Toward Macro-Insights for Suicide Prevention: Analyzing Fine-Grained Distress at Scale

Christopher M. Homan[1] **Ravdeep Johar**[1] **Tong Liu**[1]
Megan Lytle[2] **Vincent Silenzio**[2] **Cecilia O. Alm**[3]
[1] Golisano College of Computing and Information Sciences, Rochester Institute of Technology
[2] Department of Psychiatry, University of Rochester Medical Center
[3] College of Liberal Arts, Rochester Institute of Technology
{cmh§ | rsj7209$ | tl8313$ | Megan_Lytle† | Vincent_Silenzio† | coagla$}
§@cs.rit.edu $@rit.edu †@urmc.rochester.edu

Abstract

Suicide is a leading cause of death in the United States. One of the major challenges to suicide prevention is that those who may be most at risk cannot be relied upon to report their conditions to clinicians. This paper takes an initial step toward the automatic detection of suicidal risk factors through social media activity, with no reliance on self-reporting. We consider the performance of annotators with various degrees of expertise in suicide prevention at annotating microblog data for the purpose of training text-based models for detecting suicide risk behaviors. Consistent with crowdsourcing literature, we found that novice-novice annotator pairs underperform expert annotators and outperform automatic lexical analysis tools, such as Linguistic Inquiry and Word Count.

1 Introduction

Suicide is among the leading causes of death for individuals 10–44 years of age in the United States (Heron and Tejada-Vera, 2009). Indeed, while mortality rates for most illnesses decreased between 2008 and 2009, the rate of suicide increased by 2.4% (Heron and Tejada-Vera, 2009). The lifetime prevalence for suicidal ideation is 5.6–14.3% in the general population, and as high as 19.8–24.0% among youth (Nock et al., 2008).

The first step toward suicide *prevention* is to identify, ideally in consultation with clinical experts, the risk factors associated with suicide. Due to social stigma among other sociocultural factors (Crosby et al., 2011), individuals at risk for committing suicide may not always reach out to professionals or, if they do, provide them with accurate information. They may not even realize their own level of suicide risk before it is too late. Self-reporting, then, is not an entirely reliable means of detecting and assessing suicide risk, and research on suicide prevention can benefit from also exploring other channels for assessing risk.

For instance, individuals may be more inclined to seek support from informal resources, such as social media, instead of seeking treatment (Crosby et al., 2011; Bruffaerts et al., 2011; Ryan et al., 2010). Evidence suggests that youth and emerging adults usually prefer to seek help from their friends and families; however, higher levels of suicidal ideation are associated with lower levels of help-seeking from both formal or informal resources (Deane et al., 2001).

These patterns in help-seeking behavior suggest that social media might be an important channel for discovering those at risk for—and even preventing—suicide. Internet- and telecommunications-driven activity is revolutionizing the social sciences by providing data, much of it publicly available, on human activity in situ, at volumes and a level of time and space granularity never before approached. Can such data improve clinical preventative study and measures by providing access to at-risk individuals who would otherwise go undetected, and by leading to better science about suicide risk behaviors?

The stress-diathesis model for suicidal behavior (Mann et al., 1999) suggests that they might. It says that (1) objective states, such as depression or life events, as well as subjective states and traits, such as substance abuse or family history of depression, suicide, or substance abuse, are among the risk factors that contribute to suicidal ideation and (2) the presence of these factors could eventually lead to either externalizing (e.g., interper-

Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 107–117,
Baltimore, Maryland USA, June 27, 2014. ©2014 Association for Computational Linguistics

sonal violence) or internalizing aggression (e.g., attempting suicide).

Since the stress-diathesis model was developed using risk factors for suicidal behavior, and because it makes a connection between internalized and externalized acts, it is a suitable framework for analyzing publicly available linguistic data from social media outlets such as Twitter. Data from social media can be seen as a kind of natural experiment on depression and suicidal ideation that is unburdened by such sample biases as the willingness of individuals to take part in research and/or seek out formal sources of support. Moreover, this approach may provide information about individuals who are unlikely to engage in formal help-seeking behaviors, or may inform effective methods of natural helping. Thus, this macro-level approach to monitoring suicidal behaviors may have future implications not only for identifying individuals who have a higher prevalence for suicidal behaviors but it could eventually lead to additional methods for enhancing protective factors against suicide.

In this paper, we take steps toward the automatic detection of suicide risk among individuals via social media. Suicide ideation is a complex behavior and its connection to suicide itself remains poorly understood. We focus on a particular aspect of suicidality, namely *distress*. While not equivalent to suicide ideation, according to Nock et al. (2010) distress is an important risk factor in suicide, and one that is observable from microblog text, though admittedly observing suicide risk behavior is a subjective and noisy venture.

