MS Thesis Proposal

Modeling of Distress: Towards Understanding Suicidaility

Ravdeep Johar

Committee Chair: Dr. Cecilia Ovesdotter Alm

Reader: Dr. Christopher M. Homan

Observer: Dr. Megan Lytle

Department of Computer Science
B. Thomas Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, New York

May 8, 2014

Abstract

Suicide is a leading cause of death worldwide. 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 project will focus on modeling suicidaility using social media venues like Twitter, Reddit and TrevorProject. The modeling will provide an early flag system to identify users, who are at-risk using their social data.(wip)

Contents

List of Figures

List of Tables

> When you get into a tight place and everything goes against you, till it seems as though you could not hang on a minute longer, never give up then, for that is just the place and time that the tide will turn.
>
> Harriet Beecher Stowe

1 Introduction

Approximately a million people worldwide die by suicide every year (World Health Organisation, 2002), it is one of the leading causes of death among individuals aged 13-25 (Kung et al., 2008). Currently, there are no evidence based risk assessment tools for predicting or detecting suicidal behaviors and patients who seek help are assessed by the judgment of a psychological clinician, where the patient has to undergo a number of sessions to convey their distress. This leads to one of the major concerns while dealing with suicidality, that individuals, who are at risk, may not realize their level of risk or reach out for counseling, and even if they do, will they be able to provide accurate information or express their distress?

Self-reporting cannot always be relied upon for detecting suicide ideation in patients. This thesis seeks to provide an additional way to assess the level of risk by utilizing a patient's social media information. Social media venues provide an informal setting, where people can interact and share their opinions without having any qualms about social stigma. In this informal setting, individuals may reveal their mental distress and even use such venues to seek support and guidance instead of clinical treatment (Bruffaerts et al., 2011; Crosby et al., 2011; Ryan et al., 2010).

Social media websites such as Twitter, Reddit and other micro-blogging services are attractive for research involving extracting and mining of text data. These websites produce a large volume of real-time data: Twitter has on average 140 million tweets posted everyday and Reddit generates about 40 million posts every year. The type of data varies with the website, from short updates (via tweets) to a long narration of personal experience (via Reddit). The quality of data is usually inconsistent, with informal register, non-standard text such as ad hoc abbreviations, phonetic substitutions, ungrammatical structures and usage of emoticons for expressiveness. These charatericts of social media data creates an additional challenge to semantically process the text data.

Prior research relevant to understanding suicidality has primarily been conducted in psychology to associate risk factors with suicidal behavior and is still under-studied. The stress-diathesis model by Mann et al. (1999) provides evidence that risk factors such as depression, substance abuse or family violence can contribute to suicidal ideation. My research will focus on a preventive perspective of suicidaility rather than predicting if an individual will commit suicide. Specifically, this study aims to investigate if it is possible to automatically predict distress levels (*no distress, low distress* and *high distress*) based on users' social media information. Identifying users with such mental distress is an important step towards understanding and studying risk factors, which in the long term may help address suicide prevention by providing an early flagging system to identify individuals at-risk.

2 Background

Natural language processing(NLP) combines fields such as computer science, artificial intelligence and linguistics. It's concerned with understanding the human interactions using natural language. Although there huge advances in the field of Artificial Intelligence (AI) over the last few decades, the ultimate goal of making machines understand human language is still far off.

3 Related Work

Suicide is a subjective and complex phenomenon with a low base rate. Data on suicide is usually collected from healthcare organizations, large-scale studies, or self reporting Crosby et al. (2011); Horowitz and Ballard (2009). The problem these sources is that they are limited by ociocultural barriers, such as stigma and shame Crosby et al. (2011). Many researchers just focus on research related to associating risk factors and suicidaility without considering theoretical models Nock et al. (2008). The know risk factors are 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). Any patient who exhibits more than three of these factors is considered to be at-risk.

The related work for this thesis can be categorized as research on: (nlp) suicide notes, social media and sentiment analysis.The earliest work on analyzing suicide notes was conducted by Stone and Hunt (1963), using a dataset of 750 real suicide notes collected over a period of 10 years, they developed a system to detect real from simulated suicide notes(written by people from labor unions, fraternal groups, and the general community). Three patterns were observed:

- References to concrete things, persons and places were higher in real notes.
- Usage of the word "love" was higher in real notes.
- References to thought process and decision was higher in simulated notes.

Recent work on suicide notes has involved classifying simulated and completer suicide notes (Pestian et al., 2010, 2008), the authors categorize the suicide notes into ideators(who think about suicide), attempters(who attempt suicde) and completers(who have completed suicide). They utilize advanced techniques such as heterogeneous selection, parts-of-speech, Flesch and Kincaid readability score (Kincaid et al., 1975) and various machine learning algorithms to show that these algorithms can perform on par with mental health professionals in this task of distinguishing simulated vs completer suicide notes.

Research on social media has recently gained a lot of attention due to the large amount of data it generates, the most interesting research capability of this data is forecasting: predicting future events. Researchers have successfully been able to predict election results (Gayo-Avello, 2013), stock market movements (Bollen et al., 2011), flu outbreaks (Lampos and Cristianini, 2010), personality traits (Golbeck et al., 2011), mobility patterns of individuals (Song et al., 2010) and box-office revenues of movies (Asur and Huberman, 2010). These predictive models although successful, are not perfect because text data can't entirely account for real-world outcomes.

Sentiment Analysis is task of classifying the polarity of text into positive, neutral or negative. Early work consisted of classifying movie reviews (Pang et al., 2002) and product reviews (Turney, 2002) as positive or negative. Pang et al. (2002) and Snyder and Barzilay (2007) were then able to predict the ratings (based on 4, 5 or 10 point scale) of movies and restaurant based on their reviews. To classify these sentiments: Naïve Bayes, Maximum Entropy and SVM are the popular machine learning algorithms. Unlike traditional classifiers, where the neutral class is ignored, classifiers predicting sentiment are known to benefit with the addition of the neutral class. This task has of sentiment analysis can also classify human mood such as angry, happy or sad. Read (2005) and Go et al. (2009) consider the usage of emoticons (such as :) and :-/) to predict sentiment of the text from Usenet newsgroups and Twitter using a dataset of positive and negative emoticons.

Software Libraries such as Linguistic inquiry and word count(LIWC)(Pennebaker et al., 2001) and SentiWordNet (Baccianella et al., 2010) can classify a given text data can various psycholinguistic categories such as work, anger, sad, happy, money, home, etc. LIWC provides 80 different categories, which can provide insight on an individuals social and psychological state-of-mind.

Relatively less work has focused on suicide or other psychological conditions. Choudhury et al. (2013) and Prieto et al. (2014) successfully use Twitter data to detect depression and other mental health conditions, and argue that Twitter text data is viable to capture individuals psychological state. Choudhury et al. (2013) employed crowdsourcing to obtain a set of Twitter user who are clinically depressed based on standard psychometric tests. The author then retrieved their social information for last year and extracted behavioral text features and network features to predict the on-set of depression.

Prieto et al. (2014) attempts to predict health conditions such as flu, depression, pregnancy and eating disorders by extracting relevant tweets to each category using a set of regular expressions and then classifying these conditions using an SVM with mean f-measure of 0.85.

Jashinsky et al. (2013) used twitter data to identify suicide risk factors. The author used a filtering method to identify at-risk tweets using keywords and phrases created from suicide risk factors. These at-risk tweets were then grouped by state and departures from expectation were calculated. This research provided evidence that suicide risk factors have a presence of twitter and research on suicide is viable using such data.

4 NLP Techniques

4.1 Text Normalization

While dealing with text based data, it is first preprocessed and/or normalized before any semantic processing can take place.

- Tokenization: Splits text into individual words(tokens).

- Token Normalization: Process of normalizing individual tokens by ignoring caps, removing punctuations or using a stemmer/lemmentizer.

- Replace Informal Register: Using a dictionary of words collected from `noslang.com`, informal register is replaced with its meaning. For example, 'BBM' will be replaced with 'Black berry message'

4.2 Bag of Words

This is a common approach in NLP, where a sentence, for example: "My name is Jon" is represented as a set of words { 'My', 'name', 'is', 'Jon' }.

4.2.1 *n*-gram Modeling

Given a text, a *n*-gram is a contiguous sequence of *n* items. It is called as uni-gram, bi-gram and tri-gram where $n = 1, 2, 3$ respectively. For above example, the bi-gram model will be {'My name', 'name is', 'is Jon'}. Similary other *n*-gram models can be created using the bag-of-words approach.

4.3 Term Frequency-Inverse Document Frequency

tf-idf is a way to represent text data in a collection of documents using statistical weights. Each word in a document has a weight associated, which signifies the importance of the word in a document in the collection. The term frequency (tf) of a word (t) in a document (d) represented by equation 1, where $f(t, d)$ is the frequency of the term t in document d. Since longer documents can cause a bias in the weighting, tf is divided by the maximum frequency of a word in the document d.

$$\mathrm{tf}(t, d) = 0.5 + \frac{0.5 \times \mathrm{f}(t, d)}{\max\{\mathrm{f}(w, d) : w \in d\}} \tag{1}$$

Inverse Document Frequency(idf) is measure of how important a word is in the entire collection of documents and is calculated using equation 2, where, N is the number of documents in the collection and $\frac{N}{|\{d \in D : t \in d\}|}$ is the number of documents in which the word(t) occurs.

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|} \tag{2}$$

tf-idf is the multiplication of the above two terms(equation 3). This type of weighting is:

- Highest: when a word occurs many times within a small number of documents, which provides a discriminating power to these documents.

- Lower: when the word occurs fewer times in a document, or occurs in many documents, which imples the word has no discrimination power or relevance.

- Lowest: when the word occurs in all of the documents.

$$tf - idf = \mathrm{tf}(t, d) * \mathrm{idf}(t, D) \tag{3}$$

4.4 Latent Dirichlet Allocation

Latent Dirichlet Allocation(Blei et al., 2003) is an algorithm to perform Topic Modeling. It is an generative probabilistic model of a corpus of documents, where, the documents are represented over a random mixture of latent topics and each topic is characterized over a distribution of words.

```
1. Choose N Poisson
2. Choose Dir.
3. For each of theN wordswn:
        (a) Choose a topic zn Multinomial().
        (b) Choose a wordwn from p(wn jzn;),
            a multinomial probability conditioned on the topic
            zn.
```

Figure 1: Example input for annotator.

$$f\left(x_1, \cdots, x_{K-1}; \alpha_1, \cdots, \alpha_K\right) = \frac{1}{\mathrm{B}(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \tag{4}$$

4.5 Support Vector Machine

Support Vector Machine(Cortes and Vapnik, 1995) is a machine learning algorithm which can learn and recognize patterns. (Aizerman et al., 1964)

5 Dataset

I plan to make use of three datasets in the study. Two datasets have already been collected from Twitter. The first consists of New York City tweets collected over a month, and the second consists of nationwide tweets. The third dataset will be collected from Reddit using sub-reddits such as SuicideWatch, Depression, Happy, Anger, Selfharm, the goal is to collect about 2000 posts from each sub-reddit. Another possible dataset consists of on-line chat support data from the Trevor Project, an organization that seeks to prevent suicide among LGBT teens. Access to the Trevor Chat dataset is in progress with support of a collaborator.

5.1 Human Subject Research

RIT's IRB has confirmed that research involving Twitter or Reddit data, which is public, is not human subjects research, per federal definition. IRB approval for the Trevor Chat will be sought as soon as Trevor Project grants data access. My committee is experienced with human subjects research, and I have completed training in human subjects research.

6 Current Work

Currently, research already started on the first Twitter dataset, which consists of 2.5 million tweets. So far research has involved three steps: Filtering out distressed tweets, annotation and modeling.

