PERSEREC

Management Report 17-03
July 2017

An Evaluation of a Machine Learning Model for Suicide Detection and Prevention

David A. Ciani
Jessica A. Wortman, Ph.D.
Christina M. Hesse
Michael D. Robinson, MPH
Northrop Grumman Technology Services

Andrée E. Rose
Defense Personnel and Security Research Center
Office of People Analytics

Approved for Public Distribution
Defense Personnel and Security Research Center
Office of People Analytics

Management Report 17-03 **July 2017**

An Evaluation of a Machine Learning Model for Suicide Detection and Prevention

David A. Ciani, Jessica A. Wortman, Ph.D., Christina M. Hesse, Michael D. Robinson, MPH—*Northrop Grumman Technology Services*

Andrée E. Rose—*Defense Personnel and Security Research Center/OPA*

Released by – Eric L. Lang, Ph.D.

Defense Personnel and Security Research Center
Office of People Analytics
400 Gigling Rd.
Seaside, CA 93955

PREFACE

In 2015, the Defense Suicide Prevention Office (DSPO) sponsored the development and subsequent validation of an algorithm to determine the effectiveness of a fully automated approach to identify Service members at risk for suicide using publicly available social media information. In this effort, the Defense Personnel and Security Research Center (PERSEREC) reviewed the algorithm to determine how well it met its goal of identifying Service members who might be at risk of dying by suicide, and to suggest any improvements and recommendations for use of this algorithm in the future. The purpose of this report is to present the findings of the independent validation and verification of the suicide risk algorithm.

Eric L. Lang, Ph.D.
Director, PERSEREC

EXECUTIVE SUMMARY

In Fiscal Year (FY) 2015, the Defense Suicide Prevention Office (DSPO) funded the Defense Personnel and Security Research Center (PERSEREC)[1] to conduct an independent verification and validation (IV&V) of a suicide risk algorithm using publicly available social media information to identify Service members at potential risk of suicide. The suicide risk algorithm was designed to identify potentially at-risk Service members in a completely automated fashion using their publicly available social media information. The algorithm generates in a suicide risk score, which one then can compare to the actual cause of death, to determine the effectiveness of the algorithm. The IV&V presented here is intended to determine whether the algorithm met the goals of the project (i.e., correctly identified Service members at risk of dying by suicide), and to identify model improvements or recommendations for any future algorithm development.

METHODOLOGY

The algorithm development relied on a sample of 1,400 deceased Service members, half of whom died by suicide, and half of whom died by other causes. Using this sample, the social media vendor conducted a fully automated procedure to find, save, redact, and subsequently process and model any publicly available social media text data that they could link with confidence to a Service member in the sample. Note that this reduced the usable sample size to 185 individuals with text data. The vendor provided these text data, along with the program code they developed to calculate the risk score for each Service member to PERSEREC analysts who then performed the IV&V. PERSEREC analysts tested and validated the code to determine if they could reproduce the results provided by the social media vendor. Analysts also determined if they could make any improvements to the model.

RESULTS

Analysts at PERSEREC were able to run the algorithm and reproduce the results offered by the vendor. In addition, analysts ran experiments to improve model performance and settled on the following major modifications:

- revise the demographic data to use standard Department of Defense (DoD) occupation codes,
- recalculating sentiment scores (which assign scores to words to get a sense of emotional content of the text),
- improving the missing data technique, and,

[1] At the time this research effort began, PERSEREC was a division of the Defense Manpower Data Center (DMDC), a component of the Defense Human Resources Activity (DHRA). Subsequently PERSEREC transitioned to the newly established Office of People Analytics (OPA), also a component of DHRA.

- focus on the sub-sample of individuals who had publicly available social media data.

Overall, the authors conclude that this model for identifying Service members at risk for suicide is valid. While at this level of development, it is nowhere close to being able to identify all cases of death by suicide using social media data, it does identify roughly 70-80% of the cases when social media data are available for a subject.

Any future algorithm development for suicide risk should leverage the improvements presented here. In addition, future work must take into consideration changes in data access practices and policies for the largest source of data (Facebook). The authors of this study also suggest further investigation into applying these techniques at a unit or organizational level, to reduce potential privacy issues with singling out individuals based on their online activities. Results of this study suggest that any algorithm development on social media information needs to be flexible to address the fact that social media data are a rapidly changing source of information.

TABLE OF CONTENTS

INTRODUCTION 1

BACKGROUND 2

TEXT ANALYSIS IN BEHAVIORAL RESEARCH 2

MACHINE LEARNING 3

Decision Trees 4

Logistic Regression 4

METHOD AND EVALUATION STRATEGY 6

REVIEW OF VENDOR IMPLEMENTATION 6

TEST AND VALIDATE 6

CONDUCT EXPERIMENTS TO IMPROVE RESULTS 7

THE SUICIDE RISK PREDICTION MODEL 8

SUBJECT CASES 8

VENDOR IMPLEMENTATION 8

Data Loading and Preparation 8

Sentiment Score Generation 8

Keyword Score 9

Training The Classifier 9

Model Features 9

DISCUSSION AND EXPERIMENTS TO IMPROVE RESULTS 10

Data Availability and Subject Focus 10

Demographic Data 11

Sentiment Scoring Technique 11

Other Architecture Improvements 12

RESULTS 13

DISCUSSION AND RECOMMENDATIONS 15

OPTIONS TO OPERATIONALIZE 16

RECENT CHANGES TO SOCIAL MEDIA PLATFORM POLICIES 16

REFERENCES 19

APPENDIX A : LABORATORY NOTES A-1

APPENDIX B : STATISTICS FOR EVALUATING PREDICTIVE MODELS B-1

APPENDIX C : DATA SET DEMOGRAPHICS AND DESCRIPTIVE STATISTICS C-1

APPENDIX D : MODEL FEATURES D-1

LIST OF TABLES

Table 1 Performance Metrics 13

Table 2 Vendor Implementation, All Subjects Confusion Matrix 13

Table 3 Vendor Implementation, Social Media Subjects Confusion Matrix 14

Table 4 Improved with All Variables, Social Media Subjects Confusion Matrix 14
Table 5 Improved with Text Variables Only, Social Media Subjects Confusion Matrix 14

LIST OF FIGURES

Figure 1 Evaluation Process Map 6

LIST OF TABLES IN APPENDICES

Table C-1 Gender C-3
Table C-2 Marital Status C-3
Table C-3 Dependents C-4
Table C-4 Race C-4
Table C-5 Education C-4
Table C-6 Religious Affiliation C-4
Table C-7 Military Component C-5
Table C-8 Military Branch C-5
Table C-9 Service Rank C-5
Table C-10 Gender C-6
Table C-11 Military Branch C-6
Table C-12 Service Rank C-6
Table D-1 Vendor Demographic Features D-3
Table D-2 Vendor URL Metadata Features D-4

LIST OF FIGURES IN APPENDICES

Figure B-1 Confusion Matrix Diagram B-3

INTRODUCTION

Suicide among Service members continues to be a matter of serious concern for the Department of Defense (DoD), with suicide rates in the DoD population continuing to be high. Recent studies funded by the Defense Suicide Prevention Office (DSPO) and conducted by the Defense Personnel and Security Research Center (PERSEREC; a division of the Office of People Analytics) have explored the utility of publicly available social networking data in predicting suicide risk among active duty Service members (Hesse, Bryan, & Rose, 2015; Rose & Hesse, 2015; Wortman, Hesse, & Shechter, 2016; Whiteley & Rose, 2016). These research efforts have utilized manual coding of online content posted by Service members in order to identify possible psychological characteristics or emerging themes that might differentiate Service members who die by suicide from Service members who died from other causes.

