member message (odds ratio of 1.3). This indicates that amber community messages were 1.8 times more likely to receive a moderator response after triage had been introduced and moderators had been trained to use it, after controlling for all other factors in the model.

For red messages, the model was again significant $\chi^2(22, N=6264)=372.0$, $p<0.001$, and explained between 5.8% (Cox and Snell $R^2$) and 8.6% (Nagelkerke $R^2$) of the variance in moderator response and correctly classified 76.2% of messages. The strongest predictor of moderator response was the number of active moderators at the time the message was posted, recording an odds ratio of 1.5. This indicated that messages that were posted when there were more moderators online were slightly more likely to receive a response, after controlling for all other factors in the model. The second strongest predictor of moderator response was message period, specifically whether training for the triage system had been conducted (odds ratio of 1.3).

Finally, for crisis messages the model was also significant $\chi^2(22, N=739)=76.3$, $p<0.001$. It explained between 9.8% (Cox and Snell $R^2$) and 15.8% (Nagelkerke $R^2$) of the variance and correctly classified 81.2% of messages. The strongest predictor was the number of active moderators at the time the message was posted, recording an odds ratio of 1.7, followed by the number of community posts at the time the message was posted. These were the only two variables to contribute significantly to the regression model. This indicates that crisis messages posted when there were more active moderators were almost twice as likely to receive a response, after controlling for all other factors in the model. Neither the presence of the triage system nor training made statistically significant contributions to the model ($p's > 0.1$).

**Time taken for messages to receive a reply from a moderator**

Between pre-launch and post-launch periods, the median time taken for moderators to respond to crisis messages was reduced from 2 hours 23 minutes to 47 minutes (see Table 4). Reductions were also observed for red (2 hours 10 minutes to 42 minutes), amber (2 hours 9 minutes to 33 minutes) and green messages (from 38 to 22 minutes). Response latency also decreased between the post-launch and post-training periods for crisis (from 47 to 29 minutes), red (from 42 to 26 minutes) and amber (from 33 to 30 minutes), but increased for green messages (22 to 33 minutes). Cumulatively, the presence of the triage system and the training of moderators resulted in reductions of 80%, 80%, 77% and 12% for crisis, red, amber and green messages respectively.
Table 4: Time taken for moderators to respond to messages

<table>
<thead>
<tr>
<th></th>
<th>Pre-launch</th>
<th>Post-launch</th>
<th>Post-training</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>median (IQR)</td>
<td>median (IQR)</td>
<td>median (IQR)</td>
</tr>
<tr>
<td>Green</td>
<td>0:37:38 (10:24:35)</td>
<td>0:21:49 (6:54:12)</td>
<td>0:33:16 (10:35:09)</td>
</tr>
<tr>
<td>Amber</td>
<td>2:08:34 (11:33:11)</td>
<td>0:33:22 (4:05:39)</td>
<td>0:30:08 (4:48:54)</td>
</tr>
<tr>
<td>Red</td>
<td>2:09:45 (9:33:11)</td>
<td>0:42:17 (4:57:47)</td>
<td>0:26:25 (1:52:00)</td>
</tr>
<tr>
<td>Crisis</td>
<td>2:23:03 (7:23:54)</td>
<td>0:47:11 (1:40:34)</td>
<td>0:28:45 (1:15:50)</td>
</tr>
</tbody>
</table>

Further, the triage system reduced the variability of response latency for non-green messages. Between pre-launch and post-launch periods, the inter-quartile range of latency for crisis, red and amber messages was reduced by 77%, 48% and 65% respectively. Between pre-launch and post-training periods, the same measures were reduced by 83%, 80% and 58%. The inter-quartile range of latency for green messages decreased by 34% between pre-launch and post-launch, but remained steady (a 2% increase) between pre-launch and post-training.

Multiple regression analysis was used to further evaluate whether these differences in latency were statistically significant, and whether they remained after accounting for moderator and community activity at and around the time of the message post (note: adjusted p=0.004). Separate regression models were conducted for each priority level, and included the same variables as the models for response ratio (described above).

For green messages, the total variance explained by the model as a whole was 47.9% ($F_{21,28368} = 1243.7$, p<0.001). In the final model, the number of moderators online at the time the message was posted was the strongest predictor of moderator response latency ($\beta=-0.33$, p<0.001) indicating that the more moderators online, the shorter the latency in response time. Whether or not moderators had been trained to use the triage system also contributed significantly to the model indicating that response time reduced significantly from pre-launch to post-launch periods ($\beta=-0.09$, p<0.001).

