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

Abstract

Suicide is a leading cause of death in the United States and one that has grown to nearly epidemic proportions in some communities. One of the major challenges to suicide prevention is that those who may be most at risk cannot be relied upon to report their condition to clinical practice. This paper takes an initial step toward the automated 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.

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 with suicidal ideation 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 data sources.

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 trends in help-seeking behavior suggests that social media might be a rich outlet for learning about support seeking. 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?

Mann et al. (1999) developed the stress diathesis model for suicidal behavior, using many of the aforementioned risk factors. This model suggests (1) that 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, were among the risk factors that contributed to suicidal ideation, and (2) that the presence of these factors could eventually lead to either externalizing (e.g., interpersonal 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 to analyze publicly available linguistic data from social media outlets such as Twitter. Data from social media can be used as a natural experiment to examine depression and suicidal ideation without being constrained by such sample biases as the willingness of individuals to take part in research and/or seek out formal sources of support. Moreover, this natural experiment method may provide information about individuals who are unlikely to engage in formal help-seeking behaviors and eventually could be used to identify effective methods of natural helping. Hence, 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. We use methods that take advantage of lexical analysis to retrieve microblog posts (tweets) from Twitter and compare the performance of human annotators—some being experts, and some not—to rate the level of *distress* of each tweet. According to Nock et al. (2010) distress is an important risk factor in suicide, and one that is observeable from microblog text, though admittedly observing suicide risk behavior is a subjective and noisy venture. *Clinical* expert annotation, rather than general-purpose tools for content and sentiment analysis such as LIWC (Linguistic Inquiry and Word Count), provides a basis for text-based statistical modeling. We also show that expertise-based keyword retrieval, departing from knowledge about contributing risk factors, results in better interannotater 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, such as stigma and shame, among other reasons (Crosby et al., 2011). Moreover, suicide is a fundamentally subjective, complex phenomenon with a low base rate. For these reasons, data on suicide is never particularly reliable and 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 suicide (Kessler et al., 1999). According to Kessler et al. (1999), the odds of attempting suicide increased exponentially when individuals endorsed three of more risk factors (e.g., having a mood or substance abused disorder).

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 mechanism, i.e., 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; Shaffer et al., 2004; Brown et al., 2000).

More broadly in clinical contexts, evidence suggests that when it comes to judgments that involve clinical phenomena, experts and novices behave differently. For example, in a medical image inspection task, Li et al. (2012) identified differences in perceptual expertise patterns between novices (students) and clinically trained physicians. Similarly, Womack et al. (2012) identified differences in linguistic behaviors between experienced, attending dermatologists vs. resident dermatologists-in-training based on diagnostic verbal narratives. 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 both in speech and in the text analysis domain, as well as in many other modalities (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.

Level of distress is a key element to consider when evaluating at-risk behaviors with respect to suicidal behavior or depression. Lehrman et al. (2012) conducted a first 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 evolves can allow us to address macro-level analysis, and social media texts provide the additional opportunity to model the phenomena of interest at scale.

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. (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-partem—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

In this section, we describe the methods we use to label and detect distress in Twitter data. Our process involves four main phases: (1) We filter 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 train support vector machines and topic models with the annotated data, except for a held-out subset of 200 tweets. (4) Finally, we assess the effectiveness of these methods on the held-out set.

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	2000
	Unique users	1467
	Unique tokens	1714167
	Unique bigrams	9246715
	Unique trigrams	13061142
Categories distribution	LIWC sad	1370
	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 of the and thematic categories distributions of the collected dataset. The data was 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

We first preprocessed each tweet in the corpus by (a) converting all text to lower case; (b) stripping out punctuation and special characters; and (c) building a dictionary of more than 5,400 terms that captured informal Twitter registers, such as abbreviations and netspeak, based on http://www.noslang.com/dictionary.

In order to test the effectiveness of various methods of capturing useful corpus data, we used two different methods to filter for tweets that are relatively likely to center on suicide risk factors. As the first method, we used the Linguistic Inquiry and Word Count (LIWC) to capture 1,370 tweets by sampling randomly from the all tweets with at least the $2,000$th-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. These terms yielded 630 tweets.

3.2 Novice and Expert Tweet Annotation

```
978: Date: XXXX
    -3: dat man on maury is overreacting
    -2: @XXXX cedes!!! [-0:21:25]
    -1: yesssss! da weatherman was wronq
>>> @XXXX awwww thanks trae-trae
     1: rt @XXXX: abt 2 hop in a kab to
     2: @XXXX yeaa [+0:03:59]
     3: @XXXX wassup? [+0:05:28]
Msg_id: XXXX  [Distress: ND, LIWC Sad: N
```

Figure 1: Example input for annotator. Each line is one tweet. The tweet being annotated is indicated by >>>. Annotators were given context in the form of a window of three immediately preceding tweets s well as three immediately following tweets, including their respective time-stamped offsets compared to the annotation target tweet. (Tweeter information has been blanked out.)

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. Each tweet in each set was rated on a four-point scale (H, ND, LD, HD) according to the level of distress evident (Table 2).

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

Table 2: Distress-related categories used to annotate the tweets.

For the annotation process itself, each tweet was provided with a context, i.e., three tweets before and after the tweet to be annotated, along with the timestamp of these tweets and thematic category to which the tweet belonged (Figure 1).

3.3 Modeling

We represent each tweet as a collection of all unigrams, bigrams, and trigrams in the message. For example, a simple tweet "I am so happy" is represented as the following *feature vector*: {I, am, so, happy, I am, am so, so happy, I am so, am so happy}. This method allows one to construct prior probabilities on pairs and triples of consecutive words and thus model the probability spaces of arbitrarily long utterances, in a way that is natural and often effective in representing linguistic data with at least a limited context (given data sparsity concerns for longer sequences) for the purpose of classification or topic modeling.

We perform topic modeling on our dataset to compare the topics. Topic modeling is often used to analyze text data by finding topics within a corpus of documents. A topic is characterized by lexical items that are likely to occur with the topic. These models are capable of connecting words with similar meanings and distinguish words with multiple meanings. We utilize Latent Dirichlet Algorithm (LDA) (Blei et al., 2003) to create these topics. In this method the documents (in our case tweets) are represented as random mixtures over latent topic where each topic is characterized by a distribution over words. Before performing the topic modeling, the stop words and words that occur only once in the dataset are removed. The LDA algorithm then establishes three topics using 100 iterations.

We use Support Vector Machines (SVM), 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*. They have proven to be an extremely effect tool in classifying text in numerous settings, for different types of problems and with varying form of text data, including Twitter.

4 Results

Figures 2 and 3 show the distribution of the annotation labels each annotator makes. Interestly, the novices are fairly consistently highly conservative with assigning distressed labels, whereas the expert exhibits a higher sensitivity toward low distress than either of the novices. This suggests

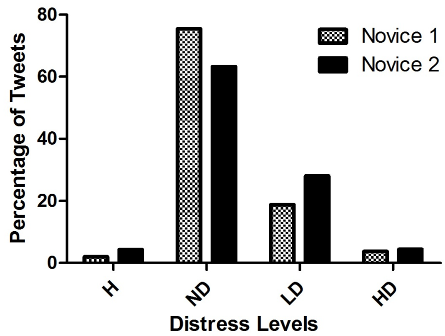

Figure 2: Distribution of distress level annotations on the common tweets (N=250, indentical set) between annotators Novice 1 and Novice 2.

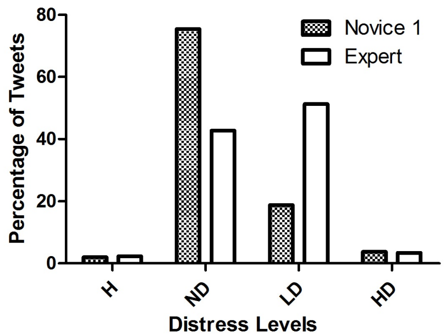

Figure 3: Distribution of distress level annotations from Novice 1 and Expert. Note the these two datasets are disjoint (N = 1000 tweets, respectively).

that it is is important in this domain to not 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 with the filtering, we do not generally pick up tweets that are of the opposing polarity than we intended, which is good.

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

Table 3: 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 4: 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 5: 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 3 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 4 thru 6 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 integrating clinically acknowledge insights.

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.

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

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

- `@XXXX that's just sad i feel for you`

And here are two examples of tweets labels 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 rather ambiguous, even when annotators have the previous and subsequent three tweets from the user of the label tweet to rely on for context. While context and time offset information was useful for annotators, distress annotation is clearly a challenging tasks, as the confusion matrices in Tables 5-6 reveal. The lower agreement levels, and particularly the fuzzy boarder between 'no distress' and 'low distress' are completely in line with prior research, discussed above, on affective language phenomena.

Beyond the targeted annotation categories of distress level, there were emerging themes of aggression, privilege and oppression, and daily struggles, among other. 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`

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

The last example also shows that tweets sometimes expressed very strong ambivalence.

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, critical information such as pictures and the context behind information that has been retweeted was not available to the annotators. For example, a few individuals retweeted in a humorous manner about what individuals should say to someone who 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 was difficult to classify such tweets.

Finally, a number of tweets seemed to show compassion or empathy for others experience stress. This suggest 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`

Table 7 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 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 column is that recording artist Chris Brown had released a new album during the collection period, which seems to explain why his name appeared as prevalent..

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 7: Topic Analysis on bigrams of High Distress and Random Tweets

Train Set	Test Set	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 8: Performance of SVN-based classification when the training and testing sets are alternately Novice 1 (N1) or the Expert (E). In each case, the test set is a held-out set of 100 randomly selected tweets and the remaining 900 tweets from that annotator were used as training data. The last row shows results when N1 and E data are combined into a train set of 1800 tweets and a test set of 200 tweets (with 50% of each set consisting of data annotated by Novice 1, and the Expert respectively). Because we are focused on Distress classification, Precision and Recall are reported for the Distress class, which combine LD and HD into a single D label in this binary classification.

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 another class. Table 8 shows the performance of the support vector when trained and tested on either on the Expert and Novice 1 training sets. Four themes emerge: (1) the SVN classifier is much more accurate when the testing and training data come from the same source (taken from 100 tweets in the annotation set that are randomly held out of the training set, so that the training and test sets are disjoint); (2) when testing and training data are from different sources the SVM suffers less of an accuracy drop when the training set is from the expert than from the novice; (3) when the train set is from Novice 1, the classifier suffers a loss in recall on the Distress class, and when the train set is from the Expert, there is a loss in precision instead. If our goal is to identify distress tweets for further scientific study of how such cases express suicide risk factors, the high precision classifier trained on Expert annotations is preferable. (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 classify whether or not tweets were related to distress in order to determine the feasibility of classifying distress to enable further study of expressed suicidal behaviors suicidal behaviors. However, due to the overlap between internal and external expressions of anger, it is difficult to classify suicidal behavior without more contextual information. 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 & he lied 2 me & i feel so used & worthless now`

A number of individuals tweeted about feeling empty, hopeless, angry, frustrated, and alone. Behaviors indicating bullying and Schadenfreude were also observed. While these are risk factors for internalizing aggression (i.e., suicidal behavior); these states are also associated with externalizing 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 an expert and a novice for classification. However, the mental state of another individual, observed from a few lines of text 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 objective fashion of 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.

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 and Future Work

This research was conducted within a multidisciplinary group with backgrounds in computer science, linguistics and psychiatry. This allowed us to understand the data from multiple perspectives, both clinical psychology and data mining. We analyzed the presence of suicidality or suicide risk behavior in microblog data collected from twitter. The dataset was filtered based on terms in thematic categories and sad scores from LIWC. Text-based features were extracted to represent the suicide risk factors. Using these features, the performance of different approaches of training a system to detect evidence of suicide risk behavior in microblog data was studied. We showed that both the methods used to automatically collect training sets, as well as the expertise level of the annotators, greatly affect the performance of automatic systems for detecting suicide risk behavior.

Our current work suggests that text-based features can lead to a system for detecting suicidal behavior. The biggest challenge was handling the noise in twitter data. For example, the prominent usage of sarcasm and schadenfreude in social context leads to false positives for distress symptoms, which at this stage can be detected only via manual examination. To avoid this, data can be collected from microblogs and support forums which provide assistance to individuals at risk. Another area for future work includes prediction at the user level that utilizes social data, via social network analysis, and geographic location.

References

Cecilia Ovesdotter Alm. 2008. Affect in text and speech.

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. *J. Mach. Learn. Res.*, 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.

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 *Sixth 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 *Sixth 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 *Proc. CSCW*.

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*. ACM.

Aniko Hannak, Eric Anderson, Lisa Feldman Barrett, Sune Lehmann, Alan Mislove, and Mirek Riedewald. 2012. Tweetinin the rain: Exploring societal-scale effects of weather on mood. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM12)Dublin2012*.

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.

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 Am. Chapter of the Assoc. for Comp. 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.

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 LREC*, volume 2010.

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. *AAAI ICWSM*, 2012:2–5.

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

Adam Sadilek, Henry A Kautz, and Vincent Silenzio. 2012. Predicting disease transmission from geo-tagged micro-blog data. In *AAAI*.

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.