Although the results of such studies have been promising, the process of manual coding is costly and labor-intensive, making it impractical for implementation on a large scale. In this study, we explore the possibility of automating the process of detecting suicide risk in Service members using publicly available social media information, including social networks. On behalf of DSPO and PERSEREC, a commercial social media vendor created a suicide risk model to try to identify Service members who might be at risk for suicide. This paper will explain the background behind the model development process as well as document an independent validation and verification (IV&V). Additionally, the report will discuss considerations for potential implementation.

The purpose of conducting an IV&V of machine learning or other statistical model is to make sure that the product actually achieves the goals and metrics that it purports to meet. This is especially important when making more subjective design decisions in the construction of a data-driven model. After evaluating the model developed by the social media vendor, the authors explored various options for improving the performance of the model and developed recommendations for potential implementers.

BACKGROUND

In FY14, DSPO funded a research effort conducted by PERSEREC that attempted to leverage the content of Service members' social networking pages to predict risk of suicide. The first phase, *Indicators of Intent to Die by Suicide: Phase 1* (Hesse, Bryan, & Rose, 2015), used a sample of 1,400 deceased Service members, half of whom died by suicide, and half of whom died from other causes. A commercial social media vendor collected all publicly available social networking data for this sample. Researchers at the University of Utah's National Center for Veteran's Studies (NCVS) then coded these data for 36 clinical indicators of suicide risk. The results revealed that clinical indicators can be used to distinguish Service members who died by suicide from those who died from other causes.

The coding involved in the creation of these data was extremely labor intensive, requiring careful training of multiple coders to read all of the available social network posts. Thus, in FY15 DSPO funded PERSEREC to explore the possibility of automating the process of obtaining text from publicly available online sources, such as social networks, and using that text to identify Service members who might be at risk for suicide. Using the same sample of 1,400 deceased Service members collected for the previous study, the social media vendor created an automated process to collect publicly available text, and then used that text to try to predict suicide risk. This technique builds upon previous research that leverages human-generated content to predict psychological characteristics, including possible risk of suicide.

TEXT ANALYSIS IN BEHAVIORAL RESEARCH

Content analysis of human-generated text has a long history in the field of psychology, beginning with manual coding of content (e.g., Gottshalk & Gleser, 1969; Gottschalk, Winget, & Gleser, 1979) and leading up to more complex, automated, computer-based approaches (e.g., Pestian, Nasrallah, Matykiewicz, Bennett, & Leenaars, 2010; Stone, Bales, Namenwirth, & Ogilvie, 1962; Tausczik & Pennebaker, 2010).

In recent years, text analytics has continued to grow in popularity and in sophistication. With growing use of social networks, online blogging, and electronic information in general, large bodies of text are increasingly available. In addition, modeling capabilities with regard to prediction of psychological states have greatly improved with the introduction of machine-learning algorithms. As a result, there is an increasingly large body of literature regarding prediction of mental health, psychopathology, and risk of events such as suicide.

Studies that are more recent have used computer science-based approaches to classify individuals into at risk of suicide or not. Using a variety of strategies, including combinations of human annotated text and machine learning, researchers have reached high levels of accuracy in identifying text that might signal suicidal intent. In fact, in one of the most successful efforts, the machine

classifier was able to reach a similar level of accuracy achieved by human coders (O'Dea, et al., 2015). This suggests that machine classifiers can be extremely effective in classifying suicidal versus non-suicidal text. However, the fact that the study in question pulled text related to suicide and focused on classifying it as whether it was concerning/strongly concerning, or safe to ignore, limits the broader applicability of this study. In other words, the goal of this algorithm was to distinguish between different types of suicidal text, rather than identifying potential risk of suicide of a single individual.

Another recent study leveraged a combination of human-annotated text (words and phrases selected because they signaled possible suicidal intent) and machine-learning algorithms to distinguish suicide-related communication (e.g., discussion of a celebrity suicide) from text that indicates suicidal ideation (Burnap, Colombo, & Scourfield, 2015). Using this approach, researchers correctly classified suicidal text approximately 70% of the time, which is similar to the results of the machine learning technique that this report will demonstrate. However, as with the study described earlier, this algorithm focused on identifying suicidal text, rather than identifying individuals at risk for suicide using available text.

In addition, although many studies have explored this question in the general population, far fewer studies have attempted to identify at-risk individuals within the population of United States military Service members. These individuals present a unique set of challenges, especially when attempting to predict possible risk of suicide because they have struggles that are specific to their roles as military personnel (Bryan, Morrow, Anestis, & Joiner, 2010). As a result, additional studies should explore the possibility of predicting suicide risk using automated processing of text within this population.

In order to describe the process of developing a suicide-risk model for Service members, this report begins by introducing machine-learning algorithms in general. Following this, the report will describe an independent verification and validation of the model developed by the vendor in detail.

MACHINE LEARNING

Recent research developed automated procedures by which computers can extract actionable information from data, a process known as machine learning (Alpaydin, 2014). This involves programming a computer to optimize a solution to a problem when given a set of data. The goal is to program computers to use example data or past knowledge to solve a particular problem. Generally, machine-learning approaches begin with a large body, or corpus, of data, as an example of the desired output, and then develop a corresponding optimal model or solution.

Machine-learning algorithms, then, are a set of procedures performed by a computer to transform a given input into an output of interest. These algorithms can be either descriptive (describe a set of data) or predictive (attempt to predict a future outcome). A model of suicide risk, for example, would be predictive, because

the goal is to predict a possible future event (suicide). Other types of algorithms might seek only to describe a set of data, rather than predict the likelihood of a future event.

One must use care when developing models using machine-learning algorithms. An issue inherent to the machine-learning process is over fitting. Over fitting is where a model becomes highly optimized to a specific dataset, namely the dataset used to train it. In order to detect when this occurs, one should evaluate the performance of a machine-learning model using a set of observations separate from the training set.

There are a number of different machine-learning algorithms and techniques. What follows is a discussion of two of them, decision trees and logistic regression, which the suicide risk model uses in its implementation.

Decision Trees

Decision tree learning is a machine-learning method that constructs a predictive model for either classification (predicting discrete outcomes) or regression (predicting a continuous output variable) applications. The input for the model is a set of predictors for each observation. For each predictor, decision tree learning calculates a simple prediction model that best approximates the outcome of interest. In each step, the model divides these data into two using a threshold that relates to a particular outcome. As a result, the use of tree diagrams to represent these types of models is common, which divides the node for each variable into two separate partitions, called child nodes or leaves. If the outcome variable is continuous (such as an average score), the decision tree is known as a regression tree. If the outcome is categorical, it is a classification tree (Lewis, 2000). Decision trees have grown in popularity in recent years because they are intuitive and can explore many possible predictor variables simultaneously.

Some decision tree methods construct multiple decisions trees. A random forest classifier is one popular approach (Loh, 2011). The random forest method constructs many possible decision trees by randomly selecting sub-samples from the overall sample, and constructing a decision tree based on a random subset of the predictors. Then, each tree in the model votes for a given outcome, with the winner selected as the model's overall prediction. This helps to correct for possible overfitting by removing those variables that are only useful on a small portion of the dataset.

Logistic Regression

One of the simplest possible algorithms or models used to solve a classification problem is a logistic regression. A logistic regression is a type of linear model that calculates the strength of the association between predictors and a binary or dichotomous outcome. In order to select the predictors that are most strongly associated with the outcome, only those variables that substantially (significantly) contribute to the prediction of the outcome are kept. In doing so, researchers can

select the most parsimonious model that does not lose useful information provided by any variables in these data. In the case of social media, a logistic regression might use text features (such as n-grams[2], groups of one or more words) to predict death by suicide. The result of a logistic regression is an equation, which one can use to calculate any single individual's outcome based on the variables selected.

