Automatic Detection of Psychological Distress Indicators in Online Forum Posts

Shirin Saleem[*1], Maciej Pacula[*1], Rachel Chasin[*2], Rohit Kumar[*1], Rohit Prasad[*1], Michael Crystal[*1], Brian Marx[*3], Denise Sloan[*3], Jennifer Vasterling[*3] and Theodore Speroff[*4]
[*1] Raytheon BBN Technologies
E-mail: {ssaleem,mpacula,rkumar,rprasad,mcrystal}@bbn.com
[*2] Massachusetts Institute of Technology
E-mail: rchasin@mit.edu
[*3] National Center for PTSD at VA Boston Healthcare Sytem and Boston University School of Medicine
Email: {brian.marx,denise.sloan,jennifer.vasterling}@va.gov
[*4] VA Tennessee Valley Healthcare System and Vanderbilt University School of Medicine
Email: ted.speroff@vanderbilt.edu

***Abstract*— The stigma associated with mental health issues makes face-to-face discussions with family members, friends, or medical professionals difficult for many people. In contrast, the Internet, due to its ubiquity and global outreach, is increasingly becoming a popular medium for distressed individuals to anonymously relate experiences. In this paper, we present a system for automatically detecting psychological distress indicators in informal text interactions on Internet discussion forums. We compare a suite of innovative features and classifiers on data downloaded from an online forum discussing psychological health issues. Psychologists annotated individual messages with a comprehensive set of distress labels derived from the Diagnostic and Statistical Manual of Mental Disorders (DSM) IV. The noisy nature of the forum posts and the large set of distress labels for multi-label text classification (many of which cannot be detected by a mere surface form analysis of the text), make the task extremely challenging. A late fusion technique combines outputs from different classifiers resulting in promising accuracy on this challenging multi-label classification problem.**

I. INTRODUCTION

Detection of psychological health disorders such as Depression, Post-Traumatic Stress Disorder (PTSD), mild Traumatic Brain Injury (mTBI), etc. is conventionally based on a series of clinically administered diagnostic interviews and tests [1]. Assessment of patients using these tests is expensive, time-consuming, and sometimes unreliable due to inaccurate self-reporting by the patient. As a result, disorders such as PTSD are often under-diagnosed and under-treated [2]. An effective method for early diagnosis and treatment would be to detect salient changes in an individual's behavior in social settings. In recent times, there has been a growing shift of social interactions to the Internet via social networking sites and online discussion forums. The Internet is an ideal medium for distressed individuals to anonymously relate experiences, seek knowledge, and reach out for help. Discussions of symptoms, thoughts and experiences are open, descriptive, and honest, making them an ideal source for training psychological distress prediction models.

While there have been a few applications of automated text and voice analytics for detecting such disorders, these studies have been limited to structured questionnaires and formal clinical records [3]. Several rule-based approaches have been explored for detecting PTSD and mTBI from clinical narratives [4,5]. However, these approaches rely on annotating individual words as positive, negative, or neutral indicators of the condition. Such annotation is laborious and requires deep subject matter expertise. Given the scarcity of experts and the dynamic nature of language, such rule-based approaches are unlikely to scale to a large population.

In this paper, we present a trainable text-classification system that automatically detects psychological distress indicators from conversational text interactions on online forums. Each forum post may contain more than one distress indicator making the problem a multi-label text classification task. Two aspects of our dataset make text classification challenging. Firstly, unlike structured interviews based on pre-defined questionnaries or clinical records, we deal with noisy forum data. The posted messages are informal, unstructured and frequently exhibit poor grammar, spelling errors, and shorthand communication. Secondly, we deal with a large number of fine-grained distress labels. Many of the labels are implicitly mentioned in the text, and are inconsistently inferred even amongst human annotators. We demonstrate this in an inter-annotator agreement study where we found only moderate agreement between annotators in the coding of these distress labels.

