Clustering semantic spaces of suicide notes and newsgroups articles

P. Matykiewicz\textsuperscript{1,2}, W. Duch\textsuperscript{2}, J. Pestian\textsuperscript{1}
\textsuperscript{1}Cincinnati Children’s Hospital Medical Center, University of Cincinnati,
\textsuperscript{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\textsubscript{m}) with Thought Markers (T\textsubscript{m}), and use machine learning to computer 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 Department (ED), is to assess the likelihood of another attempt, a more lethal one. We believe to fully assess 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\textsubscript{m}: measured by the concentration of certain biochemical markers; and Thought Markers T\textsubscript{m}: measured by artifacts of thought that have been reduced to writing or transcribe speech. It is these T\textsubscript{m} where BioNLP plays an important role. In this study we focus on the T\textsubscript{m}. 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 intend to provide insight into how best to apply linguistic tools when responding to suicidal patients. An important contribution of this effort has been the creation of a suicide notes corpus. The corpus contains 866 notes from those who have completed suicide. 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 is described below.

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 assen-
tion 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 perform as well as mental health professionals when discriminating between elicited and genuine suicide notes. 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%. These initial result have inspired to conduct the described study (Pestian et al., 2008). This time the focus is on unsupervised machine learning methods. Even though text clustering has a long history (Manning and Sch`eutze, 1999) it has never been applied to suicide notes. Sections below describe the cluster analysis process and results.

2 Data

**Suicide Notes Corpus.**

Data for the suicide note data base were collected from around the United States. They were either in a handwritten 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 typed into the database exactly as seen. A second person reviewed what was typed. The table 1 provides various 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. This was not available and so 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 newsgroup have no relevance to suicide notes. Thus the following were selected: talk.politics.guns, talk.politics.mideast, talk.politics.misc, talk.religion.misc because of an ‘a priori’ assumption of their similarity to suicide. For example, one would expect religion to have more similarity to suicide than say the middle east. Each newsgroup contains 1000 articles (newsgroup postings). Headers and quotes from other postings were removed.

3 Methods

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. Next, Jim Richardson and Benjamin Franz implementation of English stemmer was included in the pre-processing stage (Lingua::Stem module\(^2\)). In addition features that appeared in less than 10 documents or in more than 500 documents were removed. Also 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 two steps of pre-processing are used to reduce outliers.

Calculations are done using open source software called R\(^4\). 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). This approach has been shown to work well when clustering documents. (Slonim et al., 2002). 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 best extract features that represent each cluster Pearson correlation coefficient is used. The correlation coefficient $CC$ is calculated between each feature and each cluster separately $CC(w_i, c_j)$. N best

\(^1\)http://archive.ics.uci.edu/ml/

\(^2\)http://www.perl.org

\(^3\)http://search.cpan.org

\(^4\)http://www.r-project.org
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.

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

| Description              | 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) |

There are total four newsgroup data sets: 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 has 1866 documents before document and feature selection is applied. Table 2 has final number of features while Table 3 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=219). Table 2 summarizes performance of best algorithms for each data set if two clusters are chosen.

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 anonymized 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 3 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
Table 3: 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, love, want | 11 | 330 |
| religion| 1  | bank, bond, check, notifi, paper | 36 | 192 |
| religion| 2  | frank, object, observ, theor, valu | 279 | 0 |
| religion| 3  | activ, christian, jesu, koresh, net | 502 | 10 |
| religion| 4  | forgiv, hope, love, sorri, want | 12 | 395 |

two dimensional plane using multidimensional scaling (MDS) initialized by principal component analysis. Each figure has different rotation but the shape is very similar in all cases. In addition MDS shows very little mixing which is also explained by results in the table 2.

5 Conclusions

Our findings suggest that unsupervised methods can distinguish between suicide notes and newsgroups. Moreover it is possible to find consistent sub-grouping of suicide notes in presence of different types of news groups. One sub-group of suicide notes show no emotional content while the is emotionally charged. This finding is supported by work of Tuckman, 1959 who classifies suicide notes 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). Further investigation should reveal if it is possible to break down the emotional subgroup even further using NLP tools. Another question is how do positive, neutral and neg-
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).

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).

ative emotional semantic features contribute to the over all risk assessment score?

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.

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.

Christopher D. Manning and Hinrich Sch"{u}tze. 1999. Foundations of statistical natural language processing. MIT Press.

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. 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.