Detecting Distressed and Non-distressed Affect States in Short Forum Texts

Michael Thaul Lehrman **Cecilia Ovesdotter Alm** **Rubén A. Proaño**
Rochester Institute of Technology
michael.lehrman@alum.rit.edu coagla@rit.edu rpmeie@rit.edu

Abstract

Improving mental wellness with preventive measures can help people at risk of experiencing mental health conditions such as depression or post-traumatic stress disorder. We describe an encouraging study on how automatic analysis of short written texts based on relevant linguistic text features can be used to identify whether the authors of such texts are experiencing distress. Such a computational model can be useful in developing an early warning system able to analyze writing samples for signs of mental distress. This could serve as a red flag, signaling when someone might need a professional assessment by a clinician.

This paper reports on classification of *distressed* and *non-distressed* short, written excerpts from relevant web forums, using features automatically extracted from input text. Varying the value of *k* in *k*-fold cross-validation shows that both coarse-grained and fine-grained automatic classification of affect states are generally 20% more accurate in detecting affect state than randomly assigning a distress label to a text. The study also compares the importance of bundled linguistic super-factors with a 2^k factorial model. Analyzing the importance of different linguistic features for this task indicates main effects of affect word list matches, pronouns, and parts of speech in the predictive model. Excerpt length contributed to interaction effects.

1 Introduction

Many people today deal with depression, post-traumatic stress disorder, and other mental disorders involving anxiety or distress, both diagnosed and undiagnosed. The societal costs of treating mental health are staggering. Sultz and Young (2011) estimate that the total mental health care treatment costs in the United States amount to more than USD 100 billion per year. The health care system in the United States generally focuses on treating patients' illnesses rather than on preventing their occurrence, and mental health care is no exception. Mental health diagnosis typically takes place after patients already show behavioral and physical symptoms associated with mental distress. Moreover, there are 33,000 suicides every year in the United States and, according to Matykiewicz et al. (2009; referencing Kung et al. (2008)), "[i]n the United States, suicide ranks second as the leading cause of death among 25-34 year-olds and the third leading cause of death among 15-25 year-olds" (p. 179).

Diagnosing mental illnesses is difficult. For example, depression has a prevalence of 19.5%, according to Mitchell et al. (2009), and is mostly diagnosed and treated by general practitioners. However, it is diagnosed correctly in only 47.3% of cases.

Commonly, the initial assessment of mental distress does not rely on clinical tests or advanced technology, and the evaluation of a patient is typically performed through the use of standardized questionnaires. A patient's answers are then compiled and compared with disease classification guidelines, such as the International Classification of Diseases or the Diagnostic and Statistical Manual, to guide the patient's diagnosis. However, these diagnostic methods are not precise and have high rates of false positives and false negatives. For example, in the United States, half of those who received mental health treatment did not meet the diagnostic criteria for a mental disorder (Kessler et al., 2005). In addition, societal and financial barriers prevent many people from seeking medical attention. In fact, in the USA, between 1990 and 2003, two-thirds of those with mental disorders did not receive treatment (Kessler et al., 2005). Many societies around the world stigmatize and discriminate against people with mental disorders, contributing to the unwillingness of individuals to acknowledge the problem and seek help (Michels et al., 2006; Fabrega, 1991).

Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012), pages 9–18,
Montréal, Canada, June 7, 2012.

It would be helpful if, e.g., military clinicians could effectively and non-invasively analyze soldiers' writing samples, social media posts, or email correspondence to screen service members for trouble coping with combat-related stress, to complement self-reporting or patient surveys. Careful thought would be required for access to such information so that it helps and not hurts. It seems useful as additional information for doctors.

We report on an initial study in which we analyze a smaller balanced dataset and experiment with inference of affect states at two different levels of affective granularity. Our work is based on Natural Language Processing (NLP) using supervised machine learning. We also discuss 2^k factorial, a method commonly used in engineering statistics, which has been successfully applied to many domains within engineering and product design for feature selection. Our work contributes initial reference values for what can be achieved by applying four fundamental supervised classification methods and text-based features to the challenging task of automatically classifying mental affect states in short texts based on just a small dataset. We discuss performance both in terms of different experimental setups, which linguistic features matter, and how labels confuse with each other.

2 Relevant previous work

Computational linguistics approaches have been applied to a range of challenging problems with impact outside the language technology field, e.g., to predict pricing movements on the stock market (Schumaker, 2010) or opinions on political candidates in event prediction markets (Lerman et al., 2008). In psychology, psychiatry, and criminology, studies with natural language data have found differences in behaviors for mental health patients or inmates with various mental health disorders (e.g., Andreasen and Pfohl, 1976; Harvey, 1983; Ragin and Oltmanns, 1983; Fraser et al., 1986; Endres, 2004; Gawda, 2010).