Although logistic regression can be very useful, models that contain a large number of possible predictors, especially when those predictors correlate with each other, can easily overwhelm it. As a result, machine-learning algorithms that use logistic regression most often do so in combination with other techniques. In addition, logistic regression is a parametric technique, meaning that it relies on certain assumptions about the nature of these data that may or may not hold true for a given dataset. Other machine-learning approaches do not make the same assumptions and so are somewhat more flexible.

[2] N-grams are sequence of words that appear in a document. Unigrams, bi-grams, and tri-grams are special cases where N=1, 2, or 3, respectively. By processing a document and extracting all of the sequences of words that occur in the document, one can generate a set of n-grams. The trivial case is the unigram, where the set is simply a list of words that appear in the document. Often after generating the n-grams, their frequency within a document is calculated and used for further analysis.

METHOD AND EVALUATION STRATEGY

In order to conduct a quality independent validation and verification, the authors conducted developed a process to verify the technical underpinnings of the vendor implementation, document it, and then apply a fresh perspective to the technique and develop improvements. Figure 1 illustrates this process.

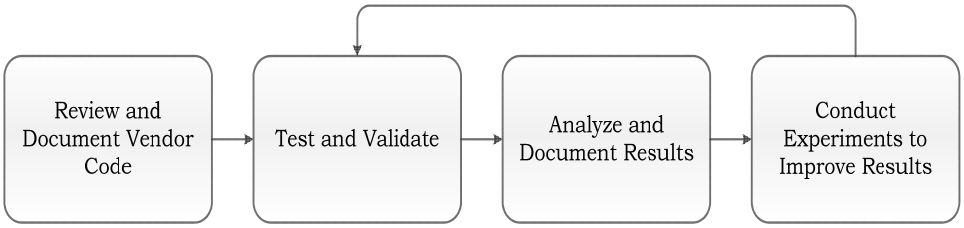

Figure 1 Evaluation Process Map

REVIEW OF VENDOR IMPLEMENTATION

The authors began by closely scrutinizing the source code developed by the vendor in order to validate the procedures implemented, looking for any obvious defects or "bugs." Additionally, the authors were able to gain a thorough understanding of the steps taken to process the data prior to building the model, as well as the model creation itself. In order to solidify understanding of the code, the authors took notes regarding the steps taken by the program. These notes were refined into a working document specifying the flow of the program, allowing the authors to examine the process and the details of the implementation. The authors also consulted technical and reference documentation to gain a thorough understanding of the library functions that the program employed. APPENDIX A contains the authors' laboratory notes captured during the documentation process.

TEST AND VALIDATE

Because machine-learning algorithms use sample data to train, it is important to have a separate data set for evaluation and validation, one that the model "hasn't seen before." Machine learning algorithms have a tendency to over-fit to small fluctuations in the training data, which results in poor predictive performance on new data. Without using a separate validation sets, researchers would not be able to detect this condition. For this effort, the authors accomplished this by randomly dividing the source data into two data sets, and training the model on the first data set (80% of the cases), and validating the model using the second data set (20% of the cases).

Using the statistical model evaluation techniques discussed in APPENDIX B, the authors examined the performance of the model. There are a number of different ways of interpreting the performance metrics. In this case, the authors chose to focus on the recall metric. The recall, or true positive rate (TPR), in this context, represents the rate at which the model successfully identifies suicide cases. This

metric does not focus on cases the model successfully identifies as non-suicide cases, or the cases the model does not correctly predict. The authors choose to focus on this metric, given that this model is likely to be implemented as a as a harm reduction tool. The concept is that not catching a suicide case is a significantly more costly error than applying an intervention to a case that turns out to be non-suicide.

In order to document the results of the testing and validation, the authors prepared confusion matrixes that tabulate the number of cases that the model correctly predicts as suicide and non-suicide, and the cases that it fails to correctly identify as suicide and non-suicide. They also prepared and reported on a selection of performance metrics including accuracy, recall, specificity, and F_1-score.

CONDUCT EXPERIMENTS TO IMPROVE RESULTS

The process of reviewing, testing, and validating the vendor implementation, left the impression on the authors there was room to improve the architecture and methods implemented by the model. To that end, they engaged in an iterative experimental process to improve various aspects of the model. After each improvement, they re-tested and validated the model and examined the performance metrics. The authors evaluated different combinations of improvements in order to optimize the results.

THE SUICIDE RISK PREDICTION MODEL

SUBJECT CASES

The cohort initially consisted of 1,400 Service members. The sample was stratified such that half, *n*=700 (50%), that died by suicide and the other half died by other means. PERSEREC research effort, *Indicators of Intent to Die by Suicide: Phase 1* (Hesse, Bryan, & Rose, 2015), originally sampled and collected data on this population. This study reuses the same cohort.

After excluding subjects for which no text data was found, 185 subjects remained, 86 (46%) of whom died by suicide. Analysis indicates that subjects with social media data were more educated and less likely to have died by compared to those with no social media data. APPENDIX C presents a detailed description of the sample's demographic characteristics.

VENDOR IMPLEMENTATION

The vendor implemented the model by using a random forest algorithm to implement a binary classifier. They implemented the model using a series of scripts in the Python programing language that leverage a number of Python libraries including Pandas, a data manipulation toolkit, and Scikit-Learn, a library that implements a number of machine learning algorithms.

Data Loading and Preparation

The program that generates the model follows a pattern typical for machine learning implementations. First, it retrieves data from a number of data sources, in this case a set of comma separated value (CSV) files containing the demographic data, as well as the social media data extracted from the vendor's automated systems. Before PERSEREC received the social media data, the vendor applied a PII removal routine in order to remove identifying information belonging to third parties.

The program then preforms a number of data cleaning and aggregation processes to make the data more suitable for use with the machine-learning algorithms. The program replaces missing data points using the mean or mode value for the variable and then transforms the categorical variables into binary indicators using a "dummy code" or one hot encoding technique.

Sentiment Score Generation

The next step taken by the program is to generate the sentiment score variables. For each word, the model obtains a net sentiment for each "sense," or meaning, of the word found in the SentiWordNet dictionary (Baccianella, Esuli, & Sebastiani, 2010). A word may have multiple senses with different sentiments. For example, the term "hot" has 21 different senses. These include both "recently stolen or smuggled" and "very popular or successful" which have markedly different sentiments.

The model then calculates a sentiment score for each chunk of text by:

(1) The program obtains the net sentiment of each sense (some senses have both a positive and negative sentiment component) by subtracting the word's negative sentiment score from its positive sentiment score.

(2) Then the program averages those net sentiments of each sense to obtain a single sentiment for the word.

Using the chunk sentiment score, the model distributes the chunk into 11 bins using a histogram function. These bins span from -1 to 1.2, with a width of 0.2 units. The vendor's model hard-codes the dimensions of the bins and the authors were unable to determine how the vendor selected these parameters. This results in 11 features for each subject, indicating the number of chunks of content at that sentiment level.

Keyword Score

The program also generates a keyword score feature by turning all of the text data into a term-document frequency matrix that counts the number of times each word occurs in each document. Then the program reduces number of terms by retaining only terms have a statistically significant correlation with death by suicide. The program uses the remaining terms to build a logistic regression model, applying it to the term frequencies for each subject, resulting in a single, numeric score.