Lehrman et al. (2012) conducted an early study on the computational modeling of distress based on short forum texts, yet left many areas wide open for continued study. For example, analysis at scale is one such open issue. More specifically, Pestian and colleagues (Matykiewicz et al., 2009; Pestian et al., 2008) used computational methods to understand suicide notes. However, when it comes to preventive contexts, such data are less insightful. For preventive health, access to real-time health-related data that dynamically evolve can allow us to address macro-level analysis. Social media provide an additional opportunity to model the phenomena of interest at scale.

We use methods that take advantage of lexical analysis to retrieve microblog posts (tweets) from Twitter and compare the performance of human annotators—one being an expert, and others not—to rate the level of distress of each tweet.

Clinical expert annotation, rather than general-purpose tools for content and sentiment analysis such as LIWC (Linguistic Inquiry and Word Count) by Pennebaker et al. (2001), provides a basis for text-based statistical modeling. We show that expertise-based keyword retrieval, departing from knowledge about contributing risk factors, results in better interannotator agreement in both novice-novice and novice-expert annotation when the keywords reflect the task at hand.

2 Related Work

Data on suicide traditionally comes from healthcare organizations, large-scale studies, or self reporting (Crosby et al., 2011; Horowitz and Ballard, 2009). These sources are limited by sociocultural barriers (Crosby et al., 2011), such as stigma and shame. Moreover, data on suicide is never particularly reliable because suicide is a fundamentally subjective, complex phenomenon with a low base rate. For these reasons, many researchers tend to focus on the relationship between risk factors and suicidal behavior, without relying heavily on theoretical models (Nock et al., 2008).

Approximately one-third of all individuals who reported suicidal ideation in their lifetime made a plan to commit suicide. Nearly three-quarters of those who reported making a suicide plan actually attempted. The odds of attempting suicide increased exponentially when individuals endorsed three or more risk factors, e.g., having a mood or substance abuse disorder (Kessler et al., 1999).

Demographics, previous suicide attempts, mental health concerns (i.e., depression, substance abuse, suicidal ideation, self-harm, or impulsivity), family history of suicide, interpersonal conflicts (i.e., family violence or bullying), and means for suicidal behavior (e.g., firearms), are commonly cited risk factors for suicidal behavior (Nock et al., 2008; Crosby et al., 2011; Gaynes et al., 2004; Harriss and Hawton, 2005; Shaffer et al., 2004; Brown et al., 2000).

Regarding the use of annotation for predictive modeling, evidence suggests that when it comes to judgments that involve clinical phenomena, experts and novices behave differently (Li et al., 2012; Womack et al., 2012). Such distinctions intuitively make sense, as the learning of medical domain knowledge requires advanced education in conjunction with substantial practical field experience.

In a task such as medical image inspection, the subtle cues that point an observer to evidence that allow them to identify a clinical condition, while accessible to experts with training and perceptual expertise to guide their exploration, are likely to be missed by novices who lack that background and clinical understanding. Such expertise can then be integrated into human-centered health-IT systems (Guo et al., 2014), in order to introduce novel ways to retrieve medical images and take advantage of an understanding of which information is useful. It is reasonable to assume that this knowledge gap also applies to other knowledge-intensive clinical domains such as mental health. In this study, we explore this question and study if novice vs. expert annotation makes a difference for identifying distress in social media texts, as well as what the impact of expert vs. novice annotation is for subsequent computational modeling with the annotated data.

Affect in language is a phenomenon that has been studied in the speech and text analysis domains, and in many others (Calvo and D'Mello, 2010). Clearly, emotion is a key element in the human experience, but it is notoriously difficult to pin down and scholars in the affective sciences lack a single agreed-upon definition for emotion. Accordingly, different theoretical constructs have been proposed to describe affect and affect-related behaviors (Picard, 1997). In addition, research on affect in language has shown that such phenomena tend to be subjective, lack real ground truth (often resulting in moderate kappa scores), and have particularly fuzzy semantics in the gray zone where neutrality and emotion meet (Alm, 2008). These kinds of problem characteristics bring with them their own set of demanding challenges from a computational perspective (Alm, 2011). Yet, the nature of such problems make them incredibly important to study, despite the challenges involved.

Sentiment analysis has been widely studied in a number of computational settings, including on various social networking sites. A rather substantial body of work already exists on the use of Twitter to study emotion (Bollen et al., 2011b; Dodds et al., 2011; Wang et al., 2012; Pfitzner et al., 2012; Kim et al., 2012; Bollen et al., 2011a; Pfitzner et al., 2012; Bollen et al., 2011c; Mohammad, 2012; Golder and Macy, 2011; De Choudhury et al., 2012a; De Choudhury et al., 2012b; De Choudhury et al., 2013; De Choudhury and Counts, 2013; Hannak et al., 2012; Thelwall et al., 2011; Pak and Paroubek, 2010). For instance, Golder and and Macy study aggregate global trends in "mood," and show, among other things, that people wake up in a relatively good mood that decays as the day progresses (Golder and Macy, 2011). Bollen et al. (2011c) show that tweets from users who took a standard diagnostic instrument for mood are often tied to current events, such as elections and holidays.