For amber messages, the total variance explained by the model as a whole was 18.8% ($F_{21,12039} = 132.4$, p<0.001). Moderator activity during the hours before, at and after a community member posted to the website, as well as triage period were entered into the model. The number of active moderators at the time of the message made the strongest significant contribution to explaining moderator response latency ($\beta=-0.35$, p<0.001). The presence of the triage system also made significant contribution to the model ($\beta=-0.08$, p<0.001) indicating that response latency decreased from the pre-launch to post-launch periods.

For red messages, the total variance explained by the model as a whole was 24.9% ($F_{21,4746} = 76.2$, p<0.001). Moderator activity during the hours before, at and after a community member posted to the website were entered into the model as well as trial period. In the final model, the number of active moderators at the time of the message made the strongest significant contribution ($\beta=-0.40$, p<0.001), indicating that the more moderators online when a community member posted, the shorter the latency in response time. The presence of the triage system also made a statistically significant contribution ($\beta=-0.14$, p=0.001) with response latency significantly decreasing from pre-launch to post-launch periods.

For crisis messages, the total variance explained by the final model as a whole was 28.2% ($F_{21,577} = 12.2$, p<0.001). The model consisted of moderator and community activity in the hours before, at and after each message was posted. The strongest predictors of response latency were the number of active moderators at the time of the message ($\beta=-0.34$, p=0.001), the number of moderator posts ($\beta=-0.26$, p=0.001), and the number of
To look more closely at the effect of triage period planned comparisons with statistical correction (adjusted alpha = 0.004) comparing pre-launch and post-launch periods (post-launch, post-training) were conducted. These showed that messages were responded to more quickly during the post-training period (i.e. combining the triage system with appropriate training) if they were labelled amber (p<0.001) or red (p<0.001). The large apparent difference in response latencies for crisis messages was significant only at the trend level (p=0.007), likely due to only having 30 crisis messages in the post-launch period.

Similar comparisons between post-launch and post-training periods showed that messages were responded to more slowly in the post-training period if they were labelled green (p<0.001) but more quickly if they were labelled red (p<0.001), but the differences for amber and crisis messages were not significant (p's>0.05).

Finally, comparisons between pre-launch and post-training periods (i.e. combining the triage system with appropriate training) showed that messages posted during the later period received replies significantly faster for all severity labels (p's<0.001).

**Discussion**

This study evaluated a triage system in terms of the accuracy with which it automatically prioritized content in online peer support, and the extent to which it improved the responsiveness of human moderators to the prioritized content. The triage algorithm achieved high accuracy (84% f-measure) in identifying content requiring moderator response. Additionally, the combination of the triage system and appropriate training resulted in modest improvements to response ratio for all priority levels other than amber, and large reductions to response latency for all priority levels other than green. Overall, the observed reductions in response latency and variability of response latency for flagged messages indicate that the triage system supported online moderator as intended.

**Accuracy of the triage algorithm**

Over the hold-out test set, the triage algorithm was more accurate in separating out flagged (i.e. non-green) messages than it was in separating urgent (i.e. red, crisis) messages. Arguably, the first of these boundaries is more important, because this determines which messages enter the sidebar (see Multimedia Appendix 1). A low recall here could cause moderators to miss posts that they should pay attention/respond to, while a low precision would increase their workload by filling the sidebar with low-priority messages. In contrast, mistakes made in separating urgent and non-urgent messages only effect the ranking of messages within this sidebar.

In addition to the above results, it was encouraging to observe a high level of agreement between moderators and the triage algorithm during the pre-launch period (i.e. the first column of Table 3), where there was a clear progression in which the likelihood of moderators responding to messages was proportional to the priority assigned by the algorithm. This agreement was in no-way due to the triage algorithm influencing moderator behaviour, since during severity labels were assigned retroactively for the pre-launch period (i.e., after the moderators' responses had been made). Conversely, it was also in no-way due to moderators influencing the algorithm, since none of the features or training data used by the algorithm were based on moderator behaviour. Thus, we are able to show that the moderators and the triage algorithm arrived at similar decisions independently.
Impact on moderator response ratio

As mentioned previously, the triage system resulted in only modest improvements to moderator response ratio. This is understandable, given that moderators already prioritized responding to urgent messages prior to the introduction of the triage system. Evidence of this response hierarchy is seen in the clear progression within the pre-launch period (i.e. the first column of Table 3), where crisis messages were more likely to receive replies than red messages, which were responded to more often than amber messages, which were in turn more likely to receive a response than green messages. Because the moderators were already behaving as desired, the potential to improve response patterns after launch and/or training was limited.