Recently, computational linguists have increasingly tackled problems in health care. For example, Zhang and Patrick (2006) automatically classified meaningful content in clinical research articles. Jha and Elhadad (2010) predicted how far breast cancer patients had progressed in their disease, based on discourse available in postings on web forums. As another example, Roark et al. (2007) explored the use of structural aspects of the language of individuals with mild cognitive impairment in assisting with such diagnostics.

More specifically in mental health, Yu et al. (2009) classified five forms of "negative life events" in text (p. 202). Pestian et al. (2008) were able to use machine learning, taking advantage of text characteristics to classify suicide notes as written by either "simulators" or "completers" as accurately as mental health experts (p. 96). The authors also found that emotional content was useful for the expert clinicians, but not for the automatic inference methods. However, this might indicate that the study did not consider an appropriate feature set. In comparison, Alm (2009) explored a more comprehensive feature set for automatic affect prediction in text. Matykiewicz et al. (2009) discriminated between suicide notes and control texts using automatic clustering techniques, and discovered sub-clusters within suicide writings. In 2011, Pestian et al. (2012) organized a challenge to determine emotions and meaningful information in notes by suicide completers. These latter investigatory efforts, while valuable, involved computationally analyzing suicide notes of individuals with advanced rather than earlier stages of mental distress.

Our work links fundamental NLP classification methods with a standard engineering statistics method. Since the publication of "Building a Better Delivery System: A New Engineering/Health Care Partnership" by the Institute of Medicine (IOM) and the National Academy of Engineering (NAE) in 2005, there has been increased attention to the potential of engineering to broadly improve U.S. health care delivery. The IOM-NAE report identifies the use of optimization techniques to support decision making as one of the most promising engineering tools and technologies that could help the health care system deliver "safe, effective, timely, patient-centered, efficient, and equitable" care (Reid et al., 2005, p. 1).

3 Conceptual model

We conceptualize the task of determining affect state as a classification problem. Formally, let t denote a text that expresses an affect state. Let k be the number of affect state classes $C = \{c_1, c_2, c_3, \ldots, c_k\}$, where c_i denotes a specific class label. The goal is to decide a mapping function $f : t \rightarrow c_i$ to

obtain an ordered labeled pair (t, c_i). The mapping is based on $F_t = \{f_1, f_2, \ldots, f_n\}$, describing n feature values, automatically extracted from the text t.

The label hierarchy is shown below in Figure 1. The coarse-grained level represents a binary classification problem: *distressed* vs. *non-distressed*. At a more fine-grained level, we distinguish four classes (see section 4 below): *high distress*, *low distress*, *response*, and *happy*.

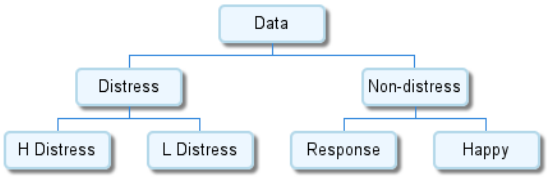

Figure 1. Class label hierarchy with two levels of granularity (binary vs. quaternary division of labels).

4 Dataset

There is currently no readily available text dataset for this problem. For this initial study, we prepared a small, annotated dataset of short written texts that represented relevant distinct, yet related, affect states. We manually collected a convenience (i.e., non-random) sample consisting of 200 posts from various public online forums dealing with mental well-being.[1] Forum posts were chosen because they are similar to other short digital social media texts, such as e-mails, online community posts, blog entries, or brief reflective writing that could be quickly gathered during a clinical session. We considered the text in the posts but not their titles.

4.1 Data annotation

Distressed and happy posts naturally divided into categories given the titles of the forums from which they were taken. Based on observation, we assumed that the distressed posts, all of which initiated new threads, were affectively distinct from *responses* to such threads, which had another polarity as they were meant to be reassuring and supportive. Therefore, we treated such responses as *non-distressed* posts. We recognize that a *response* represents a turn following an initial post. It is useful to explore how dialogic threading becomes part of affective language behaviors in social media (forums). The *happy* posts were included to represent the other extreme end of the affect spectrum.[2]

The dataset[3] was balanced such that 100 excerpts were *distressed*, 50 were *non-distressed responses*, and 50 were *non-distressed happy*. The *distressed* excerpts were then split further according to their distress intensity into *high* and *low* based on the annotator's perception, as seen in Figure 1. In an attempt to reduce personal bias, any post stating an active intent to harm someone or oneself was classified as *high distress*, while posts simply discussing bad feelings were usually classified as *low distress*. There were slightly more excerpts with low as opposed to high distress. Alm (2009) noted that expression of affect in language is often non-extreme. In a study of affective language in tales, Alm (2010) showed that affect is more often than not located in the gray zone between neutral and emotional. Table 1 shows the distribution of the excerpts according to four assigned class labels.