Training The Classifier

The program uses a K-fold method to enable the model to be both trained and evaluated using the whole data set, rather than a traditional random train-test split. The K-fold method divides the dataset into a number of subsets, called folds. The vendor uses 10 folds in this case. The program then iterates of the subsets using nine sets to train a random forest model, while generating predictions on the tenth subset. After iterating over all 10 subsets, the program will have made predictions for the entire dataset. One then can evaluate the model's performance by comparing those predictions with the known true condition.

Model Features

The model incorporates a number of different types of features:

(1) The demographic data was drawn from the SDR as well as the National Death Index (NDI) and consisted of 41 fields of data (i.e. Faith Group, Marital Status, Rank etc.) prior to the data preparation step. In the vendors implementation, each candidate feature was evaluated and only features significant (using a phi-coefficient or point bi-serial correlation coefficient) at a p-value of 0.05 were retained.

(2) Social media metadata automatically aggregated by the vendor's off-the-shelf platform extracted from URLs associated with each subject. This included data such as number of Facebook likes or number of LinkedIn connections.

(3) Unstructured textual data, also automatically aggregated by the vendor's off-the-shelf platform extracted from URLs associated with each subject. The text data was used to develop two different types of features:

   (a) A set of sentiment score features that categorized chunks of text associated with the subject themselves, third parties interacting with the subject, and unattributed text, into 10 stratified "buckets" based on their relative positive or negative sentiment (a total of 30 features).

   (b) A "suicide risk score" calculated by using a separate logistic regression model based on n-grams from the text data.

APPENDIX D presents a full list of the vendor's selected demographic and social media metadata features.

DISCUSSION AND EXPERIMENTS TO IMPROVE RESULTS

After evaluating the original model, the authors explored a number of different options for improving the performance of the model. The following section is a discussion of those experiments.

Data Availability and Subject Focus

The principal issue with the implementation of the model is that only a minority of the subjects in the training dataset have free text and social media content. This presents a unique challenge since the goal of this effort was to develop a machine-learning based model using social media data to predict suicide. This gap between 1,400 nominal subjects with demographic data only and the 273 subjects with social media data is significant. In reality, the model rolls two different prediction efforts into one: predicting suicide using demographic data and predicting suicide using social media data. Rather than addressing it, the model uses simple imputation to generate straightforward, but low quality, predictions for the large quantity of missing data.

Accordingly, the authors focused its analysis on the performance of the model using only the subjects with social media data, rather than the full dataset. Because the present mission is to evaluate the use of social media to predict suicide before it occurs, it seems reasonable to apply it only to subjects for which social media content is available, and simply note the availability (or lack thereof) of these data. In terms of an operational approach, it would be one of a number of tools in a harm-reduction toolbox, and leadership would require other tools for subjects whose online presences which not be identified. To that end, the authors filtered these data to include only subjects that had free-text online content and evaluated the model's performance accordingly.

Demographic Data

The documentation accompanying the model described a semi-automated process to identify demographic variables correlated with suicide. In the interests of testability and reproducibility, use of automated feature selection methods to select machine-learning features is encouraged. Accordingly, the authors implemented automated feature selection using SciKit-Learn's chi-squared selection methods. The authors went back to the original demographic data for the panel that had been prepared for previous studies (Hesse, Bryan, & Rose, 2015) in order to insure that the program was aggregating the categorical data appropriately. Then the authors used an automated one-hot encoding technique (also known as dummy coding) (Harris & Harris, 2013) to transform the discrete demographic variables into binary indicator variables.

The vendor's handling of one variable in particular drew the authors' attention. The developers grouped the values in the field encoding DoD occupations to reduce the number of discrete values, but the groupings used seemed to have been generated on an ad-hoc basis without significant subject matter expert input or reference to standard DoD methods of aggregating occupation data. In order to aggregate more accurately the occupation data, the authors recoded the data using the Defense Manpower Data Center's (DMDC) standardized DoD occupation codes. These values are systematic grouped in a way that is cognizant of the particularities of the DoD.

Sentiment Scoring Technique

One of the two sets of text-based features generated from the social media is a sentiment analysis. The concept is that it is possible to discriminate between suicidal and non-suicidal subjects based on the positive and negative sentiments expressed by in their postings (and the postings of third parties) on their profiles. This underlying technique has support from academic literature. Unfortunately, as implemented, the sentiment analysis does not do a good job of extracting a measure of the sentiment of the content. The model implements sentiment analysis by examining individual fragments of content extracted from the various social media profiles (the size of a comment or status update) as the unit-of-analysis. The number and length of these can vary widely between subjects.

The technique used in the vendor implementation results in an indicator that correlates highly with the mean and does not pick up much of the variation in the content. For example, one negative word in a chunk containing otherwise neutral words presents as neutral obscuring the use of a negative word.

Because the model does not normalize these features, they become confounded with the quantity of material found in the search. While the quantity of text found may be an interesting feature to explore, quantity is not what this this variable purports to measure. A proper sentiment measure captures the intensity of the sentiment (positive or negative) regardless of the volume.

In order to address this deficiency with the model, the authors investigated a number of different sentiment analysis strategies. A technique that stood out as particularly relevant to this effort was the Valence Aware Dictionary and Sentiment Reasoner (VADER), which is a lexicon and rules based sentiment analysis algorithm specifically designed to for use with social media data (Huto & Gilbert, 2014). Instead of the chunks and bins approach used previously, the model now assigns each subject three intensity scores: a positive score, a negative score, and a neutral score, based on all of the free text found taken as a whole. The model then normalizes the scores, so they do not directly correlate with the quantity of text found for a given subject.

Other Architecture Improvements

In addition to the qualitative modifications, the authors also took some time to refactor and improve the maintainability of the Python code that builds and evaluates the model. The authors refactored the code to be compatible with the Scikit-Learn Pipeline and GridSearch modules. These modules allow automated testing of different combinations of input parameters for machine-learning models. This automated testing makes it easier to find the optimal parameters. Additionally, the authors revised the feature extraction and processing code to improve clarity and maintainability. For example, the functionality for handling the demographics, sentiment analysis, and keyword scoring were broken out into separate modules.

RESULTS

In the end, the authors evaluated the performance of four models: two based on the model developed by the vendor, and two developed internally. Table 1 presents a summary of the performance metrics for the different models evaluated. In addition to the recall metric discussed earlier, the table shows each model's accuracy, precision, and F_1-score. APPENDIX B presents a discussion of the statistical methods for evaluating this type of machine learning model that may provide additional context for interpreting these figures.

Table 1
Performance Metrics

	All Subjects	Social Media Subjects Only		
	Vendor	Vendor	Improved All Variables	Improved Text Variables
Accuracy	0.57	0.70	0.70	0.76
Precision	0.57	0.63	0.62	0.73
Recall	0.60	0.75	0.81	0.69
F_1 Score	0.58	0.69	0.70	0.71

The suicide prediction model developed by the social media vendor did a minimally adequate job at classifying subjects into suicide and non-suicide groups based on the input data. As implemented, the vendor's program reports performance statistics for the full 1,400-subject population. Table 2 presents the frequencies for that classification. In functionally the same condition the vendor delivered it in, the probability of the model correctly predicting if a subject is a member of the suicide group (the recall) is 0.59 (with 0.50 being the baseline for a random chance). While low in terms of desirable outcome for a machine-learning application, it did perform better than chance. Given that, the implementation has significant room for improvement.

