Suicidal Tendencies: The Automatic Classification of Suicidal and Non-Suicidal Lyricists Using NLP

Matthew Mulholland
Montclair State University
Montclair, NJ, USA
mulhollandm2@mail.montclair.edu

Joanne Quinn
Montclair State University
Montclair, NJ, USA
quinnj11@mail.montclair.edu

Abstract

Can natural language processing be used to predict the likelihood that a musician will commit or has committed suicide? In order to explore this idea, we built a corpus of songs that includes a development set, a training set, and a test set, all consisting of different lyricists. Various vocabulary and syntactic features were then calculated in order to create a suicide/non-suicide song classifier. The features were input into the Weka machine learning suite and tested with an array of algorithms. We were able to achieve up to a 70.6% classification rate with the SimpleCart algorithm, a 12.8% increase over the majority-class baseline. Our findings suggest that syntactic and vocabulary features are useful indicators of the likelihood that a lyricist will commit or has committed suicide.

1 Introduction

Recently, research into the application of NLP to the detection of health illnesses has proved fruitful. For instance, in their study of the effects of dementia on writers, Le et al. (2011), guided by previous research, explored various hypotheses in the novels of three British writers. Their research found a decline in the type-token ratio of the novelists suffering from dementia. The use of passive constructions was also explored since it is generally believed that it represents a syntactic structure that is particularly complex and likely would be used less often by writers suffering from dementia. Indeed, they found that authors with dementia use less passives than their healthy peers. Their results indicate the potential for natural language processing in the language of mental illness. Similarly, much research into the application of NLP to depression and suicide prediction has been conducted in recent years (Pestian et al. 2012; Sohn et al. 2012).

While one might not expect depression or a suicidal tendency to affect language in the same way as an illness such as dementia, it is reasonable to assume that there will be textual indications of these mental illnesses also. In Stirman and Pennebaker (2001), word use is treated as an indicator of the mental states of suicidal and non-suicidal poets. Stirman & Pennebaker developed the Linguistic Inquiry and Word Count program to analyze over 70 language dimensions, including: polarity, affect states, death, sexuality, tense, etc. Their research found a correlation between the likelihood of a poet committing suicide and his/her disengagement from society based on the suicidal poets' heavy use of first-person singular pronouns and decreased use of the first-person plural pronouns. Interestingly, they also noted that poets who had committed suicide generally used more sexual words and references to death than their non-suicidal counterparts. Additionally, latent semantic analysis has been used to detect depression in free texts (Neuman et al. 2010). By creating a semantic field from the words commonly associated with the concept of depression, Neuman et al. were able to accurately identify depressed people through their writing. Finally, other pertinent research involves the use of concrete nouns and the lack of abstract concepts in professionally-written poetry (Kao & Jurafsky, 2012).

Based on the prior research, we anticipated that the suicidal lyricists' use of first-person singular pronouns would differ from that of the non-suicidal lyricists, possibly being significantly higher. We also hypothesized that certain features like the usage of sensual and morbid words and the passive construction would be more prevalent in the works of suicidal lyricists. Additionally, we were interested in exploring the differences among other features, such as TTR, the degree to which

International Joint Conference on Natural Language Processing, pages 680–684,
Nagoya, Japan, 14-18 October 2013.

a text is polar and in which direction, the n-gram profiles of songs in relation to the n-gram profiles of suicide/non-suicide lyricists in general, and the semantic fields of other target emotions and affect states of the two groups.

2 Methods

First, a corpus of songs by male suicide and non-suicide lyricists was constructed, which consists of training, development, and test sets. The training set is comprised of 533 songs, of which 253 were written by four lyricists who have not committed suicide and 280 by five lyricists who did commit suicide. The test set consists of 63 songs by 5 non-suicidal lyricists and 46 songs by 4 suicial lyricists. Finally, the development set consists of 168 songs, 94 from 5 non-suicidal lyricists and 74 from 6 suicidal lyricists. Finally, the test set contains 63 songs written by five non-suicide artists and 46 songs written by 4 suicide artists. Table 1 displays the composition of the sets.

