AUTOMATIC DETECTION OF PSYCHOLOGICAL DISTRESS INDICATORS AND SEVERITY ASSESSMENT IN CRISIS HOTLINE CONVERSATIONS

Maciej Pacula[1], Talya Meltzer[1], Michael Crystal[1], Amit Srivastava[1], Brian Marx[2,3]

(1) Raytheon BBN Technologies, 10 Moulton St, Cambridge, MA, U.S.A.
(2) National Center for PTSD at VA Boston Healthcare System, Boston, MA, U.S.A.
(3) Boston University School of Medicine, Boston, MA, U.S.A.

ABSTRACT

Psychological health disorders pose a growing threat to society. Disorders such as Depression, Post-Traumatic Stress Disorder (PTSD), and mild Traumatic Brain Injury (mTBI), are often under-diagnosed and under-treated. Crisis hotlines are often the last resort for people who, from the lack of proper treatment, are considering suicide or intend to harm themselves or others. This paper describes a system that automatically analyzes online crisis hotline chats to (1) extract fine-grained distress indicators that map to Diagnostic and Statistical Manual of Mental Disorders (DSM) IV codes, and to (2) perform triage classification based on the severity of distress. For distress detection, we present several approaches which leverage annotator rationales and dialogue structure to improve classification performance, demonstrating significant gains over a state-of-the-art approach from literature. For triage classification, we demonstrate early detection capability for the most severe triage code. We evaluate our work on a large corpus of chats from the U.S. Department of Veterans Affairs' online Crisis Hotline.

Index Terms— Psychological Distress, Crisis Hotline, Text Classification, Annotator Rationales, Support Vector Machines

1. INTRODUCTION

Psychological health disorders pose a growing threat to individuals, their family members and to society at large [4], [5]. Disorders such as Depression, Post-Traumatic Stress Disorder (PTSD), and mild Traumatic Brain Injury (mTBI), are often under-diagnosed and under-treated [3]. Crisis hotlines are often the last resort for people who, from the lack of proper treatment, are considering suicide or intend to harm themselves or others. Unfortunately, the responders on crisis hotlines are not necessarily mental health professionals, and as such may miss subtle indicators of distress potentially leading to delayed intervention for those who need help the most.

We present a system for automatically detecting fine-grained distress indicators in online crisis hotline chats. Specifically, our system combines state-of-the-art NLP and machine learning techniques to: (1) extract fine-grained psychological distress indicators/labels derived from Diagnostic and Statistical Manual of Mental Disorders (DSM) IV [1], and (2) assesses the severity of distress that can be used to *triage* individuals who should seek clinical help. For distress detection, we present several approaches which leverage annotator rationales and dialogue structure to improve classification performance, demonstrating significant gains over the state-of-the-art approach of Saleem et al [7]. For triage assessment, we implement and evaluate the approach used in [7], demonstrating that the resulting models are capable of detecting the most severe forms of distress early in the conversation, potentially leading to faster intervention if integrated into a (hypothetical) early warning system.

Relation to prior work: A lot of work on dialogue classification focuses on Dialogue Acts (DA) – identifying whether an utterance is a statement, question, greeting, agreement, disagreement, and so forth. Such classification is part of both understanding and managing a dialogue, by producing the next utterance. Different machine learning techniques were applied such as HMM [9] and more complex graphical models [13], SVM [10], combination of both SVM and HMM [11], and error correction on top of SVM output [12]. More recent work was done under the Partially Observable Markov Decision Processes (POMDP) framework for optimizing policies in Dialogue Management (DM), which also requires inferring probabilities of internal states [14], [15]. We are not familiar with any specific task for which dialogue classification was applied, other than producing a dialogue.

To the best of our knowledge, the only existing system for automatic detection of fine-grained psychological distress is described in [7]. Saleem et al use a multi-stage classification framework to extract fine-grained psychological distress indicators and assess treatment acuity based on users' forum posts. Our work is an extension of [7] to the dialogue domain, exploiting its unique structure. In addition, and unlike Saleem et al who worked with heavily moderated Internet forums, we are also able to include the most severe triage code in our evaluation.

2. CRISIS HOTLINE DATA

For our experiments, we used a corpus of 427 chat transcripts from the U.S. Department of Veterans Affairs'

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

Crisis Hotline, where Veterans and their families in crisis can connect with qualified responders to receive support. Hotline chats consist of series of utterances, with each utterance associated with the time it was posted and its author (responder vs. chatter). All transcripts have been de-identified and contain no protected health information (PHI). Table 1 lists corpus statistics.