Table 2
Vendor Implementation, All Subjects Confusion Matrix

N=1,400	Predicted Suicide	Predicted Non-Suicide
True Suicide	417	283
True Non-Suicide	319	381

Additionally, the authors evaluated the vendor implementation against the subset of subjects for whom social media data was available. As discussed previously, this provides a more focused approach geared towards using social media data for suicide prediction. In order to facilitate comparisons with the models developed by the authors, the same 20% validation subset the authors used when evaluating the other models. The metrics show improved performance, with recall of 0.75. Table 3 presents this model's classification decisions.

RESULTS

Table 3
Vendor Implementation, Social Media Subjects Confusion Matrix

N=37	Predicted Suicide	Predicted Non-Suicide
True Suicide	12	4
True Non-Suicide	7	14

After making the modifications to the model previously described, the authors were able to improve the performance of the model at the margin. Most of the improvement came from focusing only on subjects for which the social media vendor retrieved unstructured text data. Table 4 presents the classification results of the improved model. The recall performance increased to 0.81, the highest of the four models presented here.

Table 4
Improved with All Variables, Social Media Subjects Confusion Matrix

N=37	Predicted Suicide	Predicted Non-Suicide
True Suicide	13	3
True Non-Suicide	8	13

In addition to the primary improved model, the authors also developed a model that eliminated all of the demographic and metadata variables. This left only the variables based on the unstructured text: the sentiment scores and the suicide risk keyword score. The concept the researchers were attempting to validate was whether "less is more" with regard to this effort. While the "text only" model had a higher accuracy (0.76) than the other models, the recall performance (0.69) was less than the improved model with all variables. Given the focus on the recall statistic, the authors tentatively reject this hypothesis. Table 5 presents the results of the model's classifications.

Table 5
Improved with Text Variables Only, Social Media Subjects Confusion Matrix

N=37	Predicted Suicide	Predicted Non-Suicide
True Suicide	11	5
True Non-Suicide	4	17

DISCUSSION AND RECOMMENDATIONS

Overall, the authors conclude that this model for identifying Service members at risk for suicide is valid. While at this level of development, it is nowhere close to being able to identify all cases of death by suicide using social media data, it does identify roughly 70-80% of the cases when social media data are available for a subject. This is particularly notable given that it can operate in an automated fashion, without requiring costly human review of social media data to identify subjects who may be experiencing suicidal ideations. The result of this is that a model similar to the one evaluated here may be useful as a tool for identifying subjects for whom mental health care interventions would be helpful.

As discussed previously, one of the primary challenges facing this approach is the source data used for training the model. Machine-learning models work best when they have many data available for training. This project utilized a preselected panel of 1,400 subjects recycled from previous projects investigating suicide and social media. The vendor was only able to identify social media data for a subset of those subjects, and unstructured text for even fewer.

This process was made more difficult due to the time elapsed between the subjects' deaths, which occurred in 2010 and 2011, and the data collection, which took place in fall 2015. In the interim, there have been major changes to the social media landscape: some social media sites have redesigned and others that were once popular are not anymore. Furthermore, many social media sites used email addresses and/or phone numbers identify most accounts, which were not available for our subjects at the time of data collection.

The authors recommend that further work conducted in this area use a larger group of subjects in order to increase the chances of retrieval of social media content of interest. In conversations with the social media vendor, the authors discussed the subject of additional identifying information about the subjects. Making the association between a social networking profile and a real world identity can be difficult and the vendor suggested that they could achieve a better identification rate if subjects' email addresses and/or telephone numbers were available. Depending on the context of a potential implementation, it may also be possible to involve the subject in self-identifying their profiles.

Another aspect of maintaining a machine-learning based prediction tool is continually working to improve it. Some types of algorithms, such as neural networks have this capability built in, while others require manual intervention. If an organization implements a tool like this one in a pre-event environment, the users will not be able to verify the model's predictions (since the organization will intervene to deter suicidal behavior). This means that users cannot flag incorrectly labeled cases and feed them back into the model (the standard approach of continuous improvement of a machine-learning algorithm). Instead, the organization will periodically need to collect additional data on subjects who have

passed away and use these new data (in addition to historical data) to re-generate the model, improving the performance.

OPTIONS TO OPERATIONALIZE

There are a number of different directions the suicide prevention techniques presented here can be taken. The original concept was to implement a monitoring tool that would monitor individual subjects' social media presences for content similar to that, which precedes a suicide. This would allow for intervention with support services before the situation becomes unrecoverable. While this strategy is potentially powerful for reducing incidents of suicide, it raises a number of concerns about individual privacy and the desire not to "single out" or stigmatize individuals or their behavior within the community.

Another option for using this approach includes implementing it on an organizational level; instead of providing assessments on individuals, use social media data to look for indicators of suicide for a group of subjects that are part of an organization (such as a unit or office) and provide a "mental health barometer" for that group. This would provide leadership with useful intelligence into the readiness of their personnel and provide a signal to implement additional organization-wide interventions, if deemed appropriate.

This approach helps alleviate concerns about developing assessments of individuals and presets a lower barrier to organizational acceptance. Depending on the circumstances of implementation, an organization could also operate such a program on an opt-in basis. This approach has the added advantage of being able to involve subjects in self-identifying their social media accounts as a contribution to an effort to promote the well-being of their organization. The suicide-prevention research field has experience using this technique, through a project called Our Data Helps, which collects social media data from volunteers to help researchers learn more about the intersection between suicide and social media (Ruiz, 2016). Additionally, as social media platforms become more responsive to user concerns about privacy, obtaining consent and involving subjects in identifying their accounts is becoming more critical.

RECENT CHANGES TO SOCIAL MEDIA PLATFORM POLICIES

Since the social media vendor collected these data used for this project, there have been a number of changes in the social media privacy space that impact the methods used to collect social media data. These changes are a small part of a larger trend in society involving discussions of the privacy of data that people shared online. Government access to, and use of, these data is a particularly sensitive issue, even when aggregating open source data that are available to the public. Since data from Facebook and Twitter plays a key role in this project, and social media analytics in general, a discussion of some recent developments with these platforms follows.

Concerning Facebook, as a company they have recently made modifications to their product as well as become more aggressive with enforcement of their terms of service in ways that make it significantly more difficult to collect these types of data used in this project. In 2014, Facebook began encouraging users to be more mindful of what audiences they were sharing their social media content with, defaulting to "Friends Only" (Facebook, 2014). This severely limits these Facebook data available to the public and third parties, such as the social media vendor.

Additionally, Facebook had adopted a more aggressive stance in opposition to "scraping" data from their service. Many online services consider scraping to be a violation of their terms of service. Scraping involves using automated tools to browse and download (potentially large quantities of) webpages from a website. Automated tools then extract information from the downloaded web pages for further analysis. Scraping was the primary technique leveraged by the social media vendor to collect these data for this project. Since Facebook does not provide an Application Programing Interface (API) for third parties to access user data without user pre-authorization, scraping is the only method to gather data from Facebook for efforts similar to this without obtaining the subject's permission. If the implementation obtains the subject's authorization, official access to Facebook's APIs would be available and the data collection would be significantly more robust.

Twitter has always had a more open approach to its content than Facebook. While users do have the option of making their content private, many do not as the philosophy behind Twitter is more of a "one-to-many" publishing and sharing platform, rather than a place to connect with friends and family. As a result, it is much easier to access Twitter's public data, including access to official APIs that are much more robust than scraping-based solutions. At the same time, Twitter has recently given a less than warm response to government agencies interested in leveraging their data for analytics. Dataminer, a company that licenses Twitter's data for resale, recently had to discontinue working with a group of U.S. intelligence and law enforcement agencies on a counterterrorism-related project at Twitter's behest, after a series of negative reports in the media (Stewart & Maremont, 2016). While Twitter continues have promise as a robust data source, agencies should handle privacy and public relations considerations carefully going forward.

