The Role of Web Analytics in Mental Health: A Correlation Study of Search Engine Interests and Suicidal Behaviors

by

Jessica M. Dzurek

University of Pittsburgh

Bachelor of Science in Computer Science, 2011

Thesis submitted to the faculty of

the Bouvé College of Health Sciences

in partial fulfillment of requirements for the degree of

MASTER OF SCIENCE

In

Health Informatics

Northeastern University

August 2013

ABSTRACT

Mental illness is a growing concern within the United States as the rate of depression and suicide continue to increase annually. Information technology also continues to grow and has begun to be recognized as a useful tool to battle some of these public health challenges. This study examines the potential for leveraging web analytics as an additional resource for mental health data capture, while providing a new perspective of mental health in the population. Similarly, search engine trending is used for opinion-mining for industries such as consumer goods and marketing, this could also be leveraged for mental illness to compliment traditional techniques. This study presents work on using correlation, regression, and trend analysis to illustrate significant relationships between search engine search-term trending and nationally defined statistics such as suicide rates, mental health system utilization, and general well-being. The results of these analyses confirmed that a measureable level of correlation existed within every dataset. Correlations between the suicide rate statistics provided by the Centers of Disease Control and Prevention (CDC) and the Google Trend analysis of selected search-terms were of the highest in comparison the other datasets.

TABLE OF CONTENTS

Chapter 1. INTRODUCTION

1.1 Motivation

Modern medicine in the United States has made great strides to advance the knowledge, detection and treatment of mental illness in recent years; however, the concerns continue to grow as the statistics for mental illness and suicide rates continue to steadily climb. The advent of Internet technologies such as patient user forums, government resources, and social networking have aided in prevention through knowledge and community outreach, as well as providing additional data capture methods. Organizations ranging from non-profit to privately-owned have recognized these advantages and are beginning to expand research into different social and Internet resources.

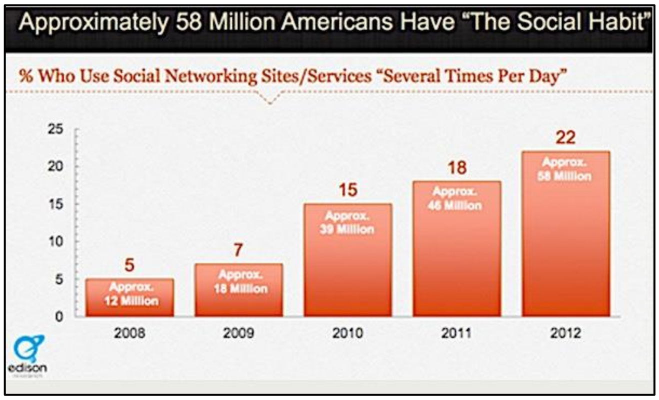
1.2 Background

Suicide has been around as long as man has been in existence. Suicide rates within the United States have steadily climbed and since 2000 have jumped to the tenth leading cause of death (CDC). It has been estimated that an American attempts suicide every 41 seconds, while every 16.7 minutes an American is successful in their attempt. Suicide has even exceeded the rate of homicide(s) in the United States. Realizing the impact of suicide on the nation, federal agencies such as the National Institute of Mental Health, Centers for Disease Control and Prevention, Substance Abuse and Mental Services Administration along with other organizations have been established to not only provide prevention and treatment, but also to understand the issue through continuous research *(Shalin, 2004).*

With the massive increase of Internet and social networking usage over the last decade in the United States, we have seen how useful these tools can be used in events leading up to suicide. There are pros and cons associated with the use of these tools and mental health. For instance, the Internet has played an important role in providing knowledge, as well as aiding in prevention and community outreach for mental illnesses such as suicide and depression. However, increased usage and expansion of social media services has also provided a place of interaction where users can be bullied and harassed, commonly referred to as Cyber-bullying. This has on occasion led to suicidal behavior.

