Automatic Detection of Psychological Distress Indicators and Severity Assessment from Online Forum Posts

Shirin Saleem[1], Rohit Prasad[1], Shiv Vitaladevuni[1], Maciej Pacula[1], Michael Crystal[1], Brian Marx[2,3], Denise Sloan[2,3], Jennifer Vasterling[2,3] and Theodore Speroff[4,5]

(1) Raytheon BBN Technologies, 10 Moulton St, Cambridge, MA, U.S.A.
(2) National Center for PTSD at VA Boston Healthcare Sytem, Boston, MA, U.S.A.
(3) Boston University School of Medicine, Boston, MA, U.S.A.
(4) VA Tennessee Valley Healthcare System, Nashville, TN, U.S.A.
(5) Vanderbilt University School of Medicine, Nashville, TN, U.S.A.

{ssaleem,rprasad,svitalad,mpacula,mcrystal}@bbn.com
{brian.marx,denise.sloan,jennifer.vasterling}@va.gov
ted.speroff@vanderbilt.edu

ABSTRACT

Psychological disorders are frequently under-diagnosed and consequently have an irreversible impact on individuals and society. The stigma associated with such disorders makes face-to-face discussions with family members and clinicians difficult for many individuals. In contrast, people openly relate experiences on Internet forums. This paper describes a novel system that analyses forum posts to: (1) detect distress indicators that directly map to the Diagnostic and Statistical Manual of Mental Disorders (DSM) IV constructs, and (2) assess the severity of distress for prioritizing individuals who should seek clinical help (i.e. triage). For distress indicator detection, we use support vector machines (SVMs) trained on a suite of innovative intra- and inter-message features. We show significant improvements in multi-label classification accuracy using human-generated rationales in support of annotated distress labels. For triage assessment, we demonstrate the effectiveness of Markov Logic Networks (MLNs) in dealing with noisy distress label detections and encoding expert rules.

KEYWORDS: Psychological Distress, Web forums, Text classification, Annotator rationales, Support Vector Machines, Probabilistic Logic, Markov Logic Networks.

Proceedings of COLING 2012: Technical Papers, pages 2375–2388,
COLING 2012, Mumbai, December 2012.

1 Introduction

Psychological health disorders pose a growing threat to individuals, their family members and to society. Disorders such as Depression, Post-Traumatic Stress Disorder (PTSD), and mild Traumatic Brain Injury (mTBI), are often under-diagnosed and under-treated (Kessler et. al, 1999). Failure to intervene early and effectively impacts individuals and their family members adversely and results in profound long-term costs to society.

The standard approach to diagnosing psychological health disorders is through a series of clinically administered diagnostic interviews and tests (Weathers et. al, 2001). However, assessment of patients using these tests is expensive and time-consuming. Furthermore, the stigma associated with mental illnesses motivates inaccurate self-reporting by affected individuals and their family members, thus making the tests unreliable.

In recent years, there has been a tremendous growth in social interactions on the Internet via social networking sites and online discussion forums. In contrast to clinical tests, the Internet is an ideal, anonymous medium for distressed individuals to relate their experiences, seek knowledge, and reach out for help. Web-forum discussions of symptoms, thoughts and experiences are open, descriptive, and honest, making them an ideal source for observing communications of individuals for assessing psychological status.

In this paper, we present a multi-stage text classification system for assessing psychological status of individuals based on their text postings on online web forums. 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 (American Psychiatric Association, 2000), and (2) assesses the severity of distress that can be used to triage individuals who should seek clinical help.

The same factors that make web-forum data interesting for observing psychological distress also make automated analysis extremely challenging. For instance, the language used in such forums is highly informal, with ill-formed, grammatically incorrect sentences, misspellings, and special character sequences such as emoticons. Vague references to emotional states, description of present vs. past traumatic experiences, and relating one's own versus other's experience all pose novel challenges to natural language processing (NLP). Additionally, any approach for psychological health analysis of text interactions must incorporate domain knowledge from expert psychologists and clinicians. Together these challenges make this domain a fascinating research area with the potential for research advances to revolutionize psychological healthcare.

1.1 Previous Work

Existing applications for automatic detection of psychological disorders have been limited to structured questionnaires and formal clinical records (Brown, et. al. 2006). In contrast, our work is focused on noisy, informal text messages from Web-forums. Text classification research on such data has primarily focused on identifying social roles in scientific forums (Wang, et. al, 2011) and sentiment analysis (Abbasi et. al, 2008). To the best of our knowledge, the work presented in this paper for assessing psychological status from web-forum text is first of its kind.