In our search for lyricists who committed suicide, we looked for lyricists who met the following prerequisite: the suicide had to be relatively unambiguous. This requirement constrained the size of the corpus a great deal. In addition, we attempted to distribute the lyrcists across the different sets in an even manner such that each set would be comprised of lyricists of a range of times and nationalities. For these reasons, it was a source of difficulty trying to create sets with unique lyricists. Although we did not require that each song be solely written by the lyricist in question due to the fact that it is often murky concerning to whom the lyrics should be attributed, we did make an attempt to exclude songs written entirely by bandmates or other musicians. Due to the lack of female lyricists who committed suicide, we were forced to consider exclusively male lyricists.

48% of the non-suicide corpus consists of songs written by one artist, Bob Dylan. Removal of 35% of those Dylan songs (and thus evening out the distribution of songs) did not significantly alter the classifier's results. The other non-suicide lyricists contributed between 32-45 songs each. The training set of suicide songs is slightly over-represented by Elliott Smith (about 33% of the songs). The four remaining suicide artists each contributed between 11% and 20% of all songs in this set. We additionally used a development set of five non-suicide lyricists (94 songs) and six suicide lyricists (74 songs), which was used to compute n-gram features.

Each song considered was searched for in online lyrics databases and was cleaned by hand, the lines being joined into punctuated sentences or phrases so that a POS-tagger and lemmatizer could be used. The lyrics were then tokenized using the OpenNLP tokenizer and lemmatized. Features were computed using Python and with some help from the UAM corpus tool (O'Donnell 2008)(which uses the Stanford Parser), especially for grammatical analysis.

Lyricist	**Suicide**	**Set**	**Songs**
Bob Dylan	n	train	123
Bob Marley	n	train	42
Mike Ness	n	train	43
Trent Reznor	n	train	45
Elliott Smith	y	train	99
Ian Curtis	y	train	40
Kurt Cobain	y	train	53
Pete Ham	y	train	34
Phil Ochs	y	train	54
Total		**train**	**533**
Ben Folds	n	test	11
Chris Bell	n	test	12
John Lennon	n	test	20
Neil Young	n	test	10
Paul Simon	n	test	10
Doug Hopkins	y	test	4
Peter Bellamy	y	test	18
Richard Manuel	y	test	10
Tom Evans	y	test	14
Total		**test**	**109**
Beck Hansen	n	dev	10
George Harrison	n	dev	24
Johnny Cash	n	dev	22
Thom Yorke	n	dev	13
Tom Petty	n	dev	25
Adrian Borland	y	dev	27
Darby Crash	y	dev	8
Jim Ellison	y	dev	12
Mel Street	y	dev	5
Michael Hutchence	y	dev	3
Stuart Adamson	y	dev	19
Total		**dev**	**168**

Table 1: Composition of Corpus Sets

3 Features

In order to create the suicide/non-suicide lyrics classifier, similar features to those used in the previous research were explored in conjunction with a set of original features. In all, there were 87 features that we explored.

3.1 Vocabulary Features

While a few of our features were based on raw counts of types, tokens, and time of song (in seconds) alone, such as TTR (type-token ratio), they are mostly used to normalize many of the features in the following two sections. Below are the vocabulary features we explored:

- by type: type/token ratio (TTR) and type/time ratio
- by token: token/time ratio

3.2 Syntactic Features

As in Le et al. (2011) and Stirman & Pennebaker (2001), we expected to find differences in the use of the passive constuction and in the proportions of the first-person pronouns to the rest of the pronouns. We expected a greater use of passive constructions in the lyrics of the suicide lyricists in comparison to the non-suicide lyricists since it might signify a greater sense of disengagement from the external world. Additionally, we hypothesized a higher proportion of first-person pronouns to other pronouns in the suicidal lyrics since a common perception about suicide cases and depressive people in general is that they are more self-centered or that they are less concerned with others.

In addition to the exploration of the first-person pronouns and passive constructions, we also looked at the differences in the usage of mental-state verbs co-occuring with the first-person singular pronoun, including the use of verbs like *think* and *feel*. Our expectation here was that the suicide writers might use such constructions more often due perhaps to a preoccupation with thoughts and feelings and an inclination to think and feel more often than act.