Class	Raw count and % of total excerpts
High Distress	39 (19.5%)
Low Distress	61 (30.5%)
Response	50 (25.0%)
Happy	50 (25.0%)
Total	**200 (100%)**

Table 1. Distribution of excerpts by four classes.

Figure 2 provides affect class distribution by source. As expected, subforum topic seems related

[1] Excerpts were culled from forums that dealt with mental health states at BreastCancer.org and reddit.com. Manually inspecting data ensured that relevant texts were included, but we also acknowledge that data obtained by such a selection process might differ from data obtained by random selection.

[2] Short *happy* post example: "I now have my foot in the door of the custom cake decorating business. I start in customer service as a cashier/barista, work my way through frosting, and then either into wedding, birthday, or sculpted cakes! I have been unemployed for 3 months now and this is huge. It means I can start saving money again, paying my bills and loans, and all the while doing something I love!"

[3] Posts were self-annotated according to the title of the forum to which they were submitted (e.g., *r/depression* posts as *distress*, and *r/happy* posts as *happy* and *non-distress*). Self-annotation acknowledges that people experience subjective differences in their tolerance levels for distress. Only distressed posts were perceptually sorted into high or low distress based on data observations. Texts were also inspected to block invalid posts, spam, or irrelevant responses.

to the distribution of intensity of distressed posts (high vs. low).

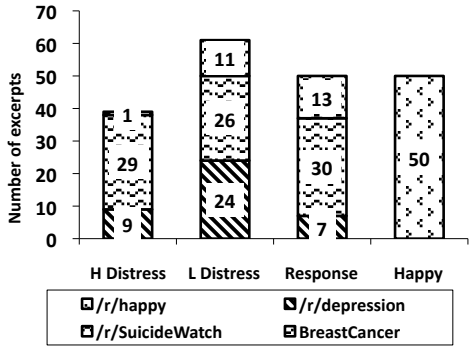

Figure 2. Excerpts by class and source.

5 Corpus linguistic analysis of dataset

Since this was an exploratory study, we conducted corpus linguistic analysis of the dataset by exploring descriptive statistics of linguistic and textual dimensions of the dataset.[4] As Table 2 shows, the collected corpus had 3,140 sentences, and totaled 49,850 words. There were on average 16 sentences or roughly 250 words in an excerpt.

Total excerpts	200
Total sentences	3,140
Total words	49,850
Average sentences per excerpt	15.70
Average words per excerpt	249.25
Average words per sentence	15.88

Table 2. Basic dataset statistics.

Table 3 shows basic statistics on text length.

Affect state and source	Sentences / excerpt	Words / excerpt
H Distress /r/SuicideWatch	19.8	300.0
H Distress /r/depression	31.1	399.7
L Distress breast cancer forum	16.5	297.5
L Distress /r/SuicideWatch	21.0	355.7
L Distress /r/depression	19.9	308.0
Response breast cancer forum	9.8	163.6
Response /r/SuicideWatch	14.3	218.2
Response /r/depression	13.4	219.9
Happy /r/happy	8.5	144.9

Table 3. Sentences and words per excerpt by affect state and source.

The statistics indicate that *happy* posts have the fewest sentences and words per excerpt, followed by the *responses*, ending with the *distressed* posts.[5]

In Table 4, we consider words per sentence as a metric independent of excerpt length, therefore avoiding potential selection bias. The average sentence length tended to be similar across forums.

Affect state and source	No. of excerpts	Words per sentence
H Distress /r/SuicideWatch	29	15.1
H Distress /r/depression	9	12.8
L Distress breast cancer forum	11	18.0
L Distress /r/SuicideWatch	26	16.9
L Distress /r/depression	24	15.5
Response breast cancer forum	13	16.6
Response /r/SuicideWatch	30	15.3
Response /r/depression	7	16.4
Happy /r/happy	50	17.1

Table 4. Length statistics by affect state and source.

We also examined exact lexical matches in polarity word lists,[6] with words having positive and negative connotation, which had been used before in Alm's work (2009). Positive words seemed favored in non-distressed posts (i.e., *responses* and *happy* posts). The opposite did not hold for *distressed* posts. Results are in Table 5.