Several rule-based approaches have been explored for detecting PTSD and mTBI from clinical narratives (Elkin et. al, 2010) (Trusko et. al, 2010). However, these approaches rely on annotating individual words as positive, negative, or neutral indicators of the condition. Such annotation is

laborious, lacks consistency, and requires deep subject matter expertise. Instead, our approach uses statistical models that do not require such laborious annotation and encode domain knowledge by learning weights for the domain rules from data.

1.2 Novel Contributions

We present several novel techniques within a multi-stage text classification framework for assessing psychological status from informal text posted on Web-forums. First, we describe a suite of features and classifiers trained on expert-annotated text to detect distress indicators. The training data itself is a first of its kind, where each message has been annotated by psychologists using a codebook of 136 distress labels that directly map to DSM-IV constructs. Since messages are often tagged with multiple distress indicators, the detection task is a multi-label classification problem with a large set of labels. Additionally, a fraction of our data is annotated with rationales that support distress labels. We show that these rationales can be effectively used to improve multi-label classification accuracy. Specifically, we observe a relative improvement of 14.6% over using plain text features. Another key contribution of this work is the use of probabilistic logic, namely Markov Logic Networks (MLNs) (Richardson and Domingos, 2006) to incorporate domain-specific rules, and handle the inherent noise in the data. We show that MLNs improve the triage classification accuracy, and provide a robust approach for inferring triage codes from noisy distress label detections as well as potentially contradictory domain rules.

2 Corpus for Experimentation

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

In consultation with psychologists, a codebook of 136 psychological distress labels spanning PTSD, mTBI, and depression symptoms was developed. Codes/labels were mostly derived from the DSM-IV guidelines (American Psychiatric Association, 2000). The labels were organized into five broad categories: Stress Exposure (e.g., Combat Exposure, Traumatic Loss, Captivity), Affect (e.g., Anger/Rage/Frustration/Contempt, Fear, Worthlessness), Behaviour (e.g., Social Isolation, Sleep problems, Excessive Drug Use), Cognition (e.g., Intrusive Thoughts and Memories, Homicide Ideation, Posttraumatic Amnesia), and Domains of Impairment (e.g., 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. Additionally, for a subset of messages, we highlighted contextual rationales to support the distress labels annotations. Figure 1 shows a snapshot of the distress labels and their hierarchy.

Expert psychologists next annotated each author in a thread with a triage code that indicates treatment acuity or the priority assigned to a referral for additional treatment. We used three triage codes in our annotation – 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. For each of these triage codes,

the treatment acuity varies from emergency intervention or urgent care evaluation for TR1 to non-urgent treatment referral for TR2 to no recommendation for treatment for TR3. Since online forums are moderated and expunged of sensitive content, we rarely observed any occurrences of TR1 in the forum posts. Our focus in this paper is hence restricted to distinguishing between codes TR2 and TR3. However, our approach is extensible to the detection of TR1 if appropriate training data were available.