Chapter 2. RELATED WORK

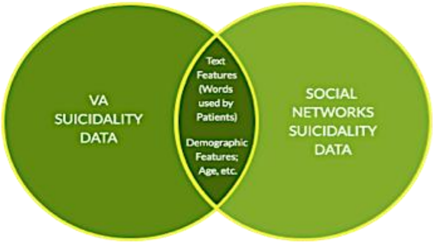
The use of social media has increased rapidly since 2010 due to the popularity of services such as MySpace, Facebook, Twitter and Google+ (figure 1.0) *(Baer, 2012).* This initially sparked the interest of retail, marketing, and political industries, but has since spread into the medical arena. Initially this data was used to monitor the spread of illness with the development of tools such as Google Flu Trends. The success of these tools and this new interest has stressed the importance of funding research focused on leveraging this data for mental health. Some featured projects and studies that dive into this realm are the Durkheim Project, which studies the use of social media to assess suicide risk among a population and the Microsoft Research Lab's "Social Media as a Measurement Tool of Depression in Populations" study.



(Figure 1 - American's having personal profiles on any Social Networking Website growth)

Google Flu trends were developed to monitor health behaviors and predict flu infections based on online search queries. The study was able to establish a high correlation between the number of people who searched for flu-related topics and the number of reported flu cases. The data derived from these search queries were compared to traditional flu surveillance systems with the finding of an overlap in the time period between the spikes in both datasets. Google Flu Trends provide data allowing estimations of flu patterns by counting these flu-related search queries in different locations around the world. This successful tool has opened up the online information floodgates for further investigation in the use of similar datasets for mental illness *(“Google Trends”)*.

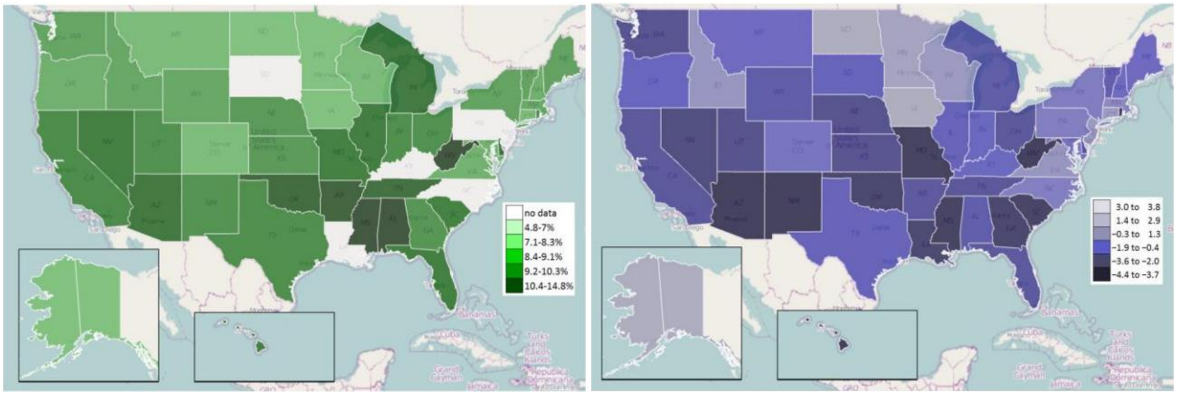
The Durkheim Project is a multidisciplinary research project, which performs analysis from text messages and social media through an opt-in program. Participants are asked to allow the monitoring of their social media activity to contribute to the project. Through the use of natural language processing (NLP) techniques, the team looks for words and phrases that might identify suicidal risk and negative behavior (Status Updates and Suicide Risk). While the project did not fully get up and running until 2012, prior research from team members around the subject area had begun in 2011. The initial investigation results indicated a statistical correlation between mental illness and text-mining methods for predicting suicidality (figure 2) *(""The Durkheim Project" Will Analyze Opt-In Data From Veterans' Social Media And Mobile Content -- Seeking Real-Time Predictive Analytics for Suicide Risk," 2012).*



(Figure 2 – Durkheim Real-time Classifier)