In fact, the introduction of the triage system appears to have initially had a detrimental effect, with response ratios dropping between pre-launch and post-launch periods for all severity labels other than crisis. This may be due to the unfamiliarity of the system and the lack of training given during this initial post-launch period. Fortunately, response ratios recovered in the post-training period, such that the end result (i.e. between pre-launch and post-training periods) was an increase across all severity labels other than amber. The reduction in response ratio for amber messages is likely due to the way the triage interface allows these messages to be clearly marked as resolved when the community has rallied around them. In the post-training period, moderators were specifically instructed to only respond to amber messages if they had been overlooked by the community.

The strongest predictor for response ratio was whether or not moderators had been trained to use the triage system (for green and amber posts) or the number of moderators active at the time of the message (for red and crisis). It is important to note that the introduction of the triage system (with training) significantly increased the likelihood that green, amber, and red messages received a moderator response after accounting for moderator and community activity. The system was not a significant predictor for the response ratios of crisis messages.

Impact on moderator response latency

Prior to the introduction of the triage system, moderators took a median time of roughly two hours to respond to non-green messages, while green messages tended to be responded to either quickly, or not at all. The informal launch of the triage system led to large reductions in response latencies for amber, red and crisis messages and modest reductions for green messages. The formal launch and associated training led to further reductions for amber, red and crisis messages, and a substantial increase for green messages. Cumulatively (i.e. between pre-launch and post-training periods) response ratios dropped by ~80% for crisis, red and amber messages, and 12% for green messages. Additionally, the variability (i.e. inter-quartile range) of response rates dropped by ~60% for amber messages and ~80% for red and crisis messages, but rose by 2% for green messages.

Across all priority levels, the strongest predictor of response latency was the number of moderators online at or around the time of the post. The formal launch of the triage system coincided with an influx of new volunteers, and their introduction had a large impact on response latency. However, it is important to note that the largest decreases in response latencies occurred between the pre-launch and post-launch periods (i.e. prior to this influx of new moderators). Additionally, the triage system and/or moderators’ training with it were shown to be a significant predictor for the reduced response latencies across all priority levels. Overall, the large reductions in response latency and variability of response latency for priority messages indicate that the triage system supports moderator behaviour as intended.
**Limitations**

This evaluation of moderator behaviour has made a key assumption that the decisions made by the triage algorithm are correct, and that it is desirable that moderators follow its recommendations. As mentioned above, the evaluation over the CLPsych 2016 dataset and the agreement observed between moderators and the algorithm during the pre-launch period are encouraging. Nevertheless, it is important that future work evaluate the impact of the mistakes that any automated system will inevitably make, particularly given the inherent subjectivity of the prioritization task.

False negatives—messages that the system does not prioritize highly enough—have the potential to be particularly problematic. As moderators come to rely on the triage system, they may neglect to look elsewhere for high-priority content, and consequently the system may counter-productively increase the chance that it is overlooked. We employed two strategies to decrease the likelihood of false negatives. The first was to reweight the triage algorithm so that it prioritizes recall ahead of precision; i.e. that it cares more about ensuring that any potentially urgent messages are included in the sidebar than ensuring that the sidebar includes only urgent messages. The second strategy was to deploy a tool that allowed ordinary community members to manually identify urgent content and add it to the triage system. The interface that enabled this crowdsourcing is described at the end of Multimedia Appendix 1. We would encourage other practitioners to provide similar safety nets if they adopt or develop an automated triage system.

Another limitation is that the analysis focused exclusively on the behaviour of moderators, and has not considered replies and support offered by community members, affiliates or moderators-in-training. Moderators may systematically avoid responding to messages that receive a strong response from the community, or encourage the community to be more self-reliant by withholding intervention when it is safe to do so. It is also possible that moderators will direct trainees to respond to urgent content rather than resolving it themselves. Our analysis has—necessarily, for the purposes of triage evaluation—assumed that a message is not resolved until it receives a reply from a moderator and that prompt replies are universally desirable. It would be more accurate to say that moderators should be kept aware of concerning content and be ready to intervene if necessary, rather than that they should always intervene as quickly as possible.