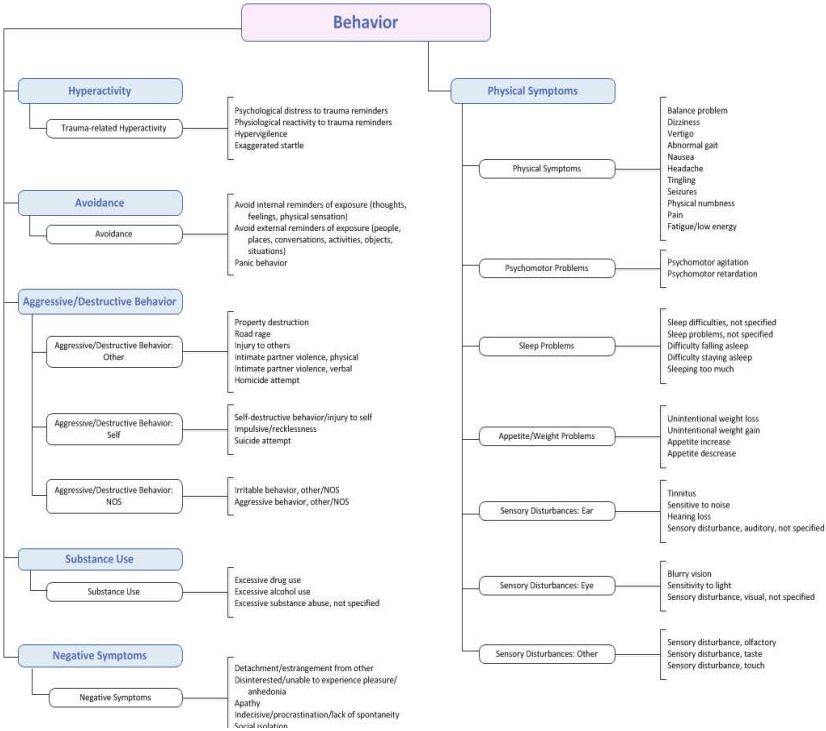

Figure 1: Snapshot of codebook of distress labels and their hierarchy.

3 Approach Overview

Figure 2 gives an overview schematic of our approach. We use a trainable multi-stage text-classification system to detect distress indicators from text interactions on Web forums and severity of distress of an author for prioritizing need for clinical care. Our system analyses the text posted by an author to first determine if it is relevant for psychological distress. If relevant, the text is further processed using multi-label classification to estimate fine-grained

psychological distress indicators. Next, information from the text and the detected distress labels is combined using domain-specific rules to estimate priority for intervention. In what follows, we describe the details for fine-grained distress detection and severity assessment.

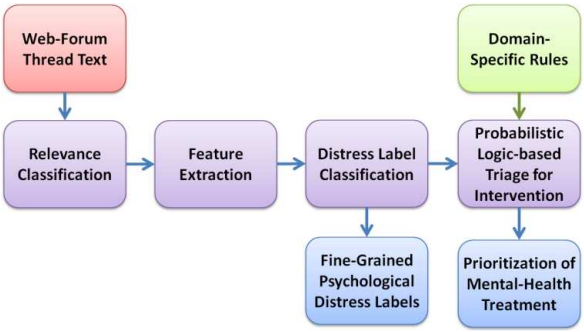

Figure 2: Schematic of Approach to Estimate Psychological Distress Labels and Prioritization of Mental-Health Intervention from Web-Forum Text.

4 Multi-label Distress Classification

4.1 Classifier

Algorithms for multi-label classification, the task of assigning one or more labels to an instance, can be grouped into two main categories: (1) problem transformation methods, and (2) algorithm adaptation methods (Tsoumakas et al. 2011). 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 using binary one-versus-all Support Vector Machines (SVMs) that detect the presence or absence of each of the fine-grained distress labels.

4.2 Features

Most systems for text classification represent documents as a bag-of-words. While this approach works well for most tasks with adequate 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 as well as 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: Unigrams – We extracted unigrams from the forum messages by first removing stop words. Next, we apply Porter stemming to remove the common morphological and inflectional endings in English. Emoticons such as 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 (Campbell and Pennebaker, 2003) 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. We hence 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 Words - Sentiment bearing words are correlated with specific distress labels (especially in the Affect category of distress labels). 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: 68 lexicons form the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et. al, 2007), and 57 lexicons from the General Inquirer (GI) system (Stone, 1966). 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 words in the message to the other messages in the thread.

A9: First Message Similarity - Real-valued feature that measures the cosine similarity of the words in the message to the words in the first message posted in the thread.

A10: Domain Phrases Derived from Rationales – (Zaidan et. al, 2008) showed improved performance in a sentiment classification task using annotator rationales within a contrastive learning framework of an SVM. Here, we use the rationales by extracting label-specific textual features from them. For every label, we first find the most frequent n-grams (n <= 5) in the highlighted rationales. We then filtered n-grams that had a high overlap ratio with other labels and also those that consisted solely of words in a pre-defined stop word list. The resulting n-grams were then used as binary features for classification. Examples of such phrases for the label Suicidal Ideation include: “thought about jumping”, “me suicidal”, “end their life”, “feel like killing myself”.

5 Psychological Triage Models for Severity Assessment

Our goal is to find authors who might require treatment or medical evaluation based on any behavioural disturbances, distress, functional impairments and/or suicidal or homicidal ideation. We explored two approaches to address this problem. The first approach uses an SVM trained on the words and predicted distress labels for the messages posted by the author. Our second approach uses Markov Logic Networks (MLNs) (Richardson and Domingos, 2006) to encode domain knowledge using probabilistic first order rules with associated weights.