We additionally examined the number of affect words present in each excerpt by considering four relevant affect word lists from Alm (2009), which were slightly expanded for this analysis (but less extensive than the polarity ones, yielding fewer matches overall).

Affect state and source	Positive	Negative
H Distress /r/SuicideWatch	18.0	20.0
H Distress /r/depression	24.0	24.0
L Distress breast cancer forum	21.0	15.0
L Distress /r/SuicideWatch	24.0	22.0
L Distress /r/depression	19.0	20.0
Response breast cancer forum	14.0	8.1
Response /r/SuicideWatch	17.0	13.0
Response /r/depression	17.0	13.0
Happy /r/happy	9.8	4.7

Table 5. Average polarity word list matches by affect state and source.

[4] We recognize that it would have been preferable to compute corpus statistics on a separate development dataset.

[5] Because only one BreastCancer.org post was classified as *high distress*, it was considered an outlier and thus excluded in presenting and discussing these tables.

[6] Positive and negative word lists contained 1915 and 2294 lexical items, respectively.

The average numbers of exact lexical matches from the word lists in all excerpts are shown in Table 6. For each affect word list (cf. columns), the highest and lowest values are in bold font. Table 6 shows that the number of average matches was low overall, and that in general, there were more matches with *sad* and *afraid* wordlists. However, *happy* posts showed slightly more overlap with the *happy* word list.

Affect state and source	Happy	Sad	Afraid	Angry
H Distress /r/SuicideWatch	0.9	1.8	2.1	**1.0**
H Distress /r/depression	1.1	**3.6**	3.3	**1.0**
L Distress breast cancer forum	**1.8**	1.6	1.9	0.5
L Distress /r/SuicideWatch	1.5	2.9	**4.0**	0.7
L Distress /r/depression	1.4	2.5	2.7	0.8
Response breast cancer forum	1.2	0.5	1.0	0.3
Response /r/SuicideWatch	1.3	2.0	2.4	0.5
Response /r/depression	**0.6**	1.1	1.3	**0.0**
Happy /r/happy	1.4	**0.4**	**0.5**	0.1

Table 6. Average emotion word list matches by affect state and source.

Lastly, because pronouns have been found important for linguistic analysis of mental health disorders or socio-cognitive processes (e.g., Andreasen and Pfohl, 1976; Pennebaker 2011), we explored this in the dataset based on the part of speech output from an NLTK-based tagger (Bird et al., 2009). Table 7 shows percentages of first-, second-, and third-person pronouns in the dataset.

Affect state and source	1st person	2nd person	3rd person
H Distress /r/SuicideWatch	77.1	0.9	22.0
H Distress /r/depression	56.1	12.0	31.9
L Distress breast cancer forum	63.0	10.9	26.1
L Distress /r/SuicideWatch	68.6	1.6	29.8
L Distress /r/depression	76.9	1.6	21.5
Response breast cancer forum	39.1	33.1	27.8
Response /r/SuicideWatch	23.1	46.1	30.8
Response /r/depression	21.3	56.9	21.8
Happy /r/happy	72.1	4.1	23.8

Table 7. % pronoun by person, affect state, and source.

There were few second-person pronouns in *distressed* and *happy* posts, but more in the responses, which had fewer first-person pronouns. This observation confirms that distressed and happy posts are *self-oriented*, but that responses, which reassure and reply to a thread initiator, are *other-oriented*. Perspective is thus another meaningful dimension of this affect dataset.

6 Computational modeling experiments

This initial study used three fundamental supervised classification methods: *Naïve Bayes*, *Maximum Entropy*, and *Decision Tree* (Bird et al., 2009). These allowed us to derive initial reference values which can be improved upon with more advanced techniques in future work. We also provide results for a fourth approach, Perkins' *Max Vote* method (2010), using the other three algorithms' predictions to give a joint prediction.

6.1 Feature set used for modeling

We developed a set of features based on the scholarly literature (e.g., Alm, 2009; Andreasen and Pfohl, 1976; Endres, 2004; Yu et al., 2009). The following features were automatically extracted from text, using Python, NLTK (Bird et al., 2009), and Perkins (2010): "bag of words" (BOW) with unique unigrams; excerpt length in sentences; excerpt and sentence lengths in words; positive vs. negative polarity word list matches; happy, sad, afraid, and angry affect word list matches; first-, second-, and third-person pronouns; and, finally, nouns, verbs, adjectives, adverbs, and pronouns.[7] Most features were initially examined both as a raw number and as a per sentence average. Features were discretized by considering how they deviated (more vs. less) from average values calculated from the corpus as a whole.[8] This resulted in 42 distinct feature types. Feature extraction was conducted the same way for train and test sets.