REFERENCES

Alpaydin, E. (2014). *Introduction to Machine Learning.* Boston: MIT Press.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *LREC*, *10*, pp. 2200-2204.

Bryan, C. J., Morrow, C. E., Anestis, M. D., & Joiner, T. E. (2010). A preliminary test of the interpersonal-psycological theory of suicidal behavior in a military sample. *Personality and Individual Differences, 48*(3), 347-350.

Burnap, P., Colombo, W., & Scourfield, J. (2015). Machine classification and analysis of suicide-related communication on Twitter. *Proceedings of the 26th ACM Conference on Hypertext & Social Media* (pp. 75-87). Association for Computing Machinery.

Facebook. (2014, May 22). *Making It Easier to Share With Who You Want.* Retrieved from Facebook Newsroom: http://newsroom.fb.com/news/2014/05/making-it-easier-to-share-with-who-you-want/

Gottschalk, L. A., & Gleser, G. C. (1969). *The measurement of psychological states through the content analysis of verbal behavior.* Univ of California Press.

Gottschalk, L. A., Winget, C. N., & Gleser, G. C. (1979). *Manual of Instructions for Using the Gottschalk-Gleser Content Analysis Scales: Anxiety, Hostility, Social Alienation-Personal Disorganization.* Univ of California Press.

Harris, D. M., & Harris, S. L. (2013). *Digital design and computer architecture* (2nd ed.). Waltham, Massachusetts: Morgan Kaufmann.

Hesse, C. M., Bryan, C., & Rose, A. E. (2015). *Indicators of Suicide Found on Social Networks: Phase 1.* Seaside: Defense Personnel and Security Research Center.

Huto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Eighth International AAAI Conference on Weblogs and Social Media.*

Lewis, R. J. (2000). An Introduction to Classification and Regression Tree (CART) Analysis. *Annual Meeting of The Society for Academic Emergency Medicine*, (pp. 1-14). San Francisco.

Loh, W. (2011). Classification and regression trees. *Data Mining and Knowledge Discovery, 1*, 14-23.

O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., & Christensen, H. (2015). Detecting suicidality on Twitter. *Internet Interventions, 2*, 183-188.

REFERENCES

Pestian, J., Nasrallah, H., Matykiewicz, P. Bennett, A., & Leenaars, A. (2010). Suicide note classification using natural language processing: A content analysis. *Biomedical Informatics Insights, 4,* 19-28.

Rose, E. A., & Hesse, C. M. (2015). *Indicators of Suicide Found on Social Networks: Phase 2.* Seaside: Defense Personnel and Security Research Center.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*(434), 473-489.

Ruiz, R. (2016, June 26). *Why scientists think your social media posts can help prevent suicide.* Retrieved from Mashable: http://mashable.com/2016/06/26/suicide-prevention-social-media/

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research, 8*(1), 3-15.

Stewart, C. S., & Maremont, M. (2016, May 8). *Twitter Bars Intelligence Agencies From Using Analytics Service.* Retrieved from The Wall Street Journal: http://www.wsj.com/articles/twitter-bars-intelligence-agencies-from-using-analytics-service-1462751682

Stone, P. J., Bales, R. F., Namenwirth, J. Z., & Ogilvie, D. M. (1962). The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science, 7*, 484-498.

Tausczik, Y.R., & Pennebaker, J.W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29,* 24-54.

Whiteley, P. J., & Rose, A. E. (2016). *Dying by Suicide: Indicators of Risk found on Social Networking Websites versus Self-reported Measures of Resilience.* Seaside: Defense Personnel and Security Research Center.

Wortman, J. A., & Shechter, O. G. (2016). *Suicide and Violent Cognitions, Emotions, and Behaviors in U.S. Military Personnel.* Seaside: Defense Personnel and Security Research Center.

APPENDIX A:

LABORATORY NOTES

The model is comprised of a set of Python scripts. The scripts process the data that the vendor captured and then trains the model using these data. The first portion of the process involved downloading data from the vendor's database and removing personally identifying information (PII). The authors were unable to run these steps in-house due to information assurance concerns and did not evaluate them beyond an inspection of the source code. The vendor provided output from those steps, which the authors used as the input of their analysis.

(1) Prepare Data

The first step in the process is to deal with missing values in these data. A separate Python class handles this. For each variable, the code imputes values differently depending on whether that variable is numeric or categorical. For numeric variables, the program uses the mean value of the variable for all missing values. For categorical variables, the program uses the most common value (mode) for all of the missing values.

Then the program converts all categorical variables to a pseudo-numeric type for use with machine-learning algorithms. For some "simple" categorical variables (gender, education, component, rank), the program manually replaces the values with integers. For the remaining categorical variables, the program uses a "dummy code" function supplied by the Pandas library to generate a numeric identifier for each category.

(2) Sentiment Scoring

Prepare the text for processing. Break up text in to blurbs (stored as pipe delimited). Blurbs are chunks of text scraped as a unit from the source material. Typically, the blurbs are the size of a single status update or comment. Tokenize each blurb and remove stop words and punctuation.

Get SentiWordNet synset for each word. The synset for each word is comprised of one or more "senses" of the word, each with a distinct positive and negative sentiment score. From these sentiment scores, calculate a composite word score over the set of senses by summing all of the positive sentiment scores, then subtracting all of the negative scores.

Then for each blurb, the program calculates a sentiment score for the blurb by taking the average of the word sentiment scores for the set of words in the blurb. Ten buckets are set up, one for each sentiment category (text/3rd party text/own text) and distribute the blurbs using a histogram procedure in those buckets based on value. Each bucket is hard coded to be "1.2 units" wide. The result is a set 30 of variables indicating the number of blurbs at a given sentiment level.

The result of this is three sets (one for each of the three text features, `text`, `3rdText`, and `ownText`)

(3) Keyword Score

The first step for generating the keyword score is to generate term-document frequency (TDF) matrixes for the three text features. A TDF matrix is a matrix with one row for each "document" (subject in this case) and one column for each term that occurs. The value is the frequency of each word in each document.

In order to select only the words that are statistically significant to our outcome, the TDF matrixes are binarized, so that rather than a representing term-frequencies, the matrixes represent the simple presence or absence of a term for a given subject. The terms with chi-squared scores (when $\chi^2 < 0.05$) are then identified.

The program then filters the TDF matrixes to include only the significant terms, retaining them for further processing during the model construction phase.

(4) Build Model

First, set up a stratified K-fold testing with ten folds. K-fold testing splits the data set into N different sets. The program builds model N times, each time with N-1 of the data sets used for training and one set used for testing. Over the N iterations, the program uses all N folds for testing. The program then compiles together the results from each fold for the final scoring.

The following procedure is repeated for each fold:

> Generate a single keyword score (for each text feature, three in total) using logistic regression. The program then trains a logistic regression model using the training set. The training set and tests sets predicted using the model.
>
> An instance of a random forest classifier is set up with the maximum number of features set to 36, the minimum sample split set to 36, and the number of estimators set to 50. The program fits and scores the model, retaining the predicted values for the test set for further evaluation.

(5) Score Model

Pool all the predicted results from the test sets together, calculate mean accuracy, and display a classification report (computes accuracy, recall, F-1 for the suicide and control groups). Calculate accuracy and recall statistics for a number of sub-populations (subjects with text, twitter, and any domain data).

Re-run the logistic regression and random forest classifier on the full set of data (holding none out for testing) and calculate the model score. Extract the importance of each feature from the classifier and output them as a list.

