Clustering semantic spaces of suicide notes and newsgroups articles

P. Matykiewicz[1,2]**, W. Duch**[2]**, J. Pestian**[1]
[1]Cincinnati Children's Hospital Medical Center, University of Cincinnati,
[2]Nicolaus Copernicus University, Toruń, Poland.

Abstract

Historically, suicide risk assessment has relied on question-and-answer type tools. These tools, built on psychometric advances, are widely used because of availability. Yet there is no known tool based on biologic and cognitive evidence. This absence often cause a vexing clinical problem for clinicians who question the value of the result as time passes. The purpose of this paper is to describe one experiment in a series of experiments to develop a tool that combines Biological Markers (B_m) with Thought Markers (T_m), and use machine learning to compute a real-time index for assessing the likelihood repeated suicide attempt in the next six-months. For this study we focus using unsupervised machine learning to distinguish between actual suicide notes and newsgroups. This is important because it gives us insight into how well these methods discriminate between real notes and general conversation.

1 Introduction

It is estimated that each year 800,000 die by suicide worldwide (World Health Organization, 2001). In the United States, suicide ranks second as the leading cause of death among 25-34 year-olds and the third leading cause of death among 15-25 year-olds (Kung et al., 2008). The challenge for those who care for suicide attempters, such as an Emergency Medicine clinicians, is to assess the likelihood of another attempt, a more lethal one. We believe to fully asses this risk a tool must be developed that measures both the biological and cognitive state of the patient. Such a tool will include Biological Markers (B_m): measured by the concentration of certain biochemical markers, Thought Markers (T_m): measured by artifacts of thought that have been reduced to writing or transcribe speech, and Clinical Markers (C_m): measured by traditional clinical risk factors. In this study we focus on the T_m because of BioNLP's important role. Here, we employ machine-learning analysis to examine suicide notes and how these notes compare to newsgroups. This is one experiment in a series of experiments that are intended to provide insight into how best to apply linguistic tools when responding to suicidal patients.

To gain insight into the suicidal mind, researchers have suggested empirically analyzing national mortality statistics, psychological autopsies, nonfatal suicide attempts and documents such as suicide notes (Shneidman and Farberow, 1957; Maris, 1981). Most suicide notes analysis has focused on classification and theoretical-conceptual analysis. Content analysis has been limited to extracting explicit information from a suicide note, e.g., length of the message, words, and parts of speech (Ogilvie et al., 1969). Classification analysis uses data such as age, sex, marital status, educational level, employment status and mental disorder (Ho et al., 1998; Girdhar et al., 2004; Chavez et al., 2006; Demirel et al., 2007). Only a very few studies have utilized theoretical-conceptual analysis , despite the assertion in the first formal study of suicide notes (Shneidman and Farberow, 1957) that such an analysis has much promise. So, the inconclusive nature of the methods of analysis has limited their application to patient care.

Our own research has taken a different approach. In particular we first wanted to determine if modern machine learning methods could be applied to free-text from those who committed suicide. Our first experiment focused on the the ability of machine learning to distinguish between real suicide notes and elicited suicide notes as well as mental health professionals. This is an important question since all current care is based on a mental health profession's interpretation. Our findings showed that mental health professionals accurately selected genuine suicide notes 50% of the time and the supervised machine learning methods were accurate 78% (Pestian et al., 2008). In this study we shift from supervised to unsupervised machine learning methods. Even though these methods have rich history we know of no research that has applied them to suicide notes. Our rationale for this study, then, is that since our ultimate goal is to create a Suicide Risk Index that incorporates biological and thought markers it is important to determine if unsupervised methods can distinguish between suicidal and non-suicidal writings. To conduct this research we developed a corpus of over 800 suicide notes from individuals who had committed suicide, as opposed to those who attempted or ideated about suicide. This is an important contribution and, as far as we know, it is the largest ever developed. It spans 70 years of notes, and now includes multiple languages. Details of this corpus are described below. We also created a corpus of data from various newsgroups that acted as non-suicidal writings. These corpora were used to conduct the analysis. The sections below describe the cluster analysis process and results.

2 Data

Suicide Notes Corpus
Data for the suicide note database were collected from around the United States. They were either in a hand written or typed written form. Once the note was acquired it was scanned into the database. Optical character recognition was attempted on the typed written notes, but not accurate, so the notes were read from the scanned version and type into the database exactly as seen. A second person reviewed what was typed. There were limitation in collecting deceased demographics. The table 1 provides vari-

ous descriptive statistics.