In our system, the MLN computes the probability of a triage code using: (1) the distribution of words in the messages posted by an author, (2) the predicted distress labels, and (3) domain-specific rules that encode dependencies between the text, distress labels and the triage. The domain-specific rules were derived from existing diagnostic criteria as follows:

1. Rules derived from Primary Care-PTSD (PC-PTSD) screening test (Prins et al. 2003) used routinely in the VA to screen for PTSD. It comprises of 4 questions which map to 10 distress labels from the codebook.
2. Rules derived from DSM-IV guidelines for PTSD. These comprise of 4 criteria consisting of questions that map to distress labels in the codebook. For example, a criterion encoded as a rule in the MLN is the presence of one or more of the trauma exposure labels and one or more of the fear/helpless labels.

hasSymptom(Helplessness, p) OR hasSymptom(Fear, p) OR hasSymptom(Horror,p) => triageCode(+t, p)
hasSymptom(Intimate family impairment, p) OR hasSymptom(Extended family impairment, p) OR hasSymptom(Friendship impairment, p) OR hasSymptom(Social impairment, p) OR hasSymptom(Occupational impairment, p) OR hasSymptom(Educational impairment, p) OR hasSymptom(Self-care impairment, p) OR hasSymptom(Financial problems, p) OR hasSymptom(Legal problems, p) => triageCode(+t, p)
hasSymptom(Sleep problems, p) OR hasSymptom(Difficulty falling asleep, p) OR hasSymptom(Anger, p) OR hasSymptom(Road rage, p) OR hasSymptom(Property destruction, p) OR hasSymptom(Concentration problem, p) OR hasSymptom(Hypervigilence, p) OR hasSymptom(Exaggerated startle, p) => triageCode(+t, p)
hasSymptom(Intrusive thoughts, p) OR hasSymptom(Nightmares, p) OR hasSymptom(Reliving event, p) OR hasSymptom(Psychological distress to trauma reminders, p) OR hasSymptom(Physiological reactivity to trauma reminders, p) => triageCode(+t, p)
hasSymptom(Nightmares, p) OR hasSymptom(Reliving event, p) OR hasSymptom(Intrusive thoughts and memories of events, p) => criterion1(True, p)

Table 1: Examples of domain-specific rules derived from DSM-IV guidelines and PC-PTSD screening tests. Here, p is variable ranging over authors of messages; and t ranges over triage codes.

MLNs have two key advantages for our application. First, the use of statistical inference provides robustness to noise in the text and label predictions, and potential contradictions in the domain-specific rules. Second, the relative weights for the domain-specific rules can be automatically learned from the training data.

We employed Alchemy, an implementation of learning and inference algorithms for MLNs, (Richardson and Domingos, 2006) for our experiments. To learn the weights of the domain-specific rules, we used discriminative training, which maximizes the conditional likelihood of target labels (in our case the triage codes) given the observed variables (in our case the message words and distress labels). Alchemy uses an approach referred to as pre-conditioner scaled conjugate gradient for discriminative weight learning (Lowd and Domingos, 2007). The inference is performed using MaxWalkSAT; see (Richardson and Domingos, 2006) for details. Table 1 shows examples of domain-specific rules incorporated in the MLN based on DSM-IV guidelines and PC-PTSD screening test.

6 Experimental Results

6.1 Inter-annotator Agreement

We performed an inter-annotator agreement study for both distress label classification and triage annotation. Annotation for distress labels was performed by four Subject Matter Experts (SMEs). We measured inter-annotator agreement among multiple annotators using the Fleiss Kappa statistic (Fleiss, 1971). 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 observed 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 (Landis and Koch, 1977). We found that the inter-annotator agreement, i.e. the Kappa values, for the individual distress labels spanned a wide range. Some of the distress labels had very good agreement, e.g., Sleep problems, and Alcohol abuse, possibly because the messages contained extensive descriptions of the distress conditions. The labels that were in poor agreement were typically those that required inference and world knowledge, e.g., Despair and Worthlessness. We will further investigate this inter-annotator agreement disparity as part of future work. Annotation for the triage classification was performed by six SMEs. We again measure the Fleiss Kappa statistic for triage codes assigned to 43 authors across 10 threads. We found this value to be 0.71, indicating good agreement.

6.2 Multi-label Distress Classification

We chose a set of 512 threads, comprising of 5000 relevant and irrelevant messages, for our multi-label distress classification experiments. We held out 90 threads for testing, and used the remaining for the training set. We collected rationales for 650 messages in training. The SVM 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 2 shows the data statistics of the experimental corpus.

Category		Train	Test
Threads		422	90
Authors		1166	260
Relevant	Messages	1868	440
	Total Words	397K	92K
	Unique Labels	118	97
	Average Number of Labels per message	2.8	2.9

Table 2: Corpus setup for Multi-Label Distress Classification

As described in Figure 2, we approached the problem of automatically detecting psychological distress indicators in forum posts in two stages. We first applied a classifier to filter out messages that have no bearing on the detection of psychological distress. Irrelevant messages include cases such as when authors choose to post very short messages that do not have any information bearing content, like a simple "Thank you", and when the topic of discussion digresses to sub-topics or tangential topics. In order to identify relevant versus irrelevant messages, we trained an SVM on the annotated forum messages, and used it to automatically recognize relevant messages in the test set. 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.

Classification performance is measured by computing the mean of the Area Under the Curve (AUC) for all labels. The AUC for each label is computed on a Receiver Operating Characteristic (ROC) curve with the false acceptance rate (FAR) bounded at 10%, and normalized such that the maximum possible AUC is 1.We also report the overall AUC number for the entire ROC curve, i.e. FAR of 100%. The labels detected for all messages posted by the same author within a thread were pooled for evaluation. For our experiments with SVMs, we used the Weka machine learning software (Hall et. al, 2009) 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. Next, the remaining features described in section 4.1 were incrementally added to the baseline feature set of the SVM classifier. Table 3 shows the performance of the SVM with the unigram TF-IDF features as well as the improvements from adding the other features. For a random classifier, the mean-AUC bounded up to False Acceptance Rate of 10% is 0.05, and the overall AUC is 0.5. No significant change in performance is seen with the incremental addition of the message level features A2-A5 and thread level features A6-A9. We retained these features since their addition did not explicitly hurt performance. Overall, the mean-AUC improves by 14.6% relative using the full set of features in section 4.1 over just the unigram words (Table 3). We see a large gain from the addition of the domain phrase features derived from rationales (A10).

We found that our approach of using the rationales by extracting label specific domain phrase features out-performed the contrastive approach in (Zaiden et. al, 2008). The latter gave a

bounded mean-AUC of 22.3, whereas our feature-based approach yielded 23.5 when added to the unigram feature set.

Feature Set	Mean AUC Bounded for 0-10% False Accept Rate	AUC (Overall)
A1	0.213	0.6757
A1, A2, A3, A4, A5	0.211	0.6699
A1, A2, A3, A4, A5, A6, A7, A8, A9	0.212	0.6699
A1, A2, A3, A4, A5, A6, A7, A8, A9, A10	0.244	0.6874

Table 3: Multi-label distress classification results with different feature sets

It is to be noted that the dataset has a high class imbalance. 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. Hence, a large number of labels perform poorly merely due to the lack of sufficient training data. In Figure 3 we also show the AUCs for all the labels. Approximately half the labels have an AUC < 0.2. The maximum value of individual AUC was found to be 0.884 for Excessive Substance Use. The top 5 labels with maximum AUC are Excessive Substance Use, Panic behavior, Nightmares or Unpleasant Dreams, Concentration Problems and Child Maltreatment. In all of these labels, there is extensive description of the distress condition in the messages. In contrast, there are many labels that are implied in the text, and are inconsistently inferred even amongst human annotators. We demonstrated this in the inter-annotator agreement study where we found only moderate agreement between annotators in the coding of these distress labels.

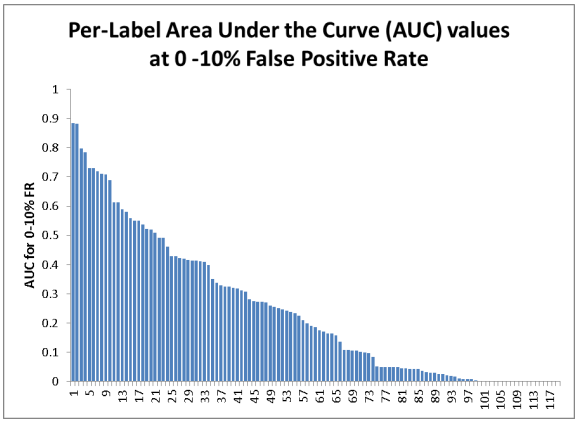

Figure 3: Per-label AUC values for false positive rate capped at 10%. The AUC is normalized such that the maximum possible value is 1.0.