These features were computed using the UAM corpus tool, which allows one to create autocoding rules and presents annotation statistics on the text level. Most of these count features were normalized by type, token, and time, but a few of them consist of ratios between features, such as first-person singular pronouns to all other pronouns. Since the latter features were occasionally affected by data sparsity, we chose to deal with undefined values resulting from zeroes in the denominator by adding 0.01 to each count so that the resulting value would not be undefined.

3.3 Semantic Class Features

Our expectations about the content of the suicide lyrics in comparison to the non-suicide lyrics was that they might deal with more negative, depressing subjects than positive ones. We also hypothesized that, as in Pennebaker & Stirman (2001), we would see a difference in the use of sexual words in the suicide lyrics (specifically, a heightened rate of sexual and death-related word usage). For these and other "semantic classes", we built word-lists consisting of terms relating to the target semantic classes.

The semantic classes considered were sensuality, action (specifically, verbs that signified some particular action), concreteness (specifically, nouns that represented concrete objects), death, love, depression, and drugs. We used the MPQA prior polarity word lexicon (Wiebe et al. 2005) to measure negative, positive, and neutral word usage. We counted the number of occurrences of these words in each song-text, normalizing the raw counts by type, token, and time. We also computed a number of features that consisted of ratios between raw counts, such as sensual words to positive words. Where applicable, we dealt with data sparsity using the same method described above, adding 0.01 to avoid undefined values.

Additionally, we used the AFINN (Nielsen, 2011) word-valence dictionary, which is a list of nearly 2,500 polar terms with associated polarity values (ranging from 5 to -5), to calculate the total and average polar value of each song. The total polar value was calculated by summing the polarity values for each polar term in a song-text while the average polar value was calculated by dividing the total polar value by the number of polar terms in a text.

3.4 N-Gram Features

A Python script was written to build unique n-gram (for n = 1 to n = 6) profiles for the classes in the development set. (This set was used exclusively for this purpose and was not needed for any other features.) These unique n-gram profiles for

suicide and non-suicide lyricists were compared to the corresponding n-gram profiles of each song in the training and test sets to find out to what extent the n-grams in a given song overlapped with either n-gram profile.

In addition to percentage of overlapping n-grams, a number of features composed of the overlapping percentages were computed. The difference between overlapping n-gram percentages for each class was calculated for each value of n. For example, if a song's bigram profile overlapped with the non-suicide bigram profile at a rate of 6% and with the suicide bigram profile at a rate of 4%, the difference would be 2%. We also computed the average overlapping n-gram percentage for each class across all values of n. Finally, we calculated the difference between the average percentages of overlap for the non-suicide n-gram profiles and that for the suicide n-gram profiles.

4 Results

All features were input into Weka and a number of different ML algorithms were run to create a classifier. As a baseline for comparison, we used the majority-class prediction rate of 57.7%. The classifier was trained on the training set and then tested on the test set of different artists' songs. The most successful algorithm was the SimpleCart algorithm, which correctly classified the songs as either suicide or non-suicide 70.6% of the time and which achieved a 0.39 Cohen's kappa value. The correct classification rate represents an increase of 12.9% over the baseline. The SimpleCart algorithm achieved a precision, recall, and F-measure of 0.71, and an ROC Area value of 0.70. In Table 4 above, we report the confusion matrix for the test set.

While our classification statistics do not reach a satisfactory level, we believe that they indicate that we are on the right track and that this task can be tackled using NLP. Of the 87 features that were calculated, a number stood out as being most useful across numerous algorithms. Included among these features are the various n-gram features, the first-person singular + mental verb features, the concrete nouns, neutral terms, sensual words, and total polar value semantic class features, and the first-person singular and passive construction syntactic features.

a	b	<– classified as
29	17	a = suicide
15	48	b = non-suicide

Table 2: Confusion Matrix.