This discovery set the foundation for the current Durkheim Project which focuses on continued progression towards the goal of providing "continuous monitoring of social network user behavioral intent enabling intervention, facilitated by social/online/mobile data sources" *("Our Project – Durkheim Project", n.d.).*

Microsoft's research program has also contributed conducting studies surrounding mental illness and social media and publishing various articles that are available publically. One such article entitled "Social Media as a Measurement Tool of Depression in Populations" discusses the usage of social media as an additional tool to accompany traditional surveying methods for understanding and detecting depression within a population. This research project saw the potential use of social media as an additional data source to provide a new level of granularity overtime and allow for larger population sample sizes. Twitter usage was used as a data source for this study by developing a corpus of 'Tweets' via a crowdsourcing methodology. A probabilistic model was used to flag if posts had an indication of depression based on emotion, activity, and language. The team was then able to create a depression indexing system based on this analysis, allowing them to expand the initial analysis to encompass demographic information and to create a predictive model. This model was validated and the predictions produced by the model were fairly accurate. A correlation was also proven between these findings and national depression statistics reported by the Centers for Disease Control and Prevention (CDC). The results of this correlation can be seen in the heap maps provided of the actual CDC data and the predicted (SMDI) depression level in *figure 3*. These maps illustrate the strong correlation between the actual depression percentage and the predicted SMDI depression level for each state. The higher the color intensity the higher the reported or predicted depression level. The correlation yielded a positive correlation of 0.5073 with a linear regression between the actual and predicted values *(De Choudhury, Counts, & Horvitz, 2013)*.



(Figure 3 – Heap map of CDC data (left) and predicted values (right))

Chapter 3. RESEARCH OBJECTIVES

3.1 Rationale for the Study

The contribution of studies such as those mentioned to this area of interest have encouraged continued exploration in the roles that social media text-mining and web analytics can play for mental illness prevention, detection and monitoring. A plethora of recent research focuses on sentiment analysis and the text-mining of social media services such as Facebook or Twitter. Data sources captured by online search engines such as Google have had far less attention in regards to mental health. The topic of online search engine usage has been selected for this study because search engine usage is one of the top ways people spend their time online as described in a 2012 study conducted by Go-Gulf.com. This study showed that while social networking still occupies the most time spent online, with normal Internet users spending approximately 22% of their time on social networking sites, search engine usage is in close second with a 21% usage rate *("How People Spend Their Time Online [Infographic]," 2012)*. In addition, this study focused on the time periods of 2008 to 2010 and 2008 to 2011, during which time there was a large growth of search engine activity, with Google experiencing a 108% usage growth *(Google Annual Search Statistics, 2013).*

While there has been a focus on the relationship of Google search volumes and health, little has been made available to the public that specifically focuses on mental health. This study hopes to contribute to this area by evaluating these search volumes with mental health statistics. Suicidal behavior reflects a certain level of health of a society in general; therefore while this study aims to provide a general view, the primary emphasis is with suicidal behavior.

3.2 Goals and Objectives

This study was performed with two goals in mind. The first goal was to determine if a relationship exists between search engine interest trends and suicidal behavior. Focus is given to answer various questions related to this goal: Is there a pattern between any change in Google Search Interest Volume and suicide rates over time? What relationships exist between these search interest and suicide, mental health services demand, or general well-being?

The second goal was to determine the plausibility of leveraging search engine inquiries and behaviors to monitor, detect, and provide intervention for depression and suicidal behaviors. If relationships can be determined between these datasets, what are the implications of this new found method of mental health data capture and insight?