APPENDIX B:

STATISTICS FOR EVALUATING PREDICTIVE MODELS

Evaluating performance is essential to understanding the predictive power and usefulness of any given predictive model. In order to evaluate a machine-learning model properly, one must set aside a portion of the sample data with known true conditions to use for evaluation (with the balance used for training).

The type of model discussed in this report is a binary classifier. The model predicts which cases belong to one of the two possible classifications. The general case uses the terms 'positive' and 'negative' to refer to these classifications.

Researchers evaluate the performance of a model of this type by comparing the known true classifications of the validation set with predicted classifications generated by the model for that same set. One then tabulates frequencies for the four possible combinations of true vs predicted condition and positive vs negative condition. This allows one to analyze how often a model agrees or disagrees with the reality of the test set.

Figure B-1 outlines the relationship between these frequencies.

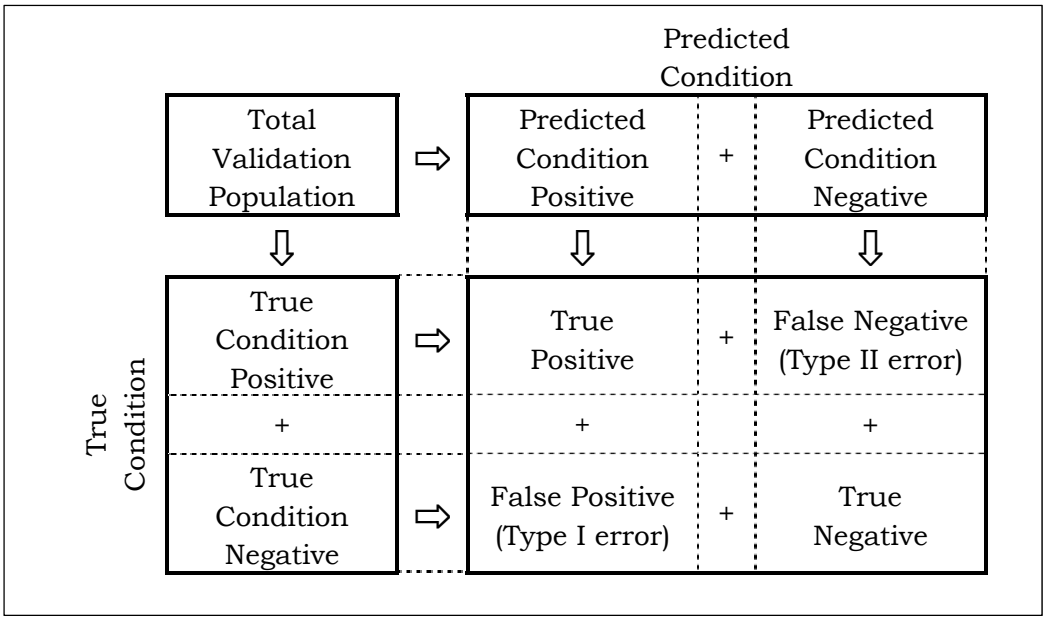

Figure B-1 Confusion Matrix Diagram

One can calculate number of metrics to assess the performance of a model. Experts commonly use three different statistics to quantify the results (O'Dea, et al., 2015).

RECALL

$$TPR = Recall = \frac{True\ Positive}{Condition\ Positive}$$

Recall, also referred to as sensitivity or as the True Positive Rate (TPR), expresses the ratio of true positive cases (those that the model and test set agree are condition positive) and the total true condition positive cases. This metric answers the question, "What percentage of the positive cases did the model correctly classify as

positive?" This is a good metric to focus on when the requirement is to maximize the number of positive cases predicted, without regard to other factors. It is effective when the cost of missing a positive case is high, and the cost of applying an intervention to false positive cases is low.

PRECISION

$$PPV = Precision = \frac{True\ Positive}{Predicted\ Positive}$$

Precision is the other primary metric used to assess classification models. Also known as positive predictive value (PPV), precision answers the question "What percentage of the predicted positive cases are actually positive cases?" One can look at this as an assessment of the "quality" of a model's positive predictions, (i.e., the rate of correct predictions). This measure is more appropriate in a situation where the goal is to make sure an intervention is applied to the smallest number of out-of-class subjects as possible. This would be appropriate in a situation where the intervention proposed is costly, and there is a desire not to apply it to out-of-class cases even if that means missing some positive cases.

F-SCORE

The F-score or F-measure is a metric that incorporates both recall and precision. Mathematically, it is the weighted harmonic mean of the precision and recall statistics. The harmonic mean is a type of average, similar to the arithmetic or geometric mean. One of its uses is to calculate averages of rates, as is the case here.

$$F_\beta = (1 + \beta^2) \times \frac{precision \times recall}{(\beta^2 \times precision) + recall}$$

By adjusting β, the one can target different levels of emphasis between precision and recall. A special, commonly used case of the F-score is F_1, where $\beta = 1$. This metric represents a balanced emphasis on both precision and recall.

APPENDIX C:

DATA SET DEMOGRAPHICS AND DESCRIPTIVE STATISTICS

The cohort consists of *N*=1,400 Service members, *n*=700 (50%) that died by suicide and *n*=700 (50%) that died by other means. The overall cohort included females with ages that ranged from 17-59, mean of 29.67 (9.922). Male ages ranged from 17-80 with a mean of 30.01 (10.176). In terms of rank, females tended to be mostly junior enlisted or non-commissioned officers with a mean score of 0.64 (0.780). Males also tended to be junior enlisted with a mean of 0.65 (0.839). With respect to education, females had a mean score of 1.65 (1.242) which means they generally had some college exposure but did not complete their education. Females were, however, slightly more educated than their male counterparts who had a mean score of 1.59 (1.268). Finally, within the entire cohort there were only 26 (9.5%) females who had any text or social media data, and 247 (90.5%) males who had any text or social media data.

The authors conducted independent samples T-tests on those who had any social media data vs. those who did not have any social media data. The comparison consisted of age, sex, rank, education, marital status, ethnicity, faith, and cause of death. The results of this test concludes that only *education* (T-test: *p*=.04, *t*=-2.921, *df* 1,395. Levene's test for equality of variances: *f*=25.231 *p*= .000) and *cause of death* (T-test: *p*=.086. *t*=1.718, *df*=1,395. Levene's test for equality of variances: *f*=7.921, *p*=.005) were statistically significant between the two groups. Individuals with social media data had a mean of 1.79 (1.444). These individuals were more educated that those without social media data as evidence of the mean score of 1.54 (1.214). Individuals who died by suicide had a mean of 0.45 (0.499), were less likely to have social media accounts, than those who died by other means mean of 0.51 (0.5).