In consultation with psychologists, we developed a codebook of 70 fine-grained psychological distress labels spanning PTSD, mTBI, and depression symptoms. Codes/labels were derived mostly from the DSM-IV guidelines [1] and span symptoms of PTSD, mTBI and depression. Our codes closely resemble those of [7], but with some unreliable and infrequent labels either removed or merged into other codes as recommended by our psychologists.

In the annotation process, a group of three psychologists independently annotated each Crisis Hotline chat with zero, one or multiple distress indicators from the codebook characterizing the psychological state of the author in accordance with the content of the chat. Additionally, psychologists highlighted contextual rationales to support their distress label annotation. Such rationales proved useful in related text classification tasks in literature [7], [8].

In addition to distress labels, each chat was also annotated with a triage code indicating treatment acuity/urgency of referral. The triage codes were: TR1 indicating current or imminent danger to self or others; TR2 indicating behavioural disturbances, distress, functional impairment and/or suicidal/homicidal ideation without any imminent danger to self or others; and TR3 where there is no evidence of current behavioural disturbance, distress or functional impairment [7]. To account for high precision and low recall of human annotators, we combined ratings from multiple annotators by computing their union (for the distress labels) and their consensus (for the triage code). To ascertain annotation quality, we measured the inter-annotator agreement using the Fleiss kappa [2]. The kappa was 0.64 for the distress labels and 0.67 for the triage codes, indicating good agreement for both categories [6].

Category		Train	Test
Transcripts		349	78
	Total Words	348K	77K
	Unique Labels	63	56
	Average Number of Labels per Transcript	5.3	5.9

Table 1: The Crisis Hotline corpus used in experiments.

3. BASELINE SYSTEM

As our baseline we implement the approach proposed by Saleem et al. for labeling fine-grained distress indicators in online forum posts [7], retrained on the Crisis Hotline data. Saleem's system translates the problem of multi-label distress classification into independent binary decisions, with a separate one-vs.-all SVM for each of the distress indicators. The classifiers are trained using a set of rich features: stemmed unigrams, normalized pronoun and punctuation counts, average sentence lengths, sentiment words from multiple categories (e.g. cognitive processes, social processes) and domain phrases derived from annotator rationales [7]. In our work, we also extract idiom features by checking constituent input n-grams against a lexicon of 20,646 English idioms and figurative phrases downloaded from various online sources, and representing each as a binary indicator variable. Unlike Saleem et al, we do not compute thread features as they only apply to internet forums. For training, feature extraction and inference, we treat each chat as a single body of text and do not break it down into constituent utterances.

4. DIALOGUE ADAPTATION

Detecting fine-grained distress indicators in dialogue presents several challenges for which the baseline system does not account. First, Crisis Hotline chats are composed of utterances whose meaning and information content w.r.t. distress indicators is influenced by the context – e.g. what question was asked immediately preceding the current utterance. By treating the chat as a single body of text, this context is lost. Furthermore, many distress indicators are triggered by specific utterances and largely uninfluenced by the rest of the conversation – treating both equally introduces significant noise. In Section 4.1 we propose a *turn-level classifier* capable of both capturing context and finding "trigger" utterances by modeling distress at the turn level.

The second challenge in dialogue classification is accounting for speaker role. A hotline responder asking about a specific form of distress does not necessarily mean the chatter exhibits that particular distress, whereas the chatter mentioning it on their own can be a good indication. In Section 4.2 we extend the turn-level classifier to capture speaker information. We then reduce problem dimensionality by capturing paraphrases in responder's utterances through clustering in Section 4.2.1.

4.1. Turn-level classification

We split each dialogue into consecutive pairs of responder-chatter utterances which we call *turns.* Neighboring utterances by the same speaker are combined so that each turn consists of exactly one responder utterance and exactly one chatter utterance. We use annotator rationales to obtain turn-level ground truth, by checking for which labels at least one annotator highlighted the turn as evidence.

Given the individual turns and their rationale-derived ground truth, we train distress models as in the baseline system but using turns as input. In the feature extraction step, we do not distinguish between responder/chatter utterances within each turn, treating them as a single body of text.