[7] Part of speech ratios were included due to an indication by Fraser et al. (1986) that verb patterns could be useful in discriminating manic patients from schizophrenics and the control group.

[8] The absence of a separate dataset for computing the averages allows a possibility of overfitting the data. However, we assume the averages are representative for similar texts and will be useful in future expanded model development.

6.2 Experiment 1: Classification at two levels

The computational experimental process is illustrated in Figure 3. In these experiments, the dataset is initially randomized and then evaluated with *k*-fold cross-validation, by repeating the classification process *k* times. Performance is thus reported as the average over *k* accuracy scores. The experiment explored five scenarios with $k = \{5, 10, 20, 100, 200\}$. The last scenario corresponds to a leave-one-out cross-validation (i.e., where the train set consists of *(N-1)* instances and the test set of one instance, and the procedure is repeated *N* times, where *N* is the total instances in the dataset).

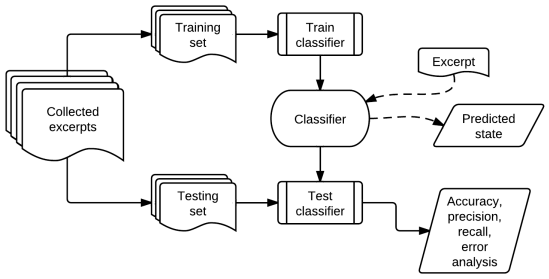

Figure 3. Computational experimentation process.

Figure 4 shows the accuracy for the coarse-grained binary classification problem which involved assigning either a *distressed* or a *non-distressed* label to a text excerpt. The majority class baseline for this is 50%, as half of the excerpts belonged to each of the two classes. Figure 4 shows that the classifiers average performance has a stable range with around 73-76% accuracy, across varying k-folds and across algorithms. This performance improves more than 20% over the majority class baseline, which is indicated by a line in Figure 4.

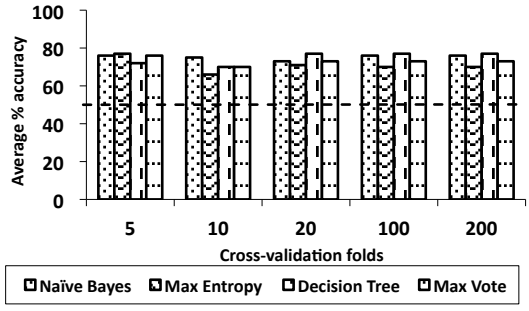

Figure 4. Classification accuracy for the coarse-grained classification scenario that considers two affect states: *distressed* and *non-distressed*.

Next, Figure 5 shows the results for classification at the fine-grained level which considers four affect classes: *high distress*, *low distress*, *response*, and *happy*. Here the majority class baseline is 30.5%. Four states yield around 54-57% accuracy. Again, that is more than a 20% improvement over the majority class baseline. The exception is Maximum Entropy, which performs poorly on this classification task.

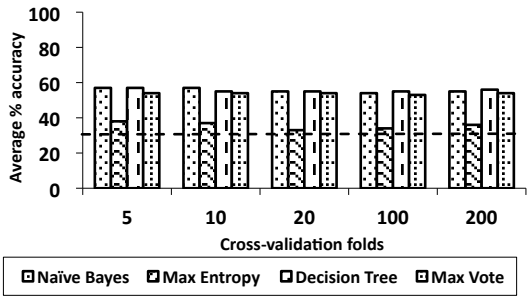

Figure 5. Classification accuracy for the fine-grained classification scenario that considers four states.

Inspecting the most relevant features from runs over the course of the study indicates that the number of second-person pronouns, which usually identified responses, and the number of verbs and fearful affect words per sentence are particularly important. In responding to a post, one uses more second-person pronouns in order to address the original poster. Again, this indicates that turn-taking impacts affective language behaviors.

A confusion matrix in Table 8 shows misclassification results for a select test fold of fine-grained classification. The shaded cells along the diagonal show how often the model correctly predicted an affect state. The other cells show where the model misclassified the affect state.