Chapter 4. RESEARCH METHODOLOGY

4.1 Correlational Explanatory Design

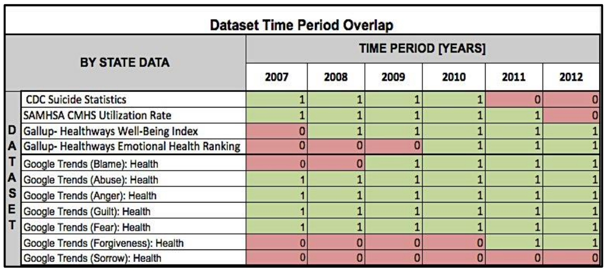
The research design used for the data collection of the various datasets adopts the correlational explanatory design methodology. The correlational research methodology observes the values of two or more variables to evaluate if a relationship exists and its strength. The main objective for correlational research is to not only to determine if a relationship exists, but also establish the direction, magnitude and forms of the observed relationships *(Cherry, Mohammad, & Brujin, 2012)*. However, it is important to note that correlation does not equal or imply causation. In other words, a strong correlation does not necessarily mean that one variable causes the other. The explanatory design simply tries to identify the association between the selected variables and digs deeper to see the extent of the variables' relationship.

4.2 Variables

Many obstacles were encountered when selecting and collecting online data sources. The initial focus was on collecting Twitter data (tweets) to conduct various sentiment analysis techniques to calculate an aggregate positive, negative, or neutral sentiment of a given region during a given time period. This proved to be difficult due to several factors that were out of the control of this study. The first roadblock was the usage and gathering limitations implemented by the API provided by Twitter, which bestowed a limit to the amount of queries and data collected by a user. Many researchers and software developers alike have provided various work-around solutions for this roadblock, however the stability of these solutions is not consistent.

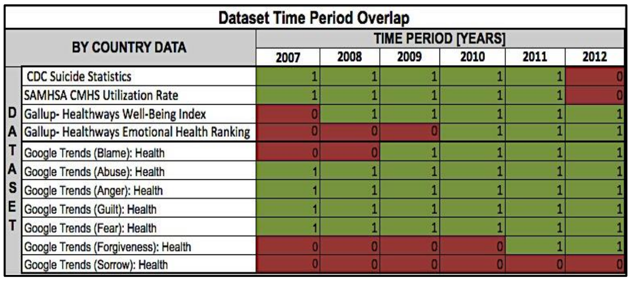
The next hurtle encountered was the lack of historically relevant data from Twitter. The implementation of GeoLocation functionality was not introduced until later in 2010, which would result in only two full years of data for analysis. One suggested strategy, as implemented in the Microsoft Research Lab's study, was to label the location of a tweet based on the location defined by the user in their personal profile. After further investigation into this work-around method to capture this dataset, it became apparent that the work required to implement a stable solution would be limited due to the resources and time constraints associated with this study.

One last challenge presented itself while establishing candidate resources. Due to the variety of sources, it was difficult to ensure overlap existed between available time periods for each dataset. It became apparent that there was a commonality between datasets of the same variable type and time period availability. The online data sources candidates were more plentiful in recent data than older data, while the defined statistic data source candidates exhibited a delay in more recent data, but supplied a reasonable amount of historical data (*figure 4*).



Dataset Time Period Overlap

BY STATE DATA		TIME PERIOD [YEARS]					
		2007	2008	2009	2010	2011	2012
DATASET	CDC Suicide Statistics	1	1	1	1	0	0
	SAMHSA CMHS Utilization Rate	1	1	1	1	1	0
	Gallup- Healthways Well-Being Index	0	1	1	1	1	1
	Gallup- Healthways Emotional Health Ranking	0	0	0	1	1	1
	Google Trends (Blame): Health	0	0	1	1	1	1
	Google Trends (Abuse): Health	1	1	1	1	1	1
	Google Trends (Anger): Health	1	1	1	1	1	1
	Google Trends (Guilt): Health	1	1	1	1	1	1
	Google Trends (Fear): Health	1	1	1	1	1	1
	Google Trends (Forgiveness): Health	0	0	0	0	1	1
	Google Trends (Sorrow): Health	0	0	0	0	0	0