Relatively little of this work has focused on suicide or related psychological conditions. Masuda et al. (2013) study suicide on mixi (a Japanese social networking service). Cheng et al. (2012) consider the ethical and political implications of online data collection for suicide prevention. Jashinsky et al. (2013) show correlations between frequency in tweets related to suicide and actual suicide in the 50 United States of America. Sadilek et al. (2014) study depression on Twitter. De Choudhury and collaborators studied depression—in general and post-partum—in Twitter (De Choudhury et al., 2012a; De Choudhury et al., 2012b; De Choudhury et al., 2013; De Choudhury and Counts, 2013) and Facebook (De Choudhury et al., 2014). Homan et al. (2014) investigate depression in TrevorSpace. A number of social theories of suicide have been proposed (Wray et al., 2011), but most of this work was with respect to offline social systems.

3 Methods

Our methods involve four main phases: (1) We filtered a corpus, obtained from Sadilek et al. (2012), of approximately 2.5 million tweets from 6,237 unique users in the New York City area that were sent during a 1-month period between May and June, 2010, into a set of 2,000 tweets that are relatively likely to be centered around suicide risk factors. (2) We annotated each of these 2,000 tweets with their level of distress, and also analyzed the annotations in detail. (3) We then trained support vector machines and topic models with the annotated data, except for a held-out subset of 200 tweets. (4) Finally, we assessed the effectiveness of these methods on the held-out data.

Source tweets	Number of tweets	2,535,706
	Unique geo-active users	6,237
	“Follows” relationships	102,739
	“Friends” relationships	31,874
Filtered tweets	Number of tweets	2,000
	Unique users	1,467
	Unique unigrams	1,714,167
	Unique bigrams	9,246,715
	Unique trigrams	1,306,1142
Categories distribution	LIWC sad	1,370
	Depressive feeling	283
	Suicide ideation	123
	Depression symptoms	72
	Self harm	67
	Family violence/discord	47
	Bullying	10
	Gun ownership	10
	Drug abuse	6
	Impulsivity	6
	Prior suicide attempts	2
	Suicide around individual	2
	Psychological disorders	2

Table 1: Summary statistics and thematic category distributions of the collected dataset. The data were collected from NYC. Geo-active users are those who geo-tag (i.e., automatically post the GPS location of) their tweets relatively frequently (more than 100 times per month).

3.1 Filtering tweets

In order to facilitate the discovery of distress-related tweets, we first (a) converted all text to lower case; (b) stripped out punctuation and special characters; and (c) mapped informal terms (such as abbreviations and netspeak) to more standard ones, based on the noslang dictionary.[1]

We then used two different methods to filter tweets that are relatively likely to center on suicide risk factors. We used LIWC to capture 1,370 tweets by sampling randomly from among the 2,000 tweets with the highest LIWC sad score. LIWC has been widely used to estimate emotion in online social networks, and specifically to mood on Twitter. This slight amount of randomness in filtering tweets this way was intended to avoid selecting obvious false positives, such as the use of “sad” in nicknames.

Next, we adopted a collection of inclusive search terms/phrases from Jashinsky et al. (2013), which was designed specifically for capturing tweets related to suicide risk factors, and applied them to our source corpus. We added to these more terms, from (Crosby et al., 2011) (see Table 2). These terms yielded 630 tweets.

depressive feeling	tired of living, leave this world, wanna die, hate my job, feeling guilty, deserve to die, desire to end own life, feeling ignored, tired of everything, feeling blue, have blues
depression symptoms	sleeping pill, have insomnia, sleep forever, sleep disorder
drug abuse	clonazepam, drug overdose, imipramine
prior suicide attempts	tried suicide
suicide ideation	commit suicide, committing suicide, feeling suicidal, want to suicide, shoot myself, a gun to head, hang myself, intention to die
self harm	hurt myself, cut myself
psychological disorders	sleep apnea
family violence discord	lost my friend, argument with wife, argument with husband, shouted at each other

Table 2: Filtering terms added to those from Jashinsky et al. (2013).

3.2 Novice and Expert Tweet Annotation

We then divided the resulting set of 2,000 filtered tweets (1,370 from the LIWC sad dimension and 630 from suicide-specific search terms), into two randomized sets of 1,000 tweets each. Both sets had the same proportion of LIWC-filtered and suicide-specific-filtered tweets. A novice annotated the first set and a counseling psychologist with experience in suicide related research annotated the second set. A second novice annotated a subset of 250 tweets of the first set, to reveal interannotator agreement between novices, as one might expect a novice without training to be less systematic. (The annotators were among the authors.) Each tweet in each set was rated on a four-point scale (H, ND, LD, HD) according to the level of distress evident (Table 3).