5 Discussion

The construction of a corpus for this type of task is beset by problems on all sides. Perhaps one of the largest issues is with the seemingly non-suicidal lyricists: whether these lyricists have passed away already or are still alive, there is no certainty that they would not have committed/will not commit suicide at some point after the point at which they are classified as non-suicidal. Perhaps one could try to use only those lyricists who died from non-suicidal causes late in life (say, after 60) since such lyricists might represent fairly safe cases, but even then there still would not be any certainty. Furthermore, there could be a situation in which a lyricist tried to commit suicide, but did succeed and news of the attempt was kept secret. Such lyricists might then be classified as non-suicides even though the amount of separation between them and the lyricists who were successful at committing suicide is next to nill. Although we acknowledge that this is a serious consideration, we believe that 1) it is a risk that we have to take in order to do this task since there seems to be no solution that guarantees 100% certainty and 2) the likelihood of committing suicide appears to be so low (even for artists) that it might not be such a bad course of action to assume that any given lyricist will not commit suicide unless he/she already has.

Though the vast majority of lyricists do not commit suicide, this fact leads directly to some of the other problems that afflicted the construction of our corpus. Since the number of lyricists who committed suicide is constrained, this leads to issues beyond the collection of as many songs as possible from such lyricists, which is a necessity. For example, the corpus of songs cannot simply be randomly split into sets because the we need to ensure that the songs we test on were not written by lyricists who composed songs that we trained on lest we merely learn the tendencies of those artists and not the abstract suicidal tendency that we are actually seeking. The same issue goes for the development set. However, even if we figured out a way to split the corpus into sets of the appropriate numbers of songs while taking into consideration

that the artists for each set must be unique, there are further factors that could skew the results. For example, we would ideally want the composition of each set of songs to be consistent from set to set in terms of the range of composition dates, the genres represented, etc. Although we attempted to take all of these factors into consideration in order to partition the data into sets, we realize that our method for doing so and the product of our labors leave much to be desired. In the future, we hope to work on refining our method so that it might optimize the partioning of our corpus.

6 Further Research

Besides expanding the corpus to include many more non-suicide lyricists and (at the very least) to include more songs from each of the suicide artists, it would perhaps be fruitful to extend the analysis to other types of features and new lexicons since it has been demonstrated that this task could be solved using NLP.

Acknowledgments

We would like to take this space to acknowledge again the use of the UAM corpus tool, which proved valuable to our analysis, the MPQA prior polarity lexicon, the AFINN word valence dictionary, the Weka machine learning suite, and the many online contributors of lyrics and word-lists. We would also like to thank Michael Flor for letting us use his own personal lemmatizer.

This material is based in part upon work supported by the National Science Foundation under Grant Numbers 0916280 and 1048406. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Justine Kao and Dan Jurafsky 2012. A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry. *NAACL-HLT* 2012, 8.

Xuan Le, Ian Lancashire, Graeme Hirst, and Regina Jokel. 2011. Longitudinal detection of dementia through lexical and syntactic changes in writing: A case study of three British novelists. *Literary and linguistic computing*, 26(4):435-61.

Y. Neuman, Y. Cohen, G. Kedma, and O. Nave. 2010. Using Web-Intelligence for Excavating the Emerging Meaning of Target-Concepts. *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Toronto, Aug-Sept 2010. IEEE Computer Society.

F. Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *Proceedings of the ESWC2011 Workshop on Making Sense of Microposts: Big things come in small pack- ages,* number 718 in CEUR Workshop Proceedings, Heraklion 2011.

Mick O'Donnell 2008. emonstration of the UAM CorpusTool for text and image annotation. *2010 Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies Demo Session - HLT* Association for Computational Linguistics, 2008.

J. P. Pestian, P. Matykiewicz, and M. Linn-Gust. 2012. Whats in a Note: Construction of a Suicide Note Corpus. *Biomedical Informatics Insights*, 2012:5 1-6.

M. F. Porter 1980. An algorithm for suffix stripping. *Program*, 14(3):130137.

S. Sohn, M. Torii, D. Li, K. Wagholikar, S. Wu, and H. Liu. 2012. A Hybrid Approach to Sentiment Sentence Classification in Suicide Notes. *Biomedical Informatics Insights*, 2012:5 (Suppl. 1) 43 50.

S. W. Stirman and J. W. Pennebaker. 2001. Word Use in the Poetry of Suicidal and Nonsuicidal Poets. *Psychosomatic Medicine* 2001, 63:517-522.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* , 39(2-3):165210.

Ian H. Witten and Eibe Frank. 1999. Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kauffman (1999).