	Predicted			
Actual	H Distress	L Distress	Response	Happy
H Distress	**7.6%**	3.0%	1.5%	3.0%
L Distress	7.6%	**4.5%**	4.5%	9.1%
Response	.	.	**28.8%**	.
Happy	10.6%	3.0%	3.0%	**13.6%**

Table 8. A select confusion matrix.[9]

Looking at the *response* class, for example, the classifier correctly classified all of the actual

[9] This table shows results from a single test of a classifier. Due to the random test set, totals do not match the corpus totals.

response excerpts. This is likely due to the importance of second-person pronouns found in particular in the *response* excerpts. However, the classifier incorrectly labeled some excerpts in each of the other classes as *response*. Although this classifier was not as accurate for the other affect classes, the accurate option was the most commonly predicted class for both *high distress* and *happy*. This was not the case for *low distress*, however, which was more often predicted as *high distress* or *happy*. This can reflect the challenge of affect analysis in the gray zone between affect and neutrality, as lower emotional intensity decreases perceptual clarity. This finding is consistent with the previous literature, discussed above. A way to deal with this issue is to combine text analysis with other data analysis.

6.3 Experiment 2: Ablation study

An ablation study was performed to assess the accuracy with different features given the four fine-grained classes, using a $k = 5$ cross-validation. We ignore bag of words, which can result in many sparse features, to examine other types.

In Table 9, the first ablation step represents only length variables; the second adds polarity variables; the third adds affect variables; the fourth adds pronoun variables; and the fifth adds part of speech variables (in each case, to the features added in previous steps). Each test was done on all four supervised classification algorithms.

The results with this split of train and test data show that each addition to the feature set improved the accuracy of the model's predictions, except the part-of-speech features. This could be due to the particular data split, the order of the ablation steps, or the ablation feature groupings. Additionally, excluding BOW features did not have a clear negative effect on performance. Considering only length averaged 25.1% accuracy across classifiers; adding five feature types resulted in 54.5%.

	Classifier type				
	NB	**ME**	**DT**	**MV**	**Mean**
Length	.260	.225	.255	.265	.251
+ polarity	.295	.295	.390	.320	.325
+ affect	.430	.395	.365	.415	.401
+ pronouns	.590	.530	.485	.580	.546
+ POS	.595	.505	.505	.575	.545

Table 9. Ablation study results: four affect states fine-grained classification scenario (NB=Naïve Bayes, ME=Max Entropy, DT=Decision Tree, MV=MaxVote).

7 Engineering statistics applied to NLP

Choosing the right feature set remains a difficult, poorly understood process. Here, we report on a separate analysis using a 2^k factorial design, which is a common method from engineering statistics that can be used to quantitatively and systematically determine the effect and interactions that different linguistic feature types have on the assessment of the affect state of a text.[10] The outcome of this factorial design is a response formula that can be used to classify excerpts.

A 2^k experimental design assumes that a decision maker wants to determine how to express the effect of k different factors and their interactions on a response of interest. Given that the factors can take any possible value, the number of necessary experiments to statistically deduce such an expression can be quite large and expensive. Instead, a 2^k design limits each factor to only two levels (a high and a low value). The minimum number of experiments needed to deduce a model that explains the direct and interaction effects of k factors is 2^k. For example, a problem in which 5 factors are assumed to affect the value of a response requires executing $2^5=32$ experiments, each with a unique arrangement of factor levels. Replications of these experiments are recommended to increase accuracy in the estimation of the term coefficients.

Having 42 candidate linguistic features that could influence an evaluator's decisions to categorize the distress state of a text would have required at least 2^{42} (over 4 trillion!) tests with different configurations of features. Therefore, we grouped related linguistic and textual features into five *super-factors*. For example, sentences per excerpt, words per excerpt, and words per sentence were all combined into a *length factor*. The super-factors chosen were: y_1 = length, y_2 = polarity, y_3 = affect, y_4 = pronoun, and y_5 = parts of speech.[11] Using five super-factors resulted in 32 (2^5) possible experimental combinations.

We assessed the 200 text excerpts based on all 42 linguistic features to get a numerical value for

[10] We adapt the regular terminology used in engineering statistics for discussing this approach. This means that k is used in a different sense in this section compared to above.

[11] BOW features were excluded here as well. The ablation study in section 6.3 also justifies their exclusion.

each super-factor. We then labeled each of these numerical values as *high* or *low*, based on the median of all 200 values for each factor and for each text. The super-factor label combinations for each of the 200 excerpts were then mapped to these 32 possible combinations. This mapping was used to generate a response formula (similar to a multi-attribute regression expression) that found the direct effect of the super-factors and their interactions on the distress evaluation.

We found that three main effects of the super-factors and four of their interactions were statistically significant. The significant super-factors were *affect*, *pronoun*, and *part of speech*. Although the main effect of *length* was not significant, its interactions with the *affect* and *pronoun* super-factors were significant.