A related limitation is that the analysis focused on moderator behaviour, without evaluating the impact this has on the community and its members. While the underlying aim of increasing moderator responsiveness to urgent content has been to improve the safety and therapeutic value of the online community, we have not measured such outcomes directly. For future studies, it will be extremely valuable to survey forum users before and after the introduction of a triage system such as this to assess whether the changes to moderator behaviour were noticeable, and whether this lead to better outcomes or an improved sense of support.

A final, more technical limitation is that some of the variables considered by our statistical models are not entirely independent. Some of the variables introduced when modelling both response ratio and latency relate to the activity levels of moderators at and around the time of a message being posted. Intuitively, greater numbers of active moderators are likely to lead to more prompt responses. Unfortunately, moderator activity levels aren’t entirely independent from response ratio or latency, since there has to be at least one active moderator during one of these windows in order for a message to receive a reply. Consequently, these models may overestimate the impact of moderator activity. The same variables are also not entirely independent from the presence of the triage system, since it is possible it has contributed to the activity levels of moderators.
Comparison with prior work

To our knowledge, this paper is the first evaluation of automated triage in online peer-support that has focused on behaviour change; i.e. the system’s ability to influence human moderators and direct their attention to where it is most needed.

There is, however, a great deal of prior work related to the machine learning/computational linguistics aspects of this study. This includes the CLPsych 2016 shared task [20] in which 15 teams of researchers competed to develop the best algorithm for prioritizing forum messages from ReachOut.com. Our classification system was directly informed by the submissions of the top performing teams [21]-[23] and was trained and evaluated on the same dataset. Encouragingly, researchers have continued to work with this data and hone their algorithms, with Cohan et al. [24] significantly outperforming all previous participants. Additionally, a second edition of the shared task ran in 2017 and attracted 15 teams of researchers [25]. Thus advances to the algorithm have already been made and are available for integration in future deployments of the triage system.

Also closely related are reports from Huh et al [26] and Delort et al [27], who both developed machine learning systems to determine whether or not a forum message requires moderation. Both differ from our work by framing the problem as binary (a message either requires moderation, or does not), whereas we have framed it as a multiclass prioritization problem. Both focus exclusively on algorithm accuracy, and do not investigate moderator behaviour change.

In addition, there is a great deal of more broadly related work on the use of machine learning to detect undesirable content and behaviour online. For example, researchers have developed algorithms to detect hate speech [28,29], cyberbullying [30-32] and the grooming activities of paedophiles [33]. There has also been much progress recently in applying natural language processing to social media to gain insights into the authors’ state of mind, or to identify and diagnose individuals who could benefit from some form of psychological intervention or assistance (see [34] for a recent review). For example, researchers have developed algorithms to detect suicide ideation [35-38], depression [39-41], post-traumatic stress disorder [41,42] and the “dark triad” of antisocial personality traits [43,44]. This work bears strong similarities to the machine-learning component of this paper, and it is very likely that the features and techniques used could be applied successfully to triaging content in online peer support.

Conclusions

This paper has described a triage system that automatically prioritizes content in online peer support to augment human moderators and help them focus their efforts on the individuals and messages that have the greatest need. Through evaluation on a dataset of manually-prioritized forum messages, we have shown that the triage algorithm is largely accurate, particularly for the critical boundary separating content that moderators need to pay attention to from content they can safely leave for peers to address on their own. Through a long-term field study, we have shown that the triage system greatly reduces the time taken for moderators to respond to prioritized content, and that moderators are largely in agreement with the triage system about which messages they should prioritize responding to.

Our underlying aim of this work in improving the responsiveness and scalability of human moderation in online peer support is to increase the safety and therapeutic value...
of these communities. However, further evaluations are needed to establish whether this is the case. Additionally, in the near future we plan to investigate the impact of mistakes made by the triage algorithm and how best to encourage moderators and peers to provide an additional safety net (and additional training data) through manual prioritization. In addition, there are many opportunities to incorporate related work in social media mining to improve the accuracy of the triage algorithm that we hope to explore.

Conflicts of interest

The authors confirm that there are no known conflicts of interest associated with this publication including any financial, personal or other relationships with other people or organizations within three years of beginning the submitted work that could inappropriately influence, or be perceived to influence this work.

References


[12] Webb M, Burns J, Collin P. Providing online support for young people with mental