For evaluation, we pool scores from all turns within a conversation and combine them using the *max* and *sum* operators. Such combination captures the fact that some labels (e.g. *Anger* and *Sleep problems*) are triggered by specific utterances (*max*), while others (e.g. *Sadness* and *Despair*) are conveyed by the overall tone of the conversation (*sum*). The *sum* operator also has the desirable property of boosting classifier confidence if a particular distress indicator is triggered by more than one utterance.

The final score for a label is a weighted sum of *max* & *sum* of turn scores, with the weight optimized through cross-validation. Despite training on rationale-derived truth, evaluation is still performed using conversation-level labels assigned by the annotators so that both the turn-level classifiers and the baseline are evaluated on identical data using identical ground truth.

4.2. Accounting for speaker role

We further extend the turn-level classifier by extracting features separately for the responder and the chatter within each turn, and then concatenating the resulting feature vectors. This allows us to distinguish words and other features depending on whether they originated from the responder or the chatter.

4.2.1. Utterance clustering

Extracting separate features for the responder and the chatter doubles the feature space, which with limited training data might lead to poor generalization of the resulting classifier. In addition, many of the responders' utterances (especially questions) tend to be paraphrases which repeat across multiple conversations. To take advantage of this repetition and to reduce the feature space, we cluster responder utterances using the Affinity Propagation algorithm [16] with cosine similarity as the affinity measure. We then use cosine similarity to each of the cluster representative utterances as a feature. The cluster features are the only features extracted for the responder's utterance within each turn, while chatter's features do not change.

5. CLASSIFIER FUSION

Since the optimal detection approach might differ among indicators, we perform a late fusion step by computing a weighted sum of scores from the baseline and all the turn-level classifiers. We optimize weights through cross-validation on a per-label basis, allowing us to choose the best classifier (or a combination of classifiers) for each label in the codebook.

6. TRIAGE CLASSIFICATION

In addition to detecting fine-grained distress indicators, we perform coarse-grained triage classification of distress severity: immediate danger to self or others (TR1), evidence of psychological disturbance but no immediate danger (TR2) and no evidence of psychological disturbance (TR3). The most severe code, TR1, is of particular importance in the Crisis Hotline setting where the expediency of intervention could determine whether the intent to harm materializes into actual harm. Unlike the moderated internet forums of [7] where the most severe posts are removed, the Crisis Hotline corpus contains 51 (12%) conversations labeled as TR1, which makes the detection of TR1 an important aspect of our work.

We implement the SVM approaches to triage classification from [7]. Both train one-vs.-all SVMs for each of the triage codes, with the first approach using only unigrams and the second adding the binary presence/absence of the automatically detected distress indicators as features.

A novel contribution of our work is the evaluation of TR1 classifier's performance not only on the entire conversation, but also on the first 25%, 50% and 75% of the utterances. This allows us to assess whether the triage models can detect severe distress early, potentially enabling their use in automated early alert systems.

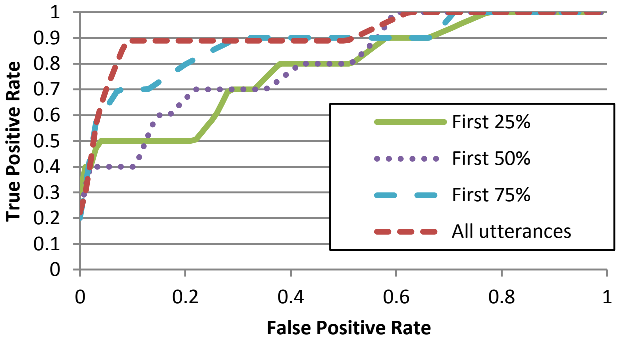

Figure 1: Receiver Operating Characteristic curves of the TR1 classifier working in a simulated early warning system where only the first 25%, 50% and 75% of utterances are available, with 100% (all utterances) shown for reference. The respective areas under the curves are 0.63, 0.77, 0.83 and 0.88.

7. EXPERIMENTAL RESULTS

As described in Section 2, we use a corpus of 427 Crisis Hotline chats for experimentation, which we randomly split into 349 training and 78 test chats. We further divide the training set into 5 cross-validation splits. We use the libSVM implementation of Support Vector Machines [17] with the RBF kernel. The kernel regularization and gamma parameters are tuned on the cross-validation set, with the final performance reported on the test set.