Dataset Time Period Overlap

BY COUNTRY DATA		TIME PERIOD [YEARS]					
		2007	2008	2009	2010	2011	2012
DATASET	CDC Suicide Statistics	1	1	1	1	1	0
	SAMHSA CMHS Utilization Rate	1	1	1	1	1	0
	Gallup- Healthways Well-Being Index	0	1	1	1	1	1
	Gallup- Healthways Emotional Health Ranking	0	0	0	1	1	1
	Google Trends (Blame): Health	0	0	1	1	1	1
	Google Trends (Abuse): Health	1	1	1	1	1	1
	Google Trends (Anger): Health	1	1	1	1	1	1
	Google Trends (Guilt): Health	1	1	1	1	1	1
	Google Trends (Fear): Health	1	1	1	1	1	1
	Google Trends (Forgiveness): Health	0	0	0	0	1	1
	Google Trends (Sorrow): Health	0	0	0	0	0	0

(Figure 4 – Dataset Availability by Time Period)

4.2.1 Independent Variables

With the realization of these struggles, a search began for another online data source that would provide stability and data that could be transformed into information to later provide insight, while working within the time and resource constraints. With the idea of Google Flu Trends and the steady climb of online search activity in recent years in mind, it was determined that a dataset produced through the use of Google Trend could be used and would meet the needs of this study. Google Trends would supply a dataset of search volume interest (SVI), an adjusted-calculated value, which spanned across location and time. This represents the amount of searches completed for specific search-terms (up to five) for a particular time, location, and category relative to the total number of searches preformed on Google. The value assigned has been normalized and scaled to values 0 to 100. For this research the time interval filter was years (2007-2012), while the location filter was set to United States and was represented in two levels of granularity, national-level (averaged) and state-level.

In order to ensure data relevancy of the Google Trends dataset negative affect terms were selected and used in the data collection process. The selected terms were derived from a list created during the 2011 research project “Binary Classifiers and Latent Sequence Models for Emotion Detection in Suicide Notes” by the National Research Council of Canada. This project analyzed each sentence of a sample of suicide notes by first annotating the sentences and applying a large-margin model to create an emotion classification, then applying a latent sequence model that determines the sentiment of a sentence and labels emotional regions. Out of the fifteen labels/terms used within the study seven terms were select based on negative emotion association and micro-averaged F-measure calculated value. The candidate terms selected for the Google Trends analysis were ‘Abuse’, ‘Anger’, ‘Blame’, ‘Fear’, ‘Forgiveness’, ‘Guilt’ and ‘Sorrow’.

4.2.2 Dependent Variables

Dependent variables were obtained to analyze national vital death statistics, the demand for mental health services, and the cumulative state of ‘well-being’ across the United States.

Since the primary focus of this research study focuses upon suicidal behavior, a dataset of the prevalence of intentional self-inflicted (suicide) cause of death was collected. This dataset consisted of State suicide rates for 2007 to 2010[1] and national

[1] *2011 Suicide Statistics for the United States by State do not currently exist at the time of this study.*

rates for 2007-2011. Rates were selected over provided counts of deaths to provide an accurate comparison, independent of a State's population.

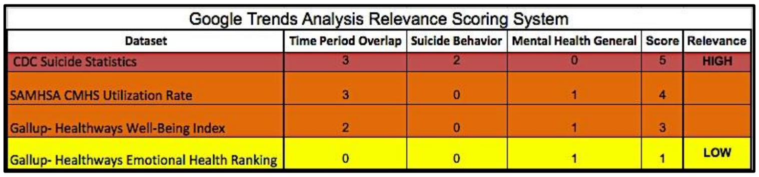
In an attempt to represent the usage of mental health services, data was collected from a series of system output table reports provided by the Substance Abuse and Mental Health Services Administration's Center for Mental Health Services (CMHS). These reports are based on the CMHS Unified Reporting System (URS) data, which is provided by the United States per annual reporting guidelines. For this study a dataset was created from the reports, which contained each State's utilization rate per 1,000 Population of clients Served by the SMHA System for 2007 to 2011.

To represent the general state or well-being of individuals within the United States or on an individual State level, the Gallup-Healthways' calculated Well-Being Index was represented in this study. This measure is calculated by surveying no less than 500 individuals daily covering six separate domains: life evaluation, emotional health, physical health, healthy behavior, work environment, and basic access. The use of these domains within the survey process aims to paint a well-rounded picture of the well-being of an individual. Each survey asks several questions within each domain and the interviewee selects a value of 0 to 10. Low values indicate 'suffering', mid values represent 'struggling', and high values show an individual is 'thriving'. The Well-Being index is then calculated by averaging the scores of each domain. This data was available for 2008-2012 on both a national and state level; however the individual domain indexes were less common for each year and level.