II. DATA COLLECTION AND ANNOTATION

Our primary data corpus consists of 512 discussion threads downloaded from an online forum for veterans with post-combat psychological issues. The forum fosters anonymous discussions between returning military personnel and their caregivers with PTSD or suspected PTSD. Note that we do not identify any individuals from their posted text nor do we trace any distress signals to a specific poster.

We first annotated the dataset with distress labels. In consultation with subject matter experts (SMEs), a codebook of 136 psychological distress labels spanning PTSD, mTBI, and depression symptoms was developed. Codes were derived from the DSM-IV guidelines [6] and also from the clinical

experience of the SMEs. The labels in the codebook were organized into five broad categories: *Stress Exposure* (eg. Combat Exposure, Traumatic Loss, Captivity), *Affect* (eg. Anger/Rage/Frustration/Contempt, Fear, Worthlessness), *Behavior* (eg. Social Isolation, Sleep problems, Excessive Drug Use), *Cognition* (eg. Intrusive Thoughts and Memories, Homicide Ideation, Posttraumatic Amnesia), and *Domains of Impairment* (eg. Legal Problems, Financial Problems, Occupational Impairment). In the annotation process, each message is first tagged to indicate if a message is relevant to assessing the author's psychological state. Each relevant message is then annotated with one or more labels from the codebook characterizing the psychological state of the author in accordance with the message content.

Annotation was performed by four SMEs. We measured inter-annotator agreement among multiple annotators using the Fleiss Kappa statistic [7]. In order to compute the overall Kappa for the distress labels, we first computed the Fleiss Kappa for each label, and then performed a weighted combination of these scores. We measured a Kappa of 0.68 for the "Relevant" tag and 0.59 for the "Distress Labels" on a set of 9 threads comprising 126 messages that were annotated by all four SMEs. In general, a Kappa of 0.41-0.60 suggests moderate agreement, and 0.61 to 0.80 suggests good agreement [8]. We found that the Kappa values for the individual distress labels spanned a wide range. The distress labels with very good agreement were those that are explicitly stated in the message; for example *Sleep problems* and *Alcohol Abuse*. The labels that were in poor agreement were mostly those that required inference and world knowledge such as *Despair* and *Worthlessness.*

III. TEXT CLASSIFICATION

We approached the problem of automatically detecting psychological distress indicators in forum posts as a two stage text classification problem. We first applied a classifier to filter out messages that have no bearing on the detection of psychological distress. Some authors choose to post very short messages that do not have any information bearing content, like a simple "Thank you". Sometimes, the topic of discussion digresses to sub-topics or tangential topics. In order to identify relevant versus irrelevant messages, we trained a Support Vector Machine (SVM) classifier [9] on the annotated forum messages. We then applied multi-label classifiers to predict one or more distress labels described by the author on the relevant messages. In this paper, we focus on this second stage of text classification, and report closed-set results on messages that we know are relevant.

Algorithms for multi-label classification, the task of assigning one or more labels to an instance, can be grouped into two main categories: a) problem transformation methods, and b) algorithm adaptation methods [10]. Problem transformation methods transform the multi-label classification problem into many single-label classification problems. Algorithm adaptation methods extend specific learning algorithms in order to handle multi-label data directly. Given the large size of our label set (118 observed labels out of 136 total), we could not find a memory-efficient way to use many of the algorithm adaptation methods. Instead, we focused on problem transformation methods. In the following subsections, we describe the features and classifiers that we investigated as well as a simple, effective classifier fusion technique to combine information from multiple classifiers.

A. Features for Classification

The majority of state-of-the-art systems in text classification represent documents as a bag-of-words. While this approach works well for most tasks in the presence of enough training data, it does not capture any semantic correlations or higher order information between words. In our experiments, we explored a variety of features that look beyond the identity of the words in the message. These include message-level features computed based on the content of individual messages and thread-level features that exploit the structure of the discussion thread and look at other messages in the thread. In all cases, the features are binary, integer, or real valued and contain no Personally Identifiably Information (PII).