We evaluate performance using three metrics: AUC, AUC10 and F1. The first two are area metrics associated with the Receiver Operating Characteristic (ROC) curves, and measure the aggregate trade-off between the true and the false positive rates (TPR and FPR). AUC is computed by integrating the ROC curve between 0% and 100% FPR, while AUC10 is bounded to 10% FPR and normalized. AUC and AUC10 values can range from 0 (worst) to 1 (perfect classifier) with chance performance being 0.5 and 0.05, respectively. AUC10 is of particular interest in the domain of distress detection where the number of false positives has to be sufficiently low for the system to be useful. In addition to AUC and AUC10, we also report the F-measure (F1) equal to the harmonic mean of precision and recall. AUC10 was the metric we used when optimizing weights for the fused classifier.

We compute AUC, AUC10 and F1 independently for each distress label/triage code and report the un-weighted mean across the label set.

7.1. Distress classification

Table 2 shows distress classification performance. The baseline classifier, without any dialogue adaptations, achieved mean AUC, AUC10 and F1 of 0.79, 0.26 and 0.28, respectively. The single biggest performance improvement came from moving to turn-level classification, resulting in AUC of 0.84 (+6% relative), AUC10 of 0.41 (+58% relative) and F1 of 0.36 (+29% relative). We did not observe an improvement from extracting features separately for the responder's and chatter's utterances within each turn without clustering, which we suspect is due to the twofold increase in dimensionality. Clustering responder's utterances with the Affinity Propagation algorithm and using cluster similarity as features, however, improved the F-measure from 0.36 to 0.41, a 14% relative increase.

Finally, fusing the baseline and the proposed classifiers resulted in an AUC of 0.86, AUC10 of 0.45 and F1 of 0.40. Both AUC scores for the fused classifier were better than any single classifier alone, and the F-measure was only higher for the classifier with cluster features (0.40 vs. 0.41). Compared to the baseline, the fused classifier attained relative improvements of 9% (AUC), 76% (AUC10) and 43% (F1).

7.2. Triage assessment

Triage classification performance is shown in Table 3. Unlike [7], we do not observe significant differences from incorporating the automatically detected distress indicators as features. Both the baseline and the system incorporating distress indicators achieve the same AUC (0.88) and almost the same AUC10 (0.53 vs. 0.54), while the baseline achieves the best F-measure (0.67 vs. 0.61).

7.2.1. Early detection of severe distress

Figure 1 shows the Receiver Operating Characteristic of the TR1 classifier (unigram features only) in a simulated early warning system, where the classifier is working with only the first 25%, 50% and 75% utterances, with performance on all utterances shown for reference. Despite working with limited data, all classifiers perform well, with AUC scores ranging from 0.63 for the first 25% utterances to 0.83 for the first 75% (reference AUC using all utterances is 0.88). Of particular note is the fact that despite only having a fraction of the information, the TR1 classifier achieves a True Positive Rate of 0.7 with only a 0.3 False Positive Rate based on only the first quarter of utterances, which makes the use of our system feasible for early detection and automated alert of the most severe forms of distress.

Method		**Mean AUC**	**Mean AUC10**	**Mean F1**
Baseline Classifier		0.79	0.26	0.28
Turn-level classification		0.84	0.41	0.36
Separate responder/chatter features	No clustering	0.83	0.41	0.36
	Affinity Propagation	0.82	0.41	**0.41**
Fused Classifier		**0.86**	**0.45**	0.40

Table 2: Distress indicator detection performance. We computed AUC, AUC10 and F1 independently for each distress indicator and averaged the results. The table shows the averaged (mean) performance across distress indicators.

Method	**Mean AUC**	**Mean AUC10**	**Mean F1**
Unigrams	**0.88**	0.53	**0.67**
Unigrams + Distress Indicators	**0.88**	**0.54**	0.61

Table 3: Triage assessment performance. We computed AUC, AUC10 and F1 independently for each of the three triage codes and then averaged the results. The table shows the averaged (mean) performance across the triage codes.

8. CONCLUSIONS AND FUTURE WORK

We introduced several novel methods for detecting fine-grained distress indicators in dialogue interactions. In particular, we presented three approaches which (a) take advantage of annotator rationales to perform turn-level inference, (b) model speaker identity by extracting features separately for the different speakers in each turn and (c) cluster utterances to reduce dimensionality and better capture paraphrases. Finally, we introduced a fused classifier that combined the best of all the approaches on a per-label basis. We demonstrated the performance of our methods on real-world Crisis Hotline chats, where we achieved AUC, AUC10 and F1 of 0.86, 0.45 and 0.40, respectively. Those numbers represent relative improvements 9%, 76% and 43% over the approach of [7] on the same data.

For triage assessment, we applied the approach of [7] to our data and demonstrated the classifier's ability to detect the most severe triage code (TR1) early in the conversation. This could enable the use of the proposed approach in automated early warning systems, potentially expediting referral for individuals most at risk of harming themselves or others.

In the future, we intend to test our system on other dialogue sources from the psychological health domain. We would also like to extend our approach to better capture dialogue structure beyond individual utterances and turns, possibly by using a graphical model that captures both inter-turn and inter-label dependencies.

ACKNOWLEDGMENTS

This material is based upon work supported by the Defense Advanced Research Project Agency (DARPA) and Space and Naval Warfare Systems Center Pacific under Contract No. N66001-11-C-4094. The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

REFERENCES

[1] American Psychiatric Association. 2000. *Diagnostic and statistical manual of mental disorders (4th ed., text rev.)*. Washington, DC

[2] J. L. Fleiss. 1971. *Measuring nominal scale agreement among many raters.* Psychological Bulletin, 76(5):378-382.

[3] R. C. Kessler, et al. 1999. *Past-year use of outpatient services for psychiatric problems in the National Comorbidity Survey*. American Journal of Psychiatry, 156(1), 115-123.

[4] N. Breslau, G. C. Davis, P. Andreski, E. Peterson. *Traumatic Events and Posttraumatic Stress Disorder in an Urban Population of Young Adults*. Arch Gen Psychiatry. 1991;48(3):216-222.

[5] M. Olfson, S. C. Marcus, B. Druss, L. Elinson, T. Tanielian, H. A. Pincus. *National Trends in the Outpatient Treatment of Depression*. JAMA. 2002; 287(2):203-209.

[6] J. R. Landis and G.G. Koch. 1977. *The measurement of observer agreement for categorical data*. . Biometrics33 (1): 159–174.

[7] S. Saleem, R. Prasad, S. Vitaladevuni, M. Pacula, M. Crystal, B. Marx, D. Sloan, J. Vasterling, T. Speroff. *Automatic Detection of Psychological Distress Indicators from Online Forum Posts*. Proceedings of the 2012 International Conference on Computational Linguistics (COLING 2012), Mumbai, India

[8] O. F. Zaidan, J. Eisner and C. Piatko. 2008. *Machine Learning with Annotator Rationales to Reduce Annotation Cost.* Proceedings of the NIPS 2008 Workshop on Cost Sensitive Learning

[9] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol van Ess-Dykema, and Marie Meteer. *Dialogue act modeling for automatic tagging and recognition of conversational speech*. Computational Linguistics, vol. 26, pp. 339–373, 2000

[10] Raul Fernandez and Rosalind W. Picard. *Dialog Act Classification from Prosodic Features Using Support Vector Machines.* In Proc. Speech Prosody, 2002

[11] Dinoj Surendran and Gina-Anne Levow. *Dialog Act Tagging with Support Vector Machines and Hidden Markov Models*. In Proc. of Interspeech, 2006

[12] Yang Liu. *Using SVM and Error-correcting Codes for Multiclass Dialog Act Classification in Meeting Corpus*. In Proc. of Interspeech, 2006

[13] G. Ji and J. Bilmes. *Dialog act tagging using graphical models*. In Proc. of ICASSP, 2005.

[14] B. Thomson and S. Young. *Bayesian update of dialogue state: A POMDP framework for spoken dialogue systems*. Computer Speech and Language, vol. 24, no. 4, pp. 562–588, 2010.

[15] Milica Gasic, Filip Jurcıcek, Blaise Thomson, Kai Yu and Steve Young. *On-line policy optimisation of spoken dialogue systems via live interaction with human subjects*. ASRU, 2011

[16] Brendan J. Frey and Delbert Dueck. *Clustering by Passing Messages Between Data Points*. Science, vol. 315, pp. 972–976, 2007.

[17] Chang, Chih-Chung and Lin, Chih-Jen. *LIBSVM: A Library for Support Vector Machines*. ACM Trans. Intell. Syst. Technol., vol. 2(3), 27:1—27:27, 2011