4.3 Procedure

4.3.1 Data Collection and Processing

This study took an iterative agile approach for capturing, organizing, and analyzing each dataset. After establishing the goals and questions to be answered within this study, each candidate data source was evaluated for relevancy and prioritized accordingly. Each dataset candidate was evaluated against three categories: time period overlap, suicide behavior and general mental health relevance. The datasets was scored ‘0’ to ‘3’ in each category and summed to establish a scoring system for prioritization (refer to figure 5).



Google Trends Analysis Relevance Scoring System

Dataset	Time Period Overlap	Suicide Behavior	Mental Health General	Score	Relevance
CDC Suicide Statistics	3	2	0	5	HIGH
SAMHSA CMHS Utilization Rate	3	0	1	4	
Gallup- Healthways Well-Being Index	2	0	1	3	
Gallup- Healthways Emotional Health Ranking	0	0	1	1	LOW

(Figure 5- Dataset Candidate Relevancy to Google Trends Analysis Scoring Results)

For ease of use and organizational purposes each dataset was transcribed into a single Microsoft Excel worksheet within a Microsoft Excel workbook representing all the datasets used in this study, with one exception of each search-term for the Google Trend analysis dataset was assigned an individual worksheet. One worksheet was also created to house all the national-level data for each dataset in one location. Each dependent variable was inputted directly into its appropriate worksheet with no manipulation or transformations of the data.

The independent variables, each selected-term’s Google Trend analysis, were copied into the proper worksheet for the state-level data results; however the national-level data was not available from the result sets, therefore needed calculated. To collect the Google Trends data, a search-term is entered into the tool with the Date filter set to a single year (2007-2012), the Location filter was set to ‘United States’, and the Category filter set to ‘Health’. The Category filter was selected to omit irrelevant data from the result set. For example, without the ‘Health’ filter applied, when generating a result set for ‘Blame’, there was a large spike in interest in 2009. After further investigation of plausible searches and causes for this spike it was discovered a popular song “Blame it” by Jamie Foxx was released in 2009, which led to many searches for music videos, mp3 downloads, and lyrics. The result set was then exported in ‘CSV’ format and State (sub-region) values were parsed out and placed into their proper worksheet. However, to obtain the national-level values a calculated-field was created to average the “weekly” search interest volume index presented in the result set for a given year.

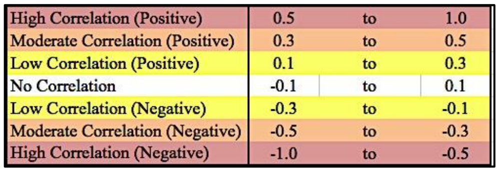
After the initial data collection process, each dataset was evaluated again for relevancy and completeness. Incomplete or irrelevant datasets or fields remained in overall workbook, but were eliminated from the analysis process. A total of five datasets or fields were eliminated: Google Trends (Forgiveness), Google Trends (Sorrow), SAMHSA Output Tables (Community Utilization), SAMHSA Output Tables (State Hospital Utilization), and Gallup-Healthways Emotional Health Ranking. These values were eliminated from the analysis set due to sparse data or weak relevancy to the study’s objectives and goals. In an attempt to prepare for the analysis stage, the average for each dataset in various overlapping time ranges (2007-2011, 2008-2010, & 2008-2011) and were calculated into the proper worksheets.

4.3.2 Analytic Process and Techniques

Various statistical analysis techniques can be implemented to test the relationship between two or more variables. However, the type and representation of the variables determines which statistical test best fits your study. It is also important to note that if two variables covary they tend to be related, although that does not always represent a causal relationship. For the purposes of this study the object is to test for patterns and relationships between each dataset.