Newsgroup Corpus
Newsgroup data was selected because it was convenient and as close to normal discourse as we could find. We understood that and ideal comparison group would be composed of Internet blogs or e-mails that were written by suicide ideators. True, a Google query of "suicide blog" yields millions of response, a review of many of these responses shows that the data are of little use for this analysis. In our opinion, the next suitable corpora was found in a 20 newsgroup collection from the University of California in Irvine (UCI) machine learning repository[1]. Most of the newsgroups have no relevance to suicide notes. Since our hypothesis is that unsupervised learning methods can tell the difference between suicidal and non-suicidal writing we selected discussions that we believed may have some similarity to suicide writings. This selection was based on reviewing the newsgroups with experts. We had conjectured that if an unsupervised method could distinguish between similar clusters those methods could distinguish between dissimilar clusters. The newsgroups ultimately selected were *talk.politics.guns*, *talk.politics.mideast*, *talk.politics.misc*, *talk.religion.misc*. Each newsgroup contains 1000 articles (newsgroup postings). Headers and quotes from other postings were removed.

3 Methods

Basic statistics are calculated using variables extracted by Linguistic Inquiry and Word Count version 2007 software (LIWC2007) (Chung and Pennebaker, 2007). J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth created an annotated dictionary. Each word in the dictionary is assigned to at least one of the following high level category: linguistic process, psychological process, personal concern, or spoken category. These categories provide an efficient and effective method for studying the various emotional, cognitive, and structural components present in individuals' verbal and written speech samples (Chung and Pennebaker, 2007; Pennebaker et al., 2001). Here it is used to analyze differences between suicide notes and news-

---

[1]http://archive.ics.uci.edu/ml/

group articles.

Feature space was prepared using open source algorithms available in *Perl* language[2]. First, Brian Duggan spell checking software that uses aspell library was used (*Text::SpellChecker* module[3]). Then, tokenizer created by Aaron Coburn was used (*Lingua::EN::Tagger* module[2]) to extract words was applied. After that, words were filtered with 319 element stop word list [4]. Next, the Richardson/Franz English stemmer was included in the pre-processing stage (*Lingua::Stem* module[2]). Features that appeared in less than 10 documents or in more than 500 documents were removed. Documents that had less than 10 features or more than 500 were removed. Finally, columns and rows were normalized to have unitary lengths. These last steps of pre-processing are used to reduce outliers.

Calculations are done using open source software called R[5]. Clustering is done with the following algorithms: expectation maximization (EM) (Witten and Frank, 2000), simple k-means with euclidean distance (SKM) (Witten and Frank, 2000), and sequential information bottleneck algorithm (sIB) (Slonim et al., 2002). The last approach has been shown to work well work well when clustering documents. Specificity, sensitivity and F1 measure are used as performance measures (Rijsbergen, 1979). Multidimensional scaling with euclidean distance measures is used for visualization purposes (Cox and Cox, 1994).

To extract features that represent each cluster, Pearson correlation coefficient is used. The correlation coefficient r is calculated between each feature and each cluster separately $r(w_i, c_j)$ where w_i is ith word and c_j is jth cluster. N best features with the highest values for each cluster are selected as most representative.

4 Results

Descriptive statistics for the data sets are listed in table 1. It shows syntactic differences between language use in suicide notes and newsgroups when *Lingua::EN::Tagger* is used.

Table 1: Descriptive statistics of suicide note corpus and newsgroups.

	suicide corpus	**newsgroups**
Sample Size	866	4000 (1000 per group)
Collection Years	1945-2009	1992-1993
Avg tokens per record (SD)	105 (154)	243 (582)
Range of tokens per record	1-1837	0-11024
Average (SD) nouns	25.21 (34.81)	77.19 (181.63)
Average (SD) pronouns	16.58 (26.69)	18.05 (63.18)
Average (SD) verbs	21.07 (32.82)	41.31 (109.23)
Average (SD) adjectives	6.43 (9.81)	16.92 (36.45)

Table 2 summarizes information about the linguistic and psychological processes of the data. The idea of "process" is derived from the Linguistic Inquiry and Word Count (LIWC2007) software (Chung and Pennebaker, 2007). This software conducts traditional natural language processing by placing various word into categories. For example, *sixltrs* includes words that are at least six letters in length. A full description of this software, dictionaries, reliability and validity tests can be found on LIWC's website. [6]. Table 2 shows that suicide notes are, in many ways, different than normal text. For our study this provides inspiration for continued research.