The obtained expression for predicting the class of an excerpt is below. Each factor is a positive or negative 1, for high or low values, respectively:

$$\text{Response} = -0.377 + 0.2062y_3 + 0.355y_4 - 0.276y_5 + 0.1983y_1y_3 + 0.1928y_3y_4 - 0.197y_1y_3y_4 - 0.1704y_1y_3y_4y_5$$

Responses can range from –2 to 1, with –2 predicting *high distress*, –1 predicting *low distress*, 0 predicting *response*, and 1 predicting *happy*. This response formula could be tested as a prediction method on future data not used in its estimation.

We further propose using the 2^k factorial mechanism to systematically reduce the super-factors into simpler features. For example, because one of the super-factors did not show a significant main effect, we can assume that its linguistic features do not individually reflect distress or non-distress. Thus, one could reconfigure new super-features, assigning new values to the 200 excerpts, and repeat the analysis and remove any super-feature whose main and secondary effects are not significant. This iterative process should halt when we have new, redefined super-features that are significant in predicting the distressed and non-distressed states of the 200 excerpts. An analysis of residuals will serve as a control mechanism to reduce the number of iterations in the process.

8 Conclusion

If there were a way to automatically identify individuals with undiagnosed mental illnesses, it would be possible to recommend a clinical visit. The problem addressed by this paper was how to discriminate related affect states via computational linguistic analysis of short online writings.

We reported on an initial dataset from forums and corpus linguistic analysis, and found patterns in the data that merit further study. To predict distress states, we used supervised classification and explored super-features' importance with a 2^k factorial design, an engineering statistics method. We approach this problem from a linguistic perspective and pay extra attention to linguistic analysis and how distress is linguistically encoded. Not only do we report on effects by forum, distress state, emotion and polarity lexicon, etc., but our 2^k factorial analysis also rigorously clarifies which linguistic feature types contribute in statistically significant ways. Additionally, the ablation study conducted largely verified these findings.

Leave-one-out cross-validation is common with small datasets; we also show that varying *k* in the cross-validation does not impact results. There are benefits with smaller datasets and shorter texts. In clinical settings, data can be especially hard to obtain, and it is useful to understand the limitations and affordances of modeling with limited data. Similarly, it is important to understand how models perform on fundamental algorithms and shallow features extracted from text that can generalize to, for example, resource-poor languages.

While this data was adequate for exploratory investigation, a larger, clinical dataset would be less prone to selection bias. Combining text with other analysis information seems key in future work. Also, more advanced algorithms could yield more accurate predictions, as could iterations of the 2^k factorial analysis. Other aspects left for future study include the relationship between the individual affect states and their predictive linguistic features and experimentation with unbalanced data scenarios. Lastly, another area to pursue is using affect features for identifying linguistic patterns unique to online communication.

Acknowledgments

This work was supported by an RIT Seed Funding Award. We thank anonymous reviewers for comments. We also thank W. McCoy and R. Lehrman.

References

Cecilia Ovesdotter Alm. 2009. Characteristics of high agreement affect annotation in text. *Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, Uppsala, Sweden, 15-16 July 2010*, 118-122.

Cecilia Ovesdotter Alm. 2009. *Affect in Text and Speech*. VDM Verlag, Saarbrücken.

Nancy J.C. Andreasen and Bruce Pfohl. 1976. Linguistic analysis of speech in affective disorders. *Archives of General Psychiatry*, 33:1361-1367.

Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural Language Processing with Python. O'Reilly Media, Sebastopol, CA. *Natural Language Toolkit*, http://www.nltk.org/book.

Anna Dixon, David McDaid, Martin Knapp, and Claire Curran. 2006. Financing mental health services in low and middle income countries: equity and efficiency concerns. *Health Policy Plan*, 21:71-82.

Johann Endres. 2004. The language of the psychopath: characteristics of prisoners' performance in a sentence completion test. *Criminal Behaviour and Mental Health*, 14:214-226.

Horacio Fabrega, Jr. 1991. Psychiatric stigma in non-Western societies. *Comprehensive Psychiatry*, 32:534-551.

William I. Fraser, Kathleen M. King, Philip Thomas, and Robert E. Kendell. 1986. The diagnosis of schizophrenia by language analysis. *British Journal of Psychiatry*, 148:275-278.

Barbara Gawda. 2010. Syntax of emotional narratives of persons diagnosed with antisocial personality. *Journal of Psycholinguistic Research*, 39:273-283.

Philip D. Harvey. 1983. Speech competence in manic and schizophrenic psychoses: The association between clinically rated thought disorder and cohesion and reference performance. *Journal of Abnormal Psychology*, 92(3):368-377.

Mukund Jha and Noémie Elhadad. 2010. Cancer stage prediction based on patient online discourse. *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010, Uppsala, Sweden*, 64-71.