Since the final datasets consist of values that are interval or ratio in nature and the first objective is to test the relationships between the independent and dependent variables, the Pearson Product-Moment Correlation (PPMC) test is a good candidate. The PPMC calculates a correlation coefficient (r), which measures the strength of a linear relationship between two variables, in this case a given dependent, and independent variable. The r-value will also provide the direction (negative or positive) of the relationship, should one exist. To perform this analysis a matrix was created with each dataset represented as a row and a column, with the coefficient value for each coordinate stored. Fortunately, Microsoft Excel has a function available for the Pearson correlation test. This function takes an array of two variables as parameters and returns the Pearson correlation coefficient (r) as the result. To test the relationships on a state-level, an average was calculated for the time range of 2008-2010 for each Google Trend search-term and used as the first array parameter. The second array parameter was the calculated average for 2008-2010 of each dependent variable: CDC provided Suicide Rate, SAMHSA Total Utilization Rate, and Gallup-Healthways' Well-Being Index. This process was slightly modified to test the relationships on a national-level. Since the

Google Trend national values already represented averages, the first parameter array contained the Google Trend yearly averages for a given term, while the second parameter array contained the yearly national values for each dependent variable. Next, the coefficient of determination (r^2) represented as a percentage was calculated to estimate the proportion of variance in dependent variable that is accounted for by the independent variable. The strength of the variable relationship was determined based on the Pearson r-value scale of -1.0 to 1.0 (figure 6). A list of correlated variables was then created for further analysis to determine the relationship form and any trend patterns.



High Correlation (Positive)	0.5	to	1.0
Moderate Correlation (Positive)	0.3	to	0.5
Low Correlation (Positive)	0.1	to	0.3
No Correlation	-0.1	to	0.1
Low Correlation (Negative)	-0.3	to	-0.1
Moderate Correlation (Negative)	-0.5	to	-0.3
High Correlation (Negative)	-1.0	to	-0.5

(*Figure 6 – Pearson Coefficient Strength Scale)*

The result list of correlated variables produced by the PPMC test provided the direction and strength of each relationship. Next, the relationship form must be determined to provide a complete understanding of the existing correlations. For this purpose a regression analysis as performed on each set of related variables. A scatter-plot graph was created for each pair of values, in each graph the x-axis was a dependent variable and the y-axis was an independent variable (Google Trend search-term). The regression analysis was only conducted on the state-level result sets, because of the small set of values available at the national-level (4 values, compared to 51 for the state-level). Once the scatter-plot graphs were created a trend-line was applied, as well as the R^2 value and graph equation. The R^2 value represents the line fit between the independent and dependent variable, while the equation provides the slope of the best-fit linear trend-line and can be used in a forecasting analysis.[2]

A trend analysis on a national-level was the last statistical analysis that was performed in this study. While trend analysis is commonly used to predict future activity, it can also be used to detect any underlying patterns or behaviors within variable datasets. For the purpose of this study, a trend analysis was done to compare the all the independent variables against a single dependent variable over time to see if any patterns existed. This analysis was performed in Microsoft Excel by creating a combination chart where a column chart was used to represent the Google Trend datasets over time and a line graph was used to show a given dependent variable trend. The goal was to see if a pattern such as both variables mirrored the same behaviors or if the behaviors were reversed (i.e. the independent variable increased while the dependent variable decreased).

4.4 Hypothesis of Expected Results

The following are the hypotheses of the expected results for the correlation, regression, and trend analysis tasks. These are discussed in terms of dependent and independent variable correlation and strength.

With respect to the correlation analysis, the higher the search interest for selected search-terms, the higher the suicide rate on both a state and national level. This will demonstrate a moderate to strong correlation at both levels for the Google Trend each search-terms. However, in the case of the correlation between SAMHSA total utilization rates and the Google Trends search-term interest, the correlation will be primarily low. As for the Gallup-Healthways Well-Being Index, there will most likely be a low to moderate correlation with the Google Trends

[2] *The equation is provided for future reference; however forecast analysis is out of scope for this study.*

search-terms because more than emotional health is taken into account for the calculation of this measure.