A1: Unigram Features – The set of words/unigrams remains the most powerful set of features and is the baseline feature set used in our experiments. In order to extract them, the data was first preprocessed to remove stop words. We also applied Porter stemming to remove the common morphological and inflectional endings in English. Emoticons or smileys were retained and used as features.

A2: Pronoun Count – Pronouns are typically discarded in most text classification applications in the pre-processing stage under the assumption that they occur too frequently to bear any information. However, in [11] it was shown that changes in the way people use pronouns when writing about traumatic experiences is a powerful predictor of changes in physician visits or an indicator of their general health. For this reason we included the normalized pronoun count as a feature.

A3: Punctuation Count – Normalized count of punctuations in the message calculated as the percentage of tokens/words in the message that are punctuations.

A4: Average Sentence Length - Average number of words in the message sentences, where sentence segmentation was determined based on punctuations and line breaks.

A5: Sentiment Features - Sentiment bearing words are correlated well with specific distress labels (especially in the Affect category of labels which includes emotions). Identifying and grouping such words in a message could positively influence the classification performance of these labels. We extracted 125 binary features indicating the presence or absence of sentiment bearing words in the message. These words were selected from two sources: 1) 68 lexicons form the Linguistic Inquiry and Word Count (LIWC) [12] and 2) 57 lexicons from the General Inquirer (GI) system [13]. The LIWC includes categories corresponding to affective and emotional processes (e.g.: positive/negative emotions), Cognitive Processes (e.g.: causation) and Social Processes (e.g.: friends) among others. The GI System includes valence categories (positive, negative) and motivation related words.

A6: Lead Author Post - Binary feature indicating whether the message was posted by the author who started the thread.
A7: First Responder Post - Binary feature indicating whether the message was posted by the author who first responded to the lead message of the thread.
A8: Thread Similarity - Real-valued feature that measures the average cosine similarity of the message to other messages in the thread.
A9: First Message Similarity - Real-valued feature that measures the cosine similarity of the message to the first message posted in the thread.
A10: Dependency Pairs as Features: Inspired by work in [14], we investigated the use of syntactic dependency relations as features for text classification. It is generally accepted in the NLP community that syntactically related pairs of words imply a semantic concept. We first parsed the text in the messages using the off-the-shelf Stanford dependency parser [15]. We then selected a subset of the dependency relations as features namely: Adjectival complement, Agent, Adjectival Modifier, Negation, Possession, Purpose Clause Modifier, Relative, Temporal Modifier and Adverbial Clause Modifier. Examples of dependency word pairs in the data are "alcohol-bad", "alcoholics-drunk", "benefits-ptsd", "flashback-sleep", "angry-blood" and "Avanza-soltab".

B. Classifiers and Combination

We trained both Support Vector Machine (SVM) and Conditional Random Fields (CRF) classifiers and combined their results. SVMs are discriminative classifiers that try to find the hyperplane that maximizes the margin between two classes which are represented as vectors in an n-dimensional space [9]. They are popularly used in text classification due their ability to inherently deal with a high dimensional feature space. We trained a binary one-versus-all SVM for each label in our dataset. CRFs are graphical models that predict the conditional probability $P(y|x)$ based on a graph that models relations between labels y and features x, with parameters for each clique in the graph [16]. We modeled each label with an unchained binary CRF thus treating each label decision independently. We performed inference using marginal label probabilities at each time step, selecting the union of top k labels and labels with probability above threshold th. We trained the CRF on individual messages in a thread. For testing, we split each message into sentences and performed inference by taking the maximum score for a label across all sentences in the message. This allowed us to exploit the fact that each label is usually generated by only a subset of the message, with the rest of the message irrelevant to that label.

Next, we implemented a top-level fusion component that combines information from multiple classifiers. Each label is assigned a score F_l

$$F_l = \sum_{i=1}^{N} (C_{il} + w_i \times S_{il}) \quad (1)$$

where C_{il} is 1 if classifier i predicts label l or 0 otherwise. S_{il} is the score assigned by classifier i for label l; w_i is the weight assigned to classifier i; and, N is the number of classifiers being fused. The term C_{il} in Equation 1 guarantees that labels predicted by multiple classifiers are given a higher score over ones that were predicted by only one classifier, even if the classifier was very confident of its prediction. To break ties between labels, we add a second term which is the weighted combination of the individual classifier scores for each label.