Table C-1
Gender

	Suicide	Non-suicide
Male	656 (94.1%)	644 (92%)
Female	41 (5.9%)	56 (8%)

Table C-2
Marital Status

	Suicide	Non-suicide
Never married	320 (45.9%)	331 (47.3%)
Married	332 (47.6%)	303 (43.3%)
Divorced	37 (5.3%)	52 (7.4%)
Legally separated	5 (0.7%)	2 (0.3%)
Widowed	1 (0.1%)	1 (0.1%)
Unknown	2 (0.3%)	11 (1.6%)

Table C-3
Dependents

	Suicide	Non-suicide
0	327 (46.9%)	328 (46.9%)
1	129 (18.5%)	133 (19%)
2	93 (13.3%)	92 (13.1%)
3	73 (10.5%)	83 (11.9%)
4 or more	62 (8.9%)	49 (7.1%)
Unknown	13 (1.9%)	15 (2.1%)

Table C-4
Race

	Suicide	Non-suicide
Caucasian	318 (45.6%)	301 (43%)
Black or African-American	36 (5.2%)	74 (10.6%)
American Indian/Alaskan Native	10 (1.4%)	13 (1.9%)
Asian	6 (0.9%)	8 (1.1%)
Unknown or Other	327 (46.9%)	304 (43.4%)

Table C-5
Education

	Suicide	Non-suicide
High School or Equivalent	516 (74%)	505 (72.1%)
Bachelor's Degree	44 (6.3%)	76 (10.9%)
Some College	51 (7.3%)	43 (6.1%)
Associates Degree	34 (4.9%)	29 (4.1%)
Less than High School Education	16 (2.3%)	20 (2.9%)
Post-Graduate or Professional Degree	19 (2.7%)	16 (2.3%)
Unknown	17 (2.4%)	11 (1.6%)

Table C-6
Religious Affiliation

	Suicide	Non-suicide
Christian	374 (53.7%)	436 (62.3%)
Atheistic, Agnostic, or No Religious Preference	218 (31.3%)	165 (23.6%)
Buddhist	5 (0.7%)	3 (0.4%)
Jewish	3 (0.4%)	1 (0.1%)
Wicca or Witchcraft	1 (0.1%)	0 (0%)
Other Eastern Religion	0 (0%)	1 (0.1%)
Unclassified or Unknown	96 (13.8%)	94 (13.4%)

Table C-1 through Table C-6 display the basic differences between the two groups, which include gender, marital status, and the number of dependents, race, education, and religious affiliation. The overwhelming majority of individuals in the non-suicide group consist of men *n*=644 (92%), that were never married *n*=331 (47.3%) or married *n*=303, (43.3%), had no dependents *n*=328 (46.9%), were Caucasian *n*=301 (43%) or were of Unknown race *n*=304 (43.4%), had at least a high school diploma or equivalent *n*=505 (72.1%), and identified themselves as Christian *n*=436 (62.3%). For Service members that died by suicide, the characteristics are similar: Service members were mostly men *n*=656 (94.1%), never married *n*=320 (45.9%) or married *n*=332 (47.6%), had no dependents *n*=327 (46.9%), were either Caucasian *n*=318 (45.6%) or of Unknown race *n*=327 (46.9%), had at least a high school diploma or equivalent *n*=516 (74%), and were Christian *n*=374 (53.7%).

Table C-7
Military Component

	Suicide	Non-suicide
Regular	318 (45.6%)	282 (40.3%)
Reserves	219 (31.4%)	234 (33.4%)
Guard	160 (23%)	184 (26.3%)

Table C-8
Military Branch

	Suicide	Non-suicide
Army	425 (61%)	418 (59.7%)
Air Force	112 (16.1%)	95 (13.6%)
Marine Corps	67 (9.6%)	88 (12.6%)
Navy	82 (11.8%)	84 (12%)
Coast Guard	11 (1.6%)	14 (2%)
Public Health	0 (0%)	1 (0.1%)

Table C-9
Service Rank

	Suicide	Non-suicide
Junior Enlisted	370 (53.1%)	354 (50.6%)
NCO	255 (36.6%)	253 (36.1%)
Officer	48 (6.9%)	64 (9.1%)
Senior Enlisted	19 (2.7%)	15 (2.1%)
Warrant Officer	5 (0.7%)	14 (2%)

Table C-7 through Table C-9 elaborate further on what military-specific demographic information each group consists of: service component, branch, and rank. For the non-suicide group regular Service members *n*=282 (40.3%), Army

n=418 (59.7%), and junior enlisted personnel *n*=354 (50.6%) composed the greatest proportion of individuals studied. For the suicide group regular Service members *n*=318 (45.6%), Army *n*=425 (61%), and junior enlisted personnel *n*=370 (53.1%) also composed the greatest proportion of individuals studied.

Table C-10
Gender

	Has Social Media Data	No Social Media Data
Male	247 (90.5%)	1,056 (93.7%)
Female	26 (9.5%)	71 (6.3%)

Table C-11
Military Branch

	Has Social Media Data	No Social Media Data
Army	166 (61%)	667 (60.1%)
Air Force	40 (15%)	167 (14.8%)
Navy	36 (13%)	130 (11.5%)
Marine Corps	26 (9.5%)	129 (11.5%)
Coast Guard	4 (1.5%)	21 (2%)
Public Health	0 (0%)	1 (0.1%)

Table C-12
Service Rank

	Has Social Media Data	No Social Media Data
Junior Enlisted	147 (54%)	577 (51.3%)
NCO	90 (33.1%)	418 (37.2%)
Officer	28 (10.3%)	84 (7.5%)
Senior Enlisted	2 (0.7%)	32 (2.8%)
Warrant Officer	5 (1.8%)	14 (1.2%)

Table C-10 through Table C-12 illustrates the differences between Service members who either had any social media data vs. did not have any social media data. There were *n*=1,127 (80.5%) individuals with no social media data and *n*=273 (19.5%) with any social media data. Individuals with no social media data found were men *n*=1,056 (93.7%) in the Army *n*=667 (60.1%) that were junior enlisted *n*=577 (51.3%). For Service members that did have any social media data: men *n*=247 (90.5%), in the Army *n*=166 (61%), in the junior enlisted ranks *n*=147 (54%) represented the greatest amount of those studied.

APPENDIX D:

MODEL FEATURES

The following is a list of the demographic and social media metadata features used in the model as implemented by the vendor.

Table D-1
Vendor Demographic Features

Discrete Fields	Distinct Values	Variable Name
Religion	Catholic/Christian, no religious preference, other, unknown	D_Religion
Marital Status	Married, unmarried, unknown	D_Married
Component	Reserve/Guard, Regular	D_Component
Ethnicity	Hispanic, Asian/Austronesian, U.S. or Canadian Tribes, None, Other, unknown	D_Ethnicity
Rank	Enlisted, warrant officer, officer	D_Rank
Gender	Male, Female	D_Gender
Education Level	Pre-high school, high school, college, grad school, unknown	D_Education
Primary Service Occupation	Admin, Support, Law, Combat, Mechanic, Tech, Medical, Engr, Transport, Other, Unknown	D_PSO_cat
Service Branch	Air Force, Coast Guard, Navy, Army, Marine Corps, Public Health Services	D_Serv
Age	Whole numbers	D_Age
Years in Service	Rounded to the nearest tenth	D_YearsInService
Age at Enlistment	Rounded to the nearest tenth	D_AgeAtEnlistment

Table D-2
Vendor URL Metadata Features

Feature	Source	Values	Variable Name
Number of Facebook Friends	Subject's Facebook profile	Integer	numFBFriends
Number of Facebook Likes	Subject's Facebook profile	Integer	numFBLikes
Number of Tweets and Retweets	Subject's Twitter profile	Integer	T_twts
Number of Twitter followers	Subject's Twitter profile	Integer	T_flwrs
Number of favorites on Twitter	Subject's Twitter profile	Integer	T_favs
Number of LinkedIn Connections	Subject's LinkedIn profile	Integer, 500 + connections is 1 bin	LI_conn
Number of LinkedIn profile elements	Subject's LinkedIn profile	Integer	LI_num
Memberships on various domains	Automated PAEI collection	binary, 1 if subject is a member, 0 otherwise	d_*
Buying interests	Automated PAEI collection	binary, 1 if subject has this interest, 0 otherwise	int_*
Domain appearance	Automated PAEI collection	binary, 1 if subject was found on this domain, 0 otherwise	d_*
Number of highly rated (confirmed) subject URLs	Automated PAEI collection	Integer	numHighs