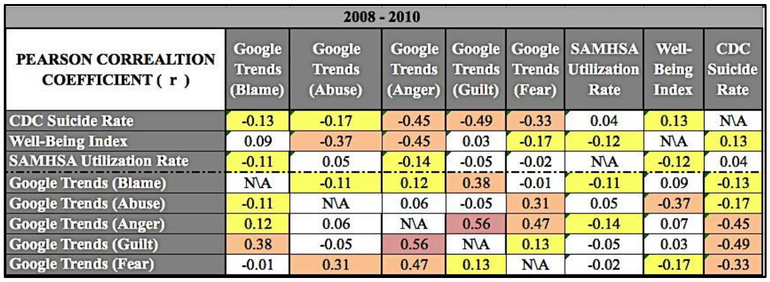
If any moderate to strong correlations exist, they will be linear in nature and the variables themselves will increase or decrease with respect to one another at a gradual rate. This is in part because of the limited time period that is being observed and presented in this study. If a larger time range would be considered dataset was analyzed the results would still be linear in nature, but the slope might be steeper.

In regards to underlying trend patterns, the only expected pattern would be between all the independent variables (Google Trends search-terms) and the suicide rate dependent variable. In reference to the first hypothesis of a strong to moderate correlation between the search-terms and suicide rates, this analysis would also be expected to be reflected in a trend analysis. The trend analysis would illustrate a similar pattern for increase or decrease over time.

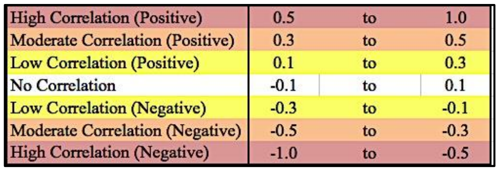
Chapter 5. RESULTS

5.1 Correlation Analysis

The state-level PPMC correlation results for the data revealed that the suicide rates provided by the Center of Disease Control and Prevention (CDC) moderately correlated with 60% of the adjusted search interests from the Google Trends datasets: r (Guilt) = -0.49, r (Anger) = -0.45, r (Fear) = -0.33. However, a weak negative correlation also existed between the two remaining search-terms, r (Abuse) = -0.17, r (Blame) = -0.13 (Refer to Table 1). However, when examining the correlation results at an averaged national level, 60% of the search-terms demonstrated a strong positive correlation: r (Guilt) = 0.95, r (Blame) = 0.77, r (Anger) = 0.55. Of the remaining search-terms ‘Fear’ revealed a moderate positive correlation of r = 0.38, while ‘Abuse’ showed a weak negative correlation with r = -0.20. The search-term correlations differed based on the level of granularity and the extended time range of Google Trends datasets (Refer to Table 2).



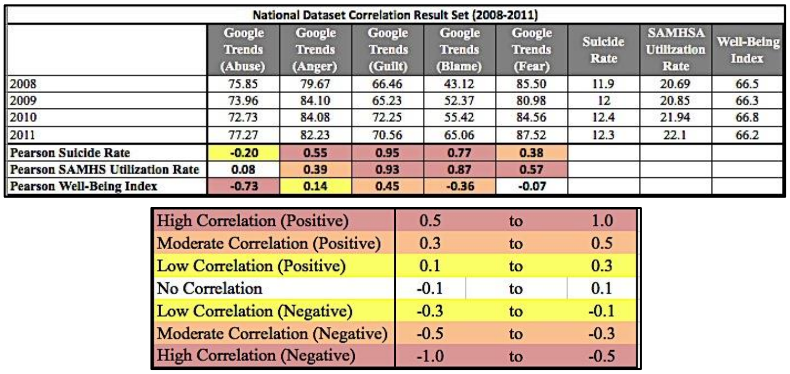
PEARSON CORREALTION COEFFICIENT (r)	2008 - 2010							
	Google Trends (Blame)	Google Trends (Abuse)	Google Trends (Anger)	Google Trends (Guilt