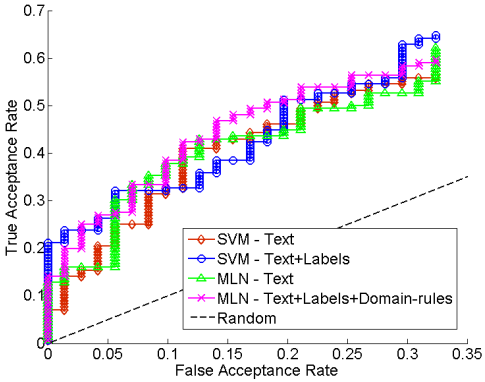

Figure 4: ROC curves for triage classification for SVM and MLN for Text and Text + Distress Labels. MLN with Text and Distress Labels combined using domain specific rules gives best results.

6.3 Triage Assessment

A subset of authors in the downloaded forum threads were tagged with triage codes, specifically 907 authors out of 1426. We used 680 of these for training the triage models, and 227 for evaluation. We compare the triage classification performance for SVM and MLN using ROC curves, i.e. rate of correct acceptance of TR2 versus false acceptance of TR3. The performance was measured using area under the curve (AUC). We capped the ROC curves to false acceptance rates less than 33% based on the fact that high false acceptance rates make the triage impractical for our application. The AUC is normalized such that the maximum possible AUC is 1. The AUC of a random/chance classifier is 0.165. Table 4 presents the AUC values for SVM and MLN for different types of inputs. As can be observed, MLNs provide statistically significant gains over SVMs by using domain-specific rules for combining information from text as well as the distress label detections. Figure 4 shows the ROC curves for the triage classification.

Method	**Area Under the Curve (AUC) with Bounded False Accept Rate of 33%**
SVM - Text	0.4090
SVM – Text + Distress Labels	0.4354
MLN – Text	0.4148
MLN – Text + Distress Labels + DSM-IV and PC-PTSD Rules	**0.4515**

Table 4: Triage classification performance AUC for ROC curves capped at false acceptance rate less than 33%. The AUC is normalized such that maximum possible value is 1.0

7. Conclusions and Future Work

In this paper, we introduced a powerful system that automatically detects psychological distress indicators from text in online forum posts, and demonstrated it in a novel domain of unconstrained web-forums. We presented multi-label classification for 136 labels of fine-grained psychological distress conditions on extremely challenging unstructured text data, and a novel approach based on probabilistic logic to employ domain-specific rules for combining information from text features and the distress label detections. We also showed that incorporating rationales from domain experts for the label annotations helps improve the multi-labeling performance, and presented a novel feature to exploit the rationale annotations.

In the future, we intend to investigate methods that exploit label dependencies. We will also investigate contextual features for classification that exploit information from previous messages within a thread. Finally, we plan to validate the system on text data from subjects diagnosed with PTSD and compare the outcomes on a control group that does not suffer from PTSD.

Acknowledgments

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

A. Abbasi, H. Chen, and A. Salem. 2008. Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. ACM Transactions on Information Systems, 26(3), no. 12.

American Psychiatric Association. 2000. Diagnostic and statistical manual of mental disorders (4th ed., text rev.). Washington, D

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

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.

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

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

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

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.

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

D. Lowd, and P. Domingos, 2007, E*ffi*cient weight learning for Markov logic networks. In PKDD-2007, 200–211.

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).

A. Prins, P. Ouimette, R. Kimerling, R. P. Cameron, D. S. Hugelshofer, J. Shaw-Hegwer, A. Thrailkill, F. D. Gusman, J. I. Sheikh. 2003. The primary care PTSD screen (PC-PTSD): development and operating characteristics. Primary Care Psychiatry, 9, 9-14

M. Richardson and P. Domingos. 2006. Markov logic networks. Machine Learning 62, 1-2, pp. 107-136.

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

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.

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

V. N. Vapnik. 1995. The Nature of Statistical Learning Theory. Springer, New York.

L. Wang, M. Lui, S. N. Kim, J. Nivre and T. Baldwin. 2011. Predicting Thread Discourse Structure over Technical Web Forums. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011), Edinburgh

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.

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