Each tweet to be annotated was provided with context in the form of the three tweets before and after the tweet to be annotated that the tweeter made, along with the timestamp of those tweets and the thematic categories to which the tweet belonged, based on the filtering process (Figure 1).

3.3 Modeling

We then mapped each tweet to a feature space composed of the unigrams, bigrams, and trigrams in the corpus. For example, a simple tweet “I am

[1] `http://www.noslang.com/dictionary`

```
978: Date: XXXX
    -3: dat man on maury is overreacting!!
        he juss doin dat cuz he on
        tv [-0:24:39]
    -2: @XXXX cedes!!! [-0:21:25]
    -1: yesssss! da weatherman was wronq
        no rainy ass prom days!! yesss
         prom is 2day guys!! class
         of 2010! [-0:02:56]
>>> @XXXX awwww thanks trae-trae
     1: rt @XXXX: abt 2 hop in a kab
        to skool i wouldn't dare spend
        over 2 dollars to get somewhere
        i dnt wanna be n da first
        place! [+0:00:57]
     2: @XXXX yeaa [+0:03:59]
     3: @XXXX wassup? [+0:05:28]
Msg_id: XXXX  [Distress: ND, LIWC Sad: No]
```

Figure 1: Example input for annotator. The tweet to be annotated is indicated by >>>. Annotators were given context in the form of the three tweets immediately preceding—and the three tweets immediately following—the tweet to be annotated that the tweeter made, along with the relative time at which each tweet was made. Each numerical label denotes one of these context tweets. (Tweeter information has been blanked out.)

Code	Distress Level
H	happy
ND	no distress
LD	low distress
HD	high distress

Table 3: Distress-related categories used to annotate the tweets.

so happy" was represented as the following *feature vector*: {I, am, so, happy, I am, am so, so happy, I am so, am so happy}. Each feature is associated with its tf-idf score (Manning et al., 2008).

We performed topic modeling on our dataset. A *topic* is a set of lexical items that are likely to occur in the same tweet. Topic models are capable of associating words with similar meanings and distinguishing among the different meanings of a single word. We used latent Dirichlet allocation (LDA) (Blei et al., 2003) to create these topics. Before doing so, we removed stop words and words that occur only once in the dataset. We then applied LDA algorithm on the data to discover three topics using 100 iterations.

We used support vector machines (SVMs) (Joachims, 1998), a machine learning method that is used to train a classification model that can assign class labels to previously unseen tweets, to assess the power of our annotations. SVMs treat each tweet as a point in an extremely high dimensional space (one dimension per uni-, bi-, and trigram in the corpus). SVMs are a form of *linear separator* that can also distinguish between nonlinearly separable classes of data by warping the feature space (though in our case we perform no such warping, or *kernelization*). They have proven to be an extremely effective tool in classifying text in numerous settings, including Twitter.

4 Results

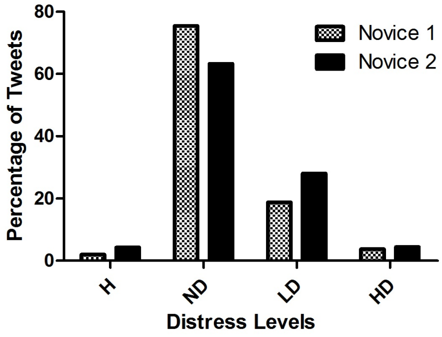

Figure 2: Distribution of distress level annotations on the tweets annotated by Novices 1 and 2 (N=250, identical set).

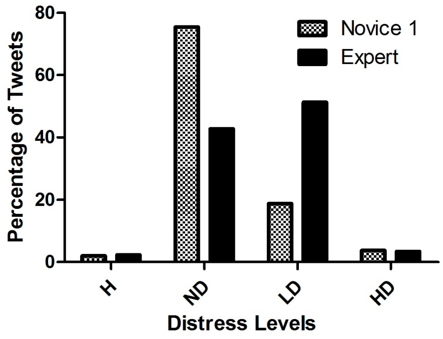

Figure 3: Distribution of distress level annotations from Novice 1 and Expert. Note the these two datasets are disjoint (N = 1000 tweets, respectively).

Figure 2 shows the distribution of annotation labels for the subset of tweets that Novices 1 and 2 both annotated, and Figure 3 compares the overall annotation distributions between Novice 1 and the Expert. Interestingly, the novices are relatively conservative, compared to the expert, in assigning distressed labels, whereas the expert exhibits a higher sensitivity toward low distress than either of the novices. This suggests that it is important in this domain not to rely too much on novice judgments, as novices are not trained to pick up on subtle cues—in contrast to the clinically trained eye.

Note that there are very few happy tweets, which confirms that our filtering was effective in removing tweets of the opposite polarity.

Filtering method	**Kappa**
LIWC sad	0.4
Thematic suicide risk factors	0.6
Both	0.5

Table 4: Cohen kappa interannotator agreement between Novice 1 and 2.

	H	ND	LD	HD
H	0	2	0	0
ND	1	85	2	1
LD	0	22	9	0
HD	0	1	0	2

Table 5: Confusion matrix between Novices 1 and 2 on annotations of the LIWC-sad-based filtered tweets.

	H	ND	LD	HD
H	4	6	0	0
ND	0	55	12	1
LD	0	12	22	5
HD	0	1	3	4

Table 6: Confusion matrix between Novices 1 and 2 on annotations of tweets filtered by Jashinsky et al. (2013)'s thematic suicide risk factors inclusion terms.

Table 4 shows the Cohen kappa score between Novices 1 and 2, when high and low distress vs. no distress and happy, are grouped in a single category and Tables 5—7 show the confusion matrices between Novices 1 and 2. In all cases the kappa score is moderate. However, it clearly improves when annotation is restricted to just those tweets filtered using the suicide-thematic inclusion terms of Jashinsky et al. (2013). This again seems to point to the usefulness of including clinical experts into the training process.

Due to their sensitive nature, we decided not to provide examples of high distress tweets. Here are two examples of tweets labeled as low distress by two annotators.

- insomnia night#56325897521365!! sheesh can't deal w/ this shit! i have class in the morning got dammit....

	H	ND	LD	HD
H	4	8	0	0
ND	1	140	14	2
LD	0	34	31	5
HD	0	2	3	6

Table 7: Confusion matrix between Novices 1 and 2 on annotations of all common tweets between the two annotators.

- @XXXX i'm still sad thoo. i feel neglected! and i miss XXXX

And here are two examples of tweets labeled as no distress by two annotators.

- i did mad push-ups tryna get that cut up look, then look at myself after a shower ... #plandidntwork; thats #whyiaintgotomiami
- my son is gonna have blues eyes and nappy hair! yes yes yes

The above examples are rather clear cut, however in many cases the tweets were more ambiguous, even when annotators had the preceding and succeeding three tweets from the user of the tweet to be annotated to rely on for context. While context and time offset information was useful for annotators, distress annotation is clearly a challenging task, as the confusion matrices in Tables 5–6 reveal. The lower agreement levels, and particularly the fuzzy border between 'no distress' and 'low distress' are completely in line with prior research, discussed above, on affective language phenomena.

Another filtering and annotation challenge involves tweets with mixed emotion, such as:

- as much as i hate my job some of the people i work with are amazing.

Beyond the targeted annotation categories of distress level, there were emerging themes of aggression, privilege and oppression, and daily struggles, among others. For instance, jobs were a popular source of distress:

- i friggin hate these bastards my job grimey ass bastards knew i wanted the day off and tell me some next shit
- hate my job wit a passion! hate every1 there.. they better do sumthin about it, or im out!

Personal bias may have impacted annotation decisions. For instance, numerous tweets contained irony and dark humor, which may result in annotators underestimating or overlooking actual distress. In addition, by pulling data from Twitter, any non-Twitter context behind the tweets is lost. For example, a few individuals retweeted in a sarcastic manner about what individuals should say to someone who is considering suicide:

- you wish!!! rt @XXXX: i think suicide is funny. especially once my mom does it
- rt @XXXX: what do i say to a person thats asking me for advice becuz they thinking bout committing suicide when i see there point? lmao

Without knowing the circumstances of the original message (beyond the provided context window) it is difficult to classify such tweets.

Finally, a number of tweets seemed to show compassion or empathy for others experiencing stress. This suggests to us the profound role that social support places in well-being and depression, that one's friends and associates can also provide clues into one's emotional state, and that social media can reveal such behavior.

- rt @XXXX: damn now what do i do? i feel empty as f$% damit!! breathe ocho, *tears* from liberty city to (cont) http://XXXX
- @XXXX that's just sad i feel for you

High Distress	Random
feel like, wanna cry, get hurt, miss 2, ima miss, win lose, tired everything, broke bitches, gun range, one person	good morning, last night, happy birthday, look like, bout 2, can't wait, video , know (cont), chris brown, jus got
commit suicide, miss you!, miss baby, feel empty, committing suicide, tired living, sleep forever, lost phone, left alone, :(miss	feel like, let know, make sure, bout go, time get, don't get, wats good, . ., don't want, jus saw
hate job, feel sad, tummy hurts, lost friend, feel helpless, leave alone, don't wanna, worst feeling, leave world, don't let	don't know, let's go, looks like, what's good, go sleep, even tho, hell yea, new single, r u?, don't wanna

Table 8: Topic analysis on bigrams of tweets labeled as high distress vs. randomly selected tweets from the larger, unlabeled dataset. The high distress tweets clearly convey strong negative affect.

Table 8 shows the results of a 3-category topic model on bigrams. The first column is taken just from tweets labeled high distress by any one of the three annotators (72 tweets total). The second column comes from a randomly-chosen sample of 2000 tweets from the 2.3 million tweet corpus. These results show that the lexical contents of the annotated tweets are recognizeably different from the random sample. By our judgement, the topical groupings in the rows of the high distress column are all clearly marked by strong negative affect, and additionally they could arguably be labeled—from top to bottom—as: "failure and defeat," "loss," and "loneliness." The rows of the second column are less clear cut, and appear to reflect a much broader scope of topics. One interesting aspect of the second, random column is that recording artist Chris Brown had released a new album during the collection period, which seems to explain why his name appeared.

Training	Testing	Precision	Recall	F-Measure
N1	N1	0.53	0.63	0.58
N1	E	0.58	0.27	0.37
E	E	0.59	0.71	0.64
E	N1	0.34	0.85	0.48
N1 + E	N1 + E	0.33	0.41	0.37

Table 9: Performance of SVM-based classification when the training and testing sets are alternately Novice 1 (N1) or the Expert (E). Because we focus on distress classification, we report precision, recall and F-measure for the distress class, which combines LD and HD into a single class with respect to binary (distress vs. non-distress) classification. In each case, a held-out set of 100 randomly selected tweets compose the test set and the remaining 900 tweets from that annotator compose the training set. The last row shows when the two training sets (respectively, test sets) are combined into a single set of 1800 (respectively, 200) tweets.

For classification, because we are most interested in being able to separate distressed from non-distressed tweets, we combine low distress and high distress into a single distress class, and no distress and happy into a non-distress class. Table 9 shows the performance of the SVM-based classifier when trained and tested on the Expert and Novice 1 training sets. Four themes emerge: (1) the SVM classifier is much more accurate (in terms of F-measure) when the testing and training data come from the same annotator (test and training data are disjoint), and the best performance comes from the expert-annotated data. (2) When testing and training data are from different annotators, the F-measure performance of the SVM is lower when the training set is from the novice rather than the expert. (3) When testing and training data are from different annotators, the SVM has lower recall and higher precision when the training set is from the novice rather than the expert. This is in part because the Expert was more sensitive to distress than Novice 1. It is premature to draw conclusions from this observation, but perhaps this shows that training with expert-labeled annotations is preferable to using novice-labeled data, espectially when our goal is to discover distressful tweets for the purpose of identifying at-risk individuals and err on the side of caution (high recall). (4) Integrating more but mixed data does not improve performance.

5 Discussion

As previously mentioned, many of the risk factors for suicidal behavior may be linked to other expressions of distress, such as aggression and interpersonal violence (Mann et al., 1999). The goal of this study is to determine the feasibility of classifying distress to enable further study of expressed suicidal behaviors. Consistent with the stress diathesis model for suicidal behavior, aggression was an emerging theme that arose from the data. Here are some examples:

- `@XXXX i don't feel sad 4 him. he gets pissed n says wat he wants then sends out fony apologies`
- `@XXXX cuz he's n a relationship with that horseface bitch &amp; he lied 2 me &amp; i feel so used &amp; worthless now`

Some individuals tweeted about feeling empty, hopeless, angry, frustrated, and alone. Behaviors indicating bullying and schadenfreude were also observed. While these are all risk factors for internalizing aggression (i.e., suicidal behavior), they are also associated with externalized aggression. In addition to overt expressions of anger and violence, many of the humorous, ironic tweets also had an aggressive undertone.

5.1 Limitations

As ground truth, we rely on tweets hand-annotated by expert and novice for classification. However, the mental state of another individual, observed from a few lines of text often written in an informal register is necessarily hard to discern and, even under less noisy conditions, extremely subjective; even the observers' personal understandings of such concepts as "distress" may differ drastically. This makes annotation quite a challenge, and does not reveal in an objective fashion a tweeter's true mental state. As we have mentioned earlier, self-reporting has its own limitations, yet it is often regarded as the gold standard for ground truth about emotional state. Part of the problem in assessing the effectiveness of self-reporting is the relative rareness by which suicide occurs, and by the inherent subjectivity of the act, which makes any data on suicide fuzzy. We hope to explore in future work the relationship between clinical observation in both on- and off-line settings and self-reporting, including the integration of natural language data of patients from clinical settings. We also hope to explore distress annotation from different perspectives and levels of context.

Higher levels of suicidal ideation have an inverse relationship with all types of help-seeking and a positive correlation with the decision to not seek support (Deane et al., 2001). Thus, we would expect suicidal individuals to generally be less active on social media than those who are not. Nevertheless, a number of studies have shown a positive correlation between online social network use and negative mood. Perhaps this means in part that individuals who are depressed are slower to disengage on- rather than off-line.

6 Conclusion

We studied the performance of different approaches to training systems to detect evidence of suicide risk behavior in microblog data. We showed that both the methods used to automatically collect training sets, as well as the expertise level of the annotator affect greatly the performance of automatic systems for detecting suicide risk factors. In general, our study and its results—from filtering via data annotation to classification—confirmed the critical importance of bringing clinical expertise into the computational modeling loop.

Acknowledgments

This work was supported by a grant of the Kodak Endowed Chair Fund from the Golisano College of Computing and Information Sciences at RIT and NSF award SES-1111016.

References

Cecilia Ovesdotter Alm. 2008. *Affect in Text and Speech*. Ph.D. thesis, University of Illinois at Urbana Champaign.

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of 49th Annual Meeting of the Assoc. for Computational Linguistics: Human Language Technologies, Portland, OR*, pages 107–112.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal Machince Learning Research*, 3:993–1022, March.

Johan Bollen, Bruno Gonçalves, Guangchen Ruan, and Huina Mao. 2011a. Happiness is assortative in online social networks. *Artificial Life*, 17(3):237–251.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011b. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Johan Bollen, Alberto Pepe, and Huina Mao. 2011c. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 450–453.

Gregory K Brown, Aaron T Beck, Robert A Steer, and Jessica R Grisham. 2000. Risk factors for suicide in psychiatric outpatients: A 20-year prospective study. *Journal of Consulting and Clinical Psychology*, 68(3):371.

Ronny Bruffaerts, Koen Demyttenaere, Irving Hwang, Wai-Tat Chiu, Nancy Sampson, Ronald C Kessler, Jordi Alonso, Guilherme Borges, Giovanni de Girolamo, Ron de Graaf, et al. 2011. Treatment of suicidal people around the world. *The British Journal of Psychiatry*, 199(1):64–70.

Rafael A. Calvo and Sidney D'Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37.

Qijin Cheng, Shu-Sen Chang, and Paul SF Yip. 2012. Opportunities and challenges of online data collection for suicide prevention. *The Lancet*, 379(9830):e53–e54.

Alex E Crosby, LaVonne Ortega, and Cindi Melanson. 2011. *Self-directed violence surveillance: Uniform definitions and recommended data elements*. Centers for Disease Control and Prevention, National Center for Injury Prevention and Control, Division of Violence Prevention.

Munmun De Choudhury and Scott Counts. 2013. Understanding affect in the workplace via social media. In *16th ACM Conference on Computer Supported Cooperative Work and Social Media (CSCW 2013)*, pages 303–316. ACM.

Munmun De Choudhury, Scott Counts, and Michael Gamon. 2012a. Not all moods are created equal! Exploring human emotional states in social media. In *6th International AAAI Conference on Weblogs and Social Media*.

Munmun De Choudhury, Michael Gamon, and Scott Counts. 2012b. Happy, nervous or surprised? Classification of human affective states in social media. In *6th International AAAI Conference on Weblogs and Social Media*.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Major life changes and behavioral markers in social media: case of childbirth. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1431–1442. ACM.

Munmun De Choudhury, Scott Counts, Eric J Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer Supported Cooperative Work & Social Computing*, pages 626–638. ACM.

Frank P Deane, Coralie J Wilson, and Joseph Ciarrochi. 2001. Suicidal ideation and help-negation: Not just hopelessness or prior help. *Journal of Clinical Psychology*, 57:901–914.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, 6(12):e26752.

Bradley N Gaynes, Suzanne L West, Carol A Ford, Paul Frame, Jonathan Klein, and Kathleen N Lohr. 2004. Screening for suicide risk in adults: A summary of the evidence for the US Preventive Services Task Force. *Annals of Internal Medicine*, 140(10):822–835.

S.A. Golder and M.W. Macy. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881.

Xuan Guo, Rui Li, Cecilia Ovesdotter Alm, Qi Yu, Jeff Pelz, Pengcheng Shi, and Anne Haake. 2014. Infusing perceptual expertise and domain knowledge into a human-centered image retrieval system: A prototype application. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 275–278. ACM.

Aniko Hannak, Eric Anderson, Lisa Feldman Barrett, Sune Lehmann, Alan Mislove, and Mirek Riedewald. 2012. Tweetin in the rain: Exploring societal-scale effects of weather on mood. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM12)*.

Louise Harriss and Keith Hawton. 2005. Suicidal intent in deliberate self-harm and the risk of suicide: The predictive power of the suicide intent scale. *Journal of Affective Disorders*, 86(2):225–233.

Melonie Heron and Betzaida Tejada-Vera. 2009. Deaths: Leading causes for 2005. *National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, 58(8):1–97.

Christopher M Homan, Naiji Lu, Xin Tu, Megan C Lytle, and Vincent Silenzio. 2014. Social structure and depression in TrevorSpace. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 615–625. ACM.

Lisa M Horowitz and Elizabeth D Ballard. 2009. Suicide screening in schools, primary care and emergency departments. *Current Opinion in Pediatrics*, 21(5):620–627.

Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. 2013. Tracking suicide risk factors through Twitter in the US. *Crisis*, pages 1–9.

T. Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning (ECML)*, pages 137–142, Berlin. Springer.

Ronald C Kessler, Guilherme Borges, and Ellen E Walters. 1999. Prevalence of and risk factors for lifetime suicide attempts in the national comorbidity survey. *Archives of General Psychiatry*, 56(7):617–626.

Suin Kim, J Bak, and Alice Oh. 2012. Do you feel what I feel? Social aspects of emotions in Twitter conversations. In *Proceedings of the AAAI International Conference on Weblogs and Social Media*.

Michael Lehrman, Cecilia Ovesdotter Alm, and Ruben Proano. 2012. Detecting distressed vs. non-distressed affect state in short forum texts. In *Proceedings of the Workshop on Language in Social Media (LSM 2012) at the Conference of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies, Montreal, Canada*, pages 9–18.

Rui Li, Jeff Pelz, Pengcheng Shi, and Anne Haake. 2012. Learning image-derived eye movement patterns to characterize perceptual expertise. In *CogSci*, pages 1900–1905.

J John Mann, Christine Waternaux, Gretchen L Haas, and Kevin M Malone. 1999. Toward a clinical model of suicidal behavior in psychiatric patients. *American Journal of Psychiatry*, 156(2):181–189.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Naoki Masuda, Issei Kurahashi, and Hiroko Onari. 2013. Suicide ideation of individuals in online social networks. *PloS one*, 8(4):e62262.

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

Saif M Mohammad. 2012. # Emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 246–255. Association for Computational Linguistics.

Matthew K Nock, Guilherme Borges, Evelyn J Bromet, Christine B Cha, Ronald C Kessler, and Sing Lee. 2008. Suicide and suicidal behavior. *Epidemiologic Reviews*, 30(1):133–154.

Matthew K Nock, Jennifer M Park, Christine T Finn, Tara L Deliberto, Halina J Dour, and Mahzarin R Banaji. 2010. Measuring the suicidal mind implicit cognition predicts suicidal behavior. *Psychological Science*, 21(4):511–517.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of Conference on Language Resources and Evaluation*.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001.

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

René Pfitzner, Antonios Garas, and Frank Schweitzer. 2012. Emotional divergence influences information spreading in twitter. *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM12)*, pages 2–5.

Rosalind W. Picard. 1997. *Affective Computing*. MIT Press, Cambridge, MA, USA.

Megan L Ryan, Ian M Shochet, and Helen M Stallman. 2010. Universal online interventions might engage psychologically distressed university students who are unlikely to seek formal help. *Advances in Mental Health*, 9(1):73–83.

Adam Sadilek, Henry A Kautz, and Vincent Silenzio. 2012. Predicting disease transmission from geo-tagged micro-blog data. In *Association for the Advancement of Articial Intelligence*.

Adam Sadilek, Christopher Homan, Walter S. Lasecki, Vincent Silenzio, and Henry Kautz. 2014. Modeling fine-grained dynamics of mood at scale. In *WSDM 2014 Workshop on Diffusion Networks and Cascade Analytics*.

David Shaffer, Michelle Scott, Holly Wilcox, Carey Maslow, Roger Hicks, Christopher P Lucas, Robin Garfinkel, and Steven Greenwald. 2004. The Columbia SuicideScreen: Validity and reliability of a screen for youth suicide and depression. *Journal of the American Academy of Child & Adolescent Psychiatry*, 43(1):71–79.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2011. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012. Harnessing Twitter 'Big Data' for automatic emotion identification. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 587–592. IEEE.

Kathryn Womack, Wilson McCoy, Cecilia Ovesdotter Alm, Cara Calvelli, Jeff B. Pelz, Pengcheng Shi, and Anne Haake. 2012. Disfluencies as extra-propositional indicators of cognitive processing. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 1–9. Association for Computational Linguistics.

Matt Wray, Cynthia Colen, and Bernice Pescosolido. 2011. The sociology of suicide. *Annual Review of Sociology*, 37:505–528.