Ronald C. Kessler, Olga Demler, Richard G. Frank, et al. 2005. Prevalence and treatment of mental disorders, 1990 to 2003. *New England Journal of Medicine*, 352:2515-2523.

Hsiang-Ching Kung, Donna L. Hoyert, Jiaquan Xu, and Sherry L. Murphy. 2008. *Deaths: Final data for 2005*. *National Vital Statistics Report*, 56:1-121.

Kevin Lerman, Ari Gilder, Mark Dredze, and Fernando Pereira. 2008. Reading the markets: Forecasting public opinion of political candidates by news analysis. *Proceedings of the 22rd International Conference on Computational Linguistics (Coling 2008)*, 473-480.

Mark Lutz. 2009. *Learning Python* (4th Edition). O'Reilly Media, Sebastopol, CA.

Pawel Matykiewicz, Wlodzilaw Duch, and John P. Pestian. 2009. Clustering semantic spaces of spaces of suicide notes and newsgroup articles. *Proceedings of the Workshop on BioNLP, Boulder, Colorado*, 179-184.

Kathleen M. Michels, Karen J. Hofman, Gerald T. Keusch, Sharon H. Hrynhow. 2006. Stigma and global health: Looking forward. *Lancet*, 367:538-539.

Alex J. Mitchell, Amol Vaze, and Sanjay Rao. 2009. Clinical diagnosis of depression in primary care: A meta analysis. *Lancet*, 374:609-619.

Elias Mossialos, Anna Dixon, Josep Figueras, and Joe Kutzin. 2002. *Funding Health Care: Options for Europe*. Open University Press, Buckingham, UK.

James W. Pennebaker. 2011. *The Secret Life of Pronouns: What our Words Say about us.* Bloomsbury Press, New York.

Jacob Perkins. 2010. *Python Text Processing with NLTK 2.0 Cookbook*. Packt Publishing, Birmingham.

John P. Pestian, Pawel Matykiewicz, Michelle Linn-Gus, Brett South, Ozlem Uzner, Jan Wiebe, Kevin B. Cohen, and Christopher Brew. (2012). Sentiment analysis of suicide notes: A shared task. *Biomedical Informatics Insights*. 5 (Suppl. 1), 3-16.

John P. Pestian, Pawel Matykiewicz, and Jacqueline Grupp-Phelan. 2008. Using natural language processing to classify suicide notes. *BioNLP 2008: Current Trends in Biomedical Natural Language Processing, Columbus, Ohio*, 96-97.

Ann Barnett Ragin and Thomas F. Oltmanns. 1983. Predictability as an index of impaired verbal communication in schizophrenic and affective disorders. *British Journal of Psychiatry*, 143:578-583.

Proctor P. Reid, W. Dale Compton, Jerome H. Grossman, and Gary Fanjiang. 2005. *Building a Better Delivery System: A New Engineering/Health Care Partnership*. The National Academies Press.

Brian Roark, Margaret Mitchell, and Kristy Hollingshead. 2007. Syntactic complexity measures

for detecting Mild Cognitive Impairment. *BioNLP 2007: Biological, translational, and clinical language processing, Prague*, 1-8.

Shekhar Saxena, Graham Thornicroft, Martin Knapp, and Harvey Whiteford. 2007. Resources for mental health: scarcity, inequity, and inefficiency. *Lancet*, 370:878-889.

Robert P. Schumaker. 2010. An analysis of verbs in financial news articles and their impact on stock price. *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, Los Angeles, California*, 3-4.

David M. Shepard, Michael C. Ferris, Gustavo H. Olivera, and T. Rockweel Mackie. 1999. Optimizing the delivery of radiation therapy to cancer patients. *SIAM Review*, 41(4):721-744.

Harry A. Sultz and Kristina M. Young. 2011. *Health care USA: Understanding its Organization and Delivery* (7th Edition). Jones and Barlett Learning, LLC, Sudbury.

C. Turrina, R. Caruso, R. Este, et al. 1994. Affective disorders among elderly general practice patients: a two-phase survey in Brescia, Italy. *British Journal of Psychiatry*, 165:533-537.

World Health Organization. 2005. *Mental Health Atlas*. WHO, Geneva, Switzerland.

Liang-Chih Yu, Chien-Lung Chan, Chung-Hsien Wu, and Chao-Cheng Lin. 2009. Mining association language patterns for negative life event classification. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 201-204.

Yitao Zhang and Jon Patrick. 2006. Extracting patient clinical profiles from case reports. *Proceedings of the 2006 Australasian Language Technology Workshop*, 167-168.