IV. Results and Discussion

We chose a set of 512 threads, comprising of 5000 relevant and irrelevant messages, for our experiments. We held out 90 threads for testing, and used the remaining for training. All system parameters were tuned based on 10-fold cross validation on the training set where threads were randomly distributed across 10 different subsets. Performance is reported on the held-out test set. Table I shows the data statistics of the experimental corpus.

TABLE I
EXPERIMENTAL DATA STATISTICS

Category		**Train**	**Test**
Threads		422	90
Relevant	Messages	1868	440
	Total Words	397K	92K
	Unique Labels	118	97
	Average Number of Labels per message	2.8	2.9

Classification performance is measured using the micro-averaged F-Measure (F) [17] measured as the harmonic mean of precision (P) and recall (R) computed over the pool of all distress labels across messages. Furthermore, the labels predicted for all messages posted by the same author within a thread were pooled for evaluation. We had the SMEs generate a set of 25 label clusters by hierarchically clustering the individual distress labels. As an example of clustering, the labels *Intrusive Thoughts and Memories*, *Nightmares//Unpleasant Dreams* and *Postraumatic Amnesia* were grouped into the *Memory Problems* cluster. We report results on both the fine-grained labels and the 25 clusters.

For our experiments with SVMs, we used the Weka machine learning software [18] with the Radial Basis Function (RBF) Kernel. We performed grid-search to find the best regularization (C) and gamma (g) parameters on the cross-validation set. For the baseline experiment with SVMs, each message was treated as a bag of words with normalized (TF-IDF) frequencies. For every label, we further optimized the threshold for classification for the best micro-averaged F-measure on the cross-validation set. Next, the remaining features described in section III-A were incrementally added to the baseline feature set of the SVM classifier. Table II shows the performance of the SVM with the unigram TF-IDF features as well as the improvements from adding the other features. A small but consistent improvement in performance is seen with the incremental addition of the message level features A2-A5 and thread level features A6-A9. The largest gain however stems from the addition of dependency features. Overall, the micro-averaged F-Measure for all 118 labels improved by 2.2% relative using all features with SVMs. In our experiments with CRFs, we used binary unigram presence as the feature vector. Training and inference were performed

using the Mallet software package [19]. We selected all labels with probabilities above 0.15, or at least 2 top labels per message based on the best micro-averaged F-measure on the cross-validation set. Results with the CRFs are also shown in Table II. We found that the CRFs perform a little worse than the SVMs with the same feature set (binary unigrams). However, when we compared the label-wise F-measure between the two classifiers, we found that while the overall F-Measure was better with the SVMs, there were certain labels where the CRFs outperformed the SVMs. Specifically, the SVMs beat the CRFs on 19 labels, the CRFs beat the SVMs on 13 labels and the two classifiers performed equally on the remaining labels.

TABLE II
MULTI-LABEL CLASSIFICATION RESULTS WITH DIFFERENT CLASSIFIERS AND FEATURE SETS

Classifier (Features)	**P**	**R**	**Micro-F**	**Micro-F (25 Clus)**
SVM (A1)	46.9	42.4	44.5	59.0
SVM (A1, A2, …, A5)	47.1	42.3	44.6	59.1
SVM (A1, A2, …, A9)	47.4	42.7	44.9	59.1
SVM (A1, A2, …, A10)	49.3	42.2	45.5	60.4
CRF (A1)	46.0	42.0	43.8	57.1
SVM (A1, …, A10) + CRF (A1)	53.4	42.1	47.1	60.7