6.1 Annotation

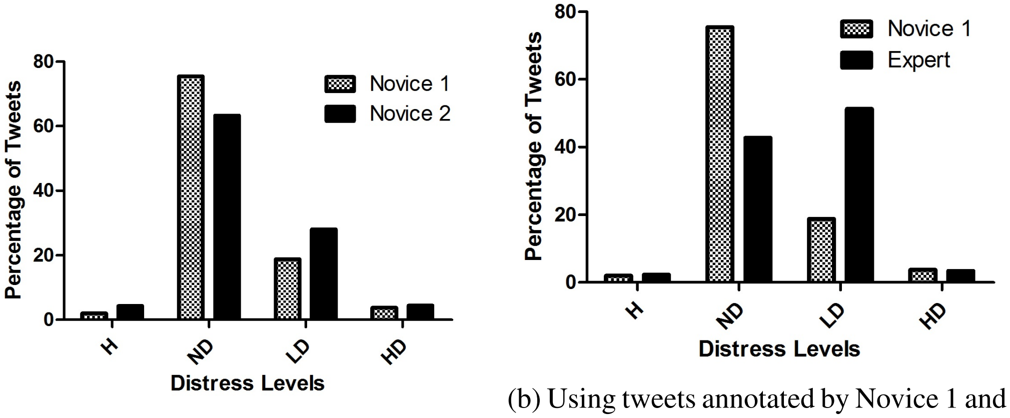
The 2000 possible at-risk tweets were annotated by two novices and an expert, where the expert is a clinical psychologist with relevant experience in research with suicidality.The 2000 tweets were divided into two sets(1000 each). Novice one annotated the entire first set, novice two annotated 250 tweets from the first set and the expert annotated all the tweets in the second set. Each tweet was provided with a context of three tweets before and after the at-risk tweet, time stamps for all seven tweets and the thematic category to which the current tweet belongs to. Figure 2 provides an example of what each each tweet(as seen by the annotators). The annotators were asked to categorize the distress level on a four-point scale(as in table 1) and check if the tweet belonged to its thematic category or not. Figure 3 shows the distributions of the annotations. An interesting observation is that, both novices were conservative/ reluctant to assign distressed labels compared to the expert. This shows that novices were unable pick up on subtly cues in the text that the expert could to identify tweets as distressed. Also the low number of happy tweets suggests that the filtering of tweets works well.

```
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 2: Example input for annotator. The text within ">>>" and "<<<" is the current tweet

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

Table 1: Distress-related categories used to annotate the tweets.



(a) Using tweets annotated by Novices 1 and 2 (N=250, identical set).

(b) Using tweets annotated by Novice 1 and Expert. Note the these two datasets are disjoint (N = 1000 tweets, respectively)

Figure 3: Distribution of distress level annotations

6.2 Modeling and Results

Topic Modeling:

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 2: 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.

SVM:

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 3: 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.

7 Solution Design and Implementation

7.1 Visulizations

7.2 Social Network

7.3 Modeling for Reddit Data

using topic modelling and other features

7.4 Outcome

8 Roadmap

8.1 Deliverables

There will be two principal deliverables for this thesis. First, a thesis report and presentation describing the hypothesis, development, functionality, visualizations, results, evaluation and conclusion. Second, the code written for developing this system will be made available and reported for evaluation.

8.2 Timeline

Date	Task
15, 2014	Complete thesis proposal
15, 2014	Complete research work and implementation of algorithms
15, 2014	Finish implementing the software
15, 2014	Finish writing report
15, 2014	Defense

References

Aizerman, M. A., Braverman, E. A., and Rozonoer, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning. In *Automation and Remote Control,*, number 25 in Automation and Remote Control,, pages 821–837.

Asur, S. and Huberman, B. A. (2010). Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *in Proc. of LREC*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal Machince Learning Research*, 3:993–1022.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Bruffaerts, R., Demyttenaere, K., Hwang, I., Chiu, W.-T., Sampson, N., Kessler, R. C., Alonso, J., Borges, G., de Girolamo, G., de Graaf, R., et al. (2011). Treatment of suicidal people around the world. *The British Journal of Psychiatry*, 199(1):64–70.

Choudhury, M. D., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting depression via social media. In Kiciman, E., Ellison, N. B., Hogan, B., Resnick, P., and Soboroff, I., editors, *ICWSM*. The AAAI Press.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Crosby, A. E., Ortega, L., and Melanson, C. (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.

Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from twitter data. *Social Science Computer Review*, 31(6):649–679.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *Processing*, pages 1–6.

Golbeck, J., Robles, C., Edmondson, M., and Turner, K. (2011). Predicting personality from twitter. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 149–156. IEEE.

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

Jashinsky, J., Burton, S. H., Hanson, C. L., West, J., Giraud-Carrier, C., Barnes, M. D., and Argyle, T. (2013). Tracking suicide risk factors through Twitter in the US. *Crisis*, pages 1–9.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document.

Kung, H.-C., Hoyert, D. L., Xu, J., and Murphy, S. L. (2008). Deaths: final data for 2005. *Natl Vital Stat Rep*, 56(10):1–120.

Lampos, V. and Cristianini, N. (2010). Tracking the flu pandemic by monitoring the social web. In *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*, pages 411–416.

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

Nock, M. K., Borges, G., Bromet, E. J., Cha, C. B., Kessler, R. C., and Lee, S. (2008). Suicide and suicidal behavior. *Epidemiologic Reviews*, 30(1):133–154.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001.

Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A., and Leenaars, A. (2010). Suicide note classification using natural language processing: A content analysis. *Biomedical Informatics Insights*, 3:19–28.

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

Prieto, V. M., Matos, S., lvarez, M., Cacheda, F., and Oliveira, J. L. (2014). Twitter: A good place to detect health conditions. *PLoS ONE*, 9(1):e86191.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, pages 43–48. Association for Computational Linguistics.

Ryan, M. L., Shochet, I. M., and Stallman, H. M. (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.

Snyder, B. and Barzilay, R. (2007). Multiple aspect ranking using the good grief algorithm. In *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL*, pages 300–307.

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science*, 327(5968):1018–1021.

Stone, P. J. and Hunt, E. B. (1963). A computer approach to content analysis: Studies using the general inquirer system. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*, AFIPS '63 (Spring), pages 241–256, New York, NY, USA. ACM.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

World Health Organisation (2002). *World Report on Violence and Health: Summary.* WHO, Geneva.