Table 2: Mean and standard deviation in linguistic and psychological processes. Selected categories with smallest p-values (<0.0001) are shown.

	suicide	**guns**	**mideast**	**politics**	**religion**
artcl	3.31 (2.79)	7.80 (3.52)	7.37 (3.34)	7.21 (3.40)	7.07 (3.51)
sixltrs	14.20 (7.34)	21.22 (6.32)	23.24 (7.03)	22.41 (7.13)	21.37 (7.87)
prnoun	16.75 (6.82)	11.96 (5.15)	10.64 (4.92)	11.77 (5.18)	13.21 (5.76)
prepos	10.61 (4.35)	12.13 (3.97)	12.89 (3.89)	12.21 (3.97)	11.75 (4.07)
verb	14.69 (5.99)	12.75 (4.72)	11.54 (4.74)	12.72 (4.63)	13.54 (4.97)
biolog	2.70 (3.04)	0.93 (1.27)	0.85 (1.50)	1.59 (2.08)	1.10 (1.75)
affctiv	7.71 (5.39)	4.83 (2.87)	4.77 (3.45)	4.90 (3.18)	5.10 (3.93)
cognitv	12.68 (5.76)	16.14 (5.93)	14.72 (5.62)	16.00 (5.49)	17.14 (6.17)
social	10.45 (5.86)	8.10 (4.20)	8.43 (4.71)	8.76 (4.37)	9.06 (5.17)

The four newsgroup data sets are combined as follows: *talk.politics.guns* + suicide notes = guns, *talk.politics.mideast* + suicide notes = mideast, *talk.politics.misc* + suicide notes = politics,

talk.religion.misc + suicide notes = religion. Each data set contained 1866 documents before document and feature selection is applied. Table 3 has final number of features while table 4 has final number of documents. In general sIB clustering algorithm performed best for all data sets with respect to F1 measure (mean = 0.976, sd = 0.008). The average score also did not change when the number of clusters varied from two to six (mean = 0.973, sd = 0.012). Performance of k-means and expectation maximization algorithm was much worse. If number of clusters was varied between two and six for different data sets the algorithms achieved F1 measure 0.146 lower than sIB (SKM mean = 0.831, sd = 0.279, EM mean = 0.824, sd = 0.219). Table 3 summarizes performance of best algorithms for each data set if two clusters are chosen.

Table 3: Best clustering algorithms for each newsgroup when clustered with suicide notes in case of two clusters (alg = clustering algorithm, sens = sensitivity, spec = specificity, F1 = F1 measure, #f = number of features, sIB = sequential information bottleneck, SKM = simple k-means).

dataset	alg	sens	spec	F1	#f
guns	sIB	.9689	.9834	.9721	1658
mideast	sIB	.9837	.9942	.9877	2023
politics	SKM	.9705	.9889	.9769	1694
religion	sIB	.9787	.9700	.9692	1553

If the desired number of clusters is increased to four then two major sub-groups are discovered in suicide notes: emotional (represented by words like: *love*, *forgive*, *hope*, and *want*) and non-emotional (represented by words like: *check*, *bank*, and *notify*). Example of the first type of note might be (suicide note was annonymized and misspellings left unchanged):

Jane I am bitterly sorry for what I have done to you. Please try to forgive me. I can't live without you and you don't want me. I can't blame you though. But I love you very much. I didn't act like it but I did and still do. Please try to be happy, Jane. That is all I ask. I try hope for the best for you and I guess that is all there is for me to say. Good by. John Johnson. Please mail this to Mom. Mrs. Jane Johnson. Cincinnati, OH.

Example of a non-emotional suicide note might be:

There is no use living in pains. That arthritis and hardening of the arteries are too much for me. There are two hundred and five dollars in the bank, and here are fifty- five dollars and eight cents. I hope that will be enough for my funeral. You have to notify the Old Age Assistance Board. Phone - 99999.

Table 4 shows best five ranked features for each cluster for each data set according to correlation coefficient CC. Features are in the order of rank so that feature with the highest CC is first. Even though that we use different newsgroups as control groups same sub-groups of suicide notes are discovered. sIB is the most stable and best performing algorithm in this experiment so it was used to discover those clusters. Stemmed word that appear in best five ranked features in at least three data sets are marked bold.

Figures 1, 2, 3, and 4 show high-dimensional document/stemmed word feature space projected on a two dimensional plane using multidimensional scaling (MDS) initialized by principal component analysis. Each figure has different rotation but the shapes are similar. In addition MDS shows very little mixing of suicide notes and newsgroups which is also explained by results in the table 3.

Figure 1: MDS showing suicide notes and *talk.politics.guns* articles (**s** character in the figure means suicide note while **a** character depicts newsgroup article, colors are used as cluster numbers).
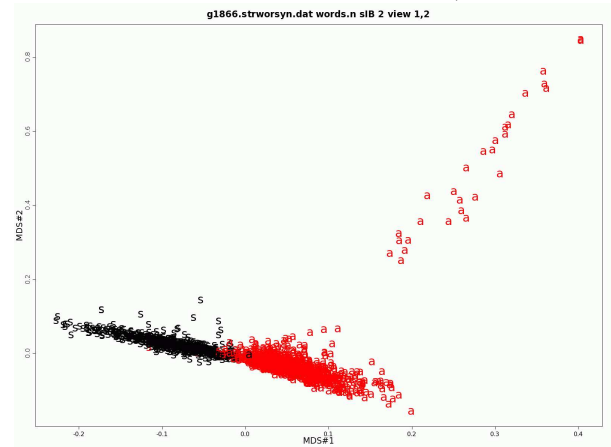
Table 4: Best five features when four clusters are created by the sIB algorithm (#c = cluster number, #a = number of newsgroup articles in a cluster, #s = number of suicide notes in a cluster). Stemmed word that appear in best five ranked features in at least three data sets are marked bold.

dataset	#c	stemmed words	#a	#s
guns	1	address, **bank**, bond, **notifi**, testam	28	204
guns	2	clinton, fbi, foreign, jim, spea	318	2
guns	3	**forgiv**, god, **hope**, **love**, **want**	4	381
guns	4	crime, firearm, gun, law, weapon	541	8
mideast	1	appressian, armenia, armenian, ohanu, proceed	464	5
mideast	2	arab, congress, isra, israel, jew	379	4
mideast	3	**bank**, **check**, funer, insur, testam	10	233
mideast	4	**forgiv**, good, **hope**, **love**, **want**	2	355
politics	1	compound, disclaim, fbi, govern, major	593	12
politics	2	clayton, cramer, optilink, relat, uunet	274	1
politics	3	**bank**, box, **check**, funer, **notifi**	11	258
politics	4	**forgiv**, good, **hope**, life, **love**	11	330
religion	1	**bank**, bond, **check**, **notifi**, paper	36	192
religion	2	frank, object, observ, theori, valu	279	0
religion	3	activ, christian, jesu, koresh, net	502	10
religion	4	**forgiv**, **hope**, **love**, sorri, **want**	12	395

Figure 2: MDS showing suicide notes and *talk.politics.mideast* articles (**s** character in the figure means suicide notes while **a** character depicts newsgroup article, colors are used as cluster numbers).
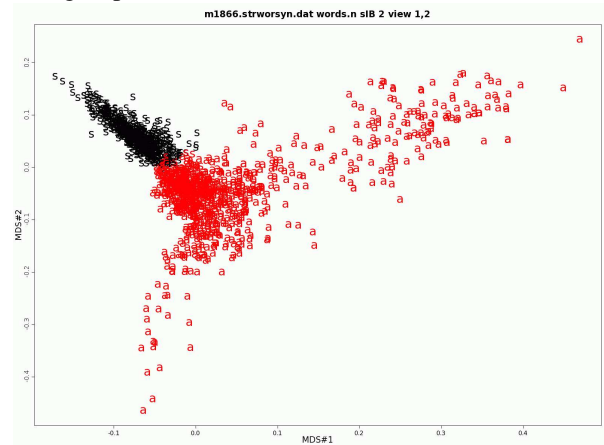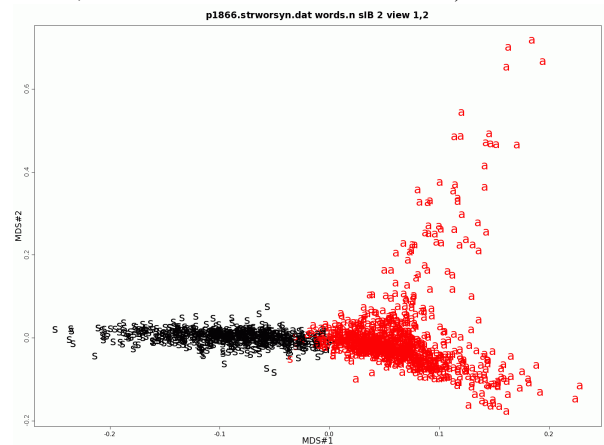

Figure 3: MDS showing suicide notes and *talk.politics.misc* articles (**s** character in the figure means suicide note while **a** character depicts newsgroup article, colors are used as cluster numbers).

5   Conclusions

Our findings suggest that unsupervised methods can distinguish between suicide notes and newsgroups, our proxy for general discussion. This is important because it is helpful in determining if NLP can be useful when integrating thought markers with biological and clinical markers ($f(B_m, T_m, C_m)$). In other words, can an NLP tools accurately distinguish between suicidal and normal thought markers ($T_m^S \neq T_m^N$)? Moreover these unsupervised methods have shown an ability to find sub-groups of suicide notes even when other types of newsgroups are present. In our analysis, one subgroup showed no emotional content while the other was emotionally charged. This finding is consistent with Tuckman's, 1959 work that showed suicide notes fall into six emotional categories: emotionally neutral, emotionally positive, emotionally negative directed inward, emotionally negative directed outward, emotionally negative directed inward and outward (Tuckman et al., 1959). The next step in developing a Suicide Risk Index is to conduct a clinical trail in the Emergency Department that will collect B_m, T_m, C_m and test multiple methods for computing the Suicide

Figure 4: MDS showing suicide notes and *talk.religion.misc* articles (**s** character in the figure means suicide note while **a** character depicts newsgroup article, colors are used as cluster numbers).
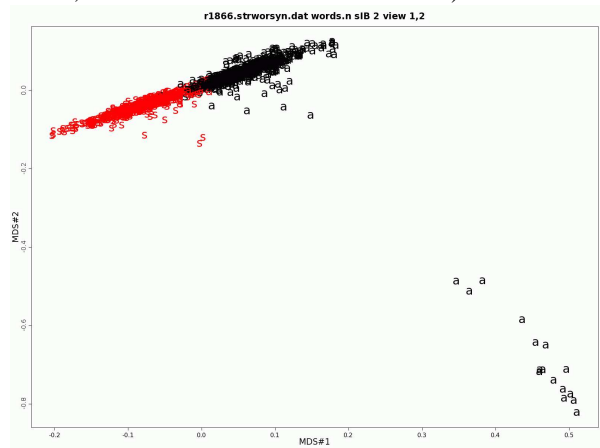
Risk Index.

References

A. Chavez, D. Paramo-Castillo, A. Leenaars, and L. Leenaars. 2006. Suicide notes in mexico: What do they tell us? Suicide and Life-Threatening Behavior, 36:709–715.

C.K. Chung and J.W. Pennebaker, 2007. The psychological functions of function words, pages 343–359. New York: Psychology Press.

T. F. Cox and M. A. A. Cox. 1994. Multidimensional Scaling. Chapman and Hall.

B. Demirel, T. Akar, A. Sayin, S. Candansayar, and A Leenaars. 2007. Farewell to the world: Suicide notes from turkey. Suicide and Life-Threatening Behavior, 38:123–128.

S. Girdhar, A. Leenaars, T.D. Dogra, L. Leenaars, and G. Kumar. 2004. Suicide notes in india: what do they tell us? Archives of Suicide Research, 8:179–185.

T. Ho, P. Yip, C. Chiu, and P. Halliday. 1998. Suicide notes: what do they tell us? Acta Psychiatrica Scandinavica, 98:467–473.

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

R. Maris. 1981. Pathways to suicide. John Hopkins University Press, Baltimore, MD.

D Ogilvie, P. Stone, and E. Shneidman. 1969. Some characteristics of genuine versus simulated suicide notes. Bulletin of Suicidology, 1:17–26.

J. W. Pennebaker, M. E. Francis, and R. J. Booth. 2001. Linguistic Inquiry and Word Count: LIWC. Lawrence Erlbaum Associates, Mahwah, NJ, 2nd edition.

J. P. Pestian, P. Matykiewicz, J. Grupp-Phelan, S. Arszman-Lavanier, J. Combs, and Robert Kowatch. 2008. Using natural language processing to classify suicide notes. In AMIA Annual Symposium Proceedings, volume 2008. American Medical Informatics Association.

C. J. Van Rijsbergen. 1979. Information Retrieval. Butterworth-Heinemann, Newton, MA, USA.

E Shneidman and N Farberow. 1957. Clues to Suicide. McGraw Hill Paperbacks.

Noam Slonim, Nir Friedman, and Naftali Tishby. 2002. Unsupervised document classification using sequential information maximization. In Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 129–136.

Jacob Tuckman, Robert J. Kleiner, and Martha Lavell. 1959. Emotional content of suicide notes. Am J Psychiatry, 116(1):59–63.

Ian H. Witten and Eibe Frank. 2000. Data mining: practical machine learning tools and techniques with Java implementations. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

World Health Organization, 2001. Burden of mental and behavioral disorders, pages 19–45. World Health Organization, Geneva.