The weights w_i for system combination were chosen by exhaustively searching the interval {0 – 10} with a step size of 0.1 to optimize the micro-averaged F-measure on the cross-validation set. The performance of system combination with the best SVM and CRF configuration is shown in Table II. There was predictably no improvement in recall with system combination given that no new labels were predicted. However a relative improvement of 8.3% in precision and 3.5% in micro-averaged F-measure is seen over the best individual classifier. The macro-averaged F-measure [17] for the combined system was 9.9 for all 118 labels. Note that this low value stems from the class imbalance in our dataset. The most frequently occurring label – Anger/Rage/ Frustration/ Contempt has 698 training examples whereas half of the labels have less than 20 examples in training. A large number of labels hence perform poorly. Our results are comparable to multi-label classification results reported in [10] on datasets of similar cardinality. We also generated a Receiver Operating Characteristic (ROC) by plotting the true positive rate against the false positive rate for the system combination output and found the Area Under the Curve (AUC) to be 0.73.

V. Conclusions and Future Work

In this paper, we introduced a system that automatically detects psychological distress indicators from text in online forum posts. We explored different text classification algorithms and features and combined their top-level results thereby leveraging their individual strengths. In the future, we intend to investigate methods that exploit label dependencies and also leverage domain rules for the rarer labels. We also plan to investigate contextual features that exploit information from previous messages within the thread. Finally, we hope to extend the system to use the output labels to predict diagnostic criteria for PTSD, mTBI, or depression.

Acknowledgment

This paper is based upon work supported by the DARPA DCAPS Program. The views expressed here are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

[1] F. W. Weathers, T. M. Keane, and J. R. T. Davidson. 2001. *Clinician-Administered PTSD Scale: A review of the first ten years of research*. Depression and Anxiety, Vol 13(3), 132-156.

[2] 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.

[3] S. H. Brown, et al. 2006. *eQuality: Electronic Quality Assessment from Narrative Clinical Reports*. Mayo Clinic Proceedings, vol. 81, pp. 1472-1481.

[4] P. L. Elkin, et al. 2010. *The Health Archetype Language (HAL-42): Interface considerations*. International Journal of Medical Informatics, vol. 79, pp. 71-75.

[5] B. Trusko, et al. 2010. *Are Post Traumatic Stress Disorder Mental Health Terms Found in SNOMED-CT Medical Terminology?* Journal of Traumatic Stress, vol. 23, pp. 794-801.

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

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

[8] J. R. Landis and G.G. Koch. 1977. *The measurement of observer agreement for categorical data*. . Biometrics 33 (1): 159–174.

[9] V. N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer, New York.

[10] G. Tsoumakas, I. Katakis, I. Vlahavas. 2011. *Random k-Labelsets for Multi-Label Classification*. IEEE Transactions on Knowledge and Data Engineering, IEEE, 23(7), pp. 1079-1089

[11] R. S. Campbell and J. W. Pennebaker. 2003. *The secret life of pronouns: Flexibility in writing style and physical health*. Psychological Science, 14, 60-65.

[12] J. W. Pennebaker, R. J. Booth, M. E. Francis. 2007. *Linguistic Inquiry and Word Count (LIWC2007): A text analysis program*. Austin, TX: LIWC (www.liwc.net).

[13] P. J. Stone. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press

[14] V. Nastase, J. Sayyad and M. Fernanda. 2007. *Using Dependency Relations for Text Classification*. Caropreso University of Ottawa SITE Technical Report TR-2007-12

[15] M.C. de Marneffe, B. MacCartney and C. D. Manning. 2006. *Generating Typed Dependency Parses from Phrase Structure Parses.* Proceedings of *LREC*

[16] J. Lafferty, A. Mccallum, and F. Pereira. 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. In Proceedings of 18th International Conference on Machine Learning, pp 282-289.

[17] Y. Yang. 1999. *An evaluation of statistical approaches to text categorization.* Journal of Information Retrieval, 78-88

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten. 2009. *The WEKA Data Mining Software: An Update*. SIGKDD Explorations, Volume 11, Issue 1

[19] A. McCallum. 2002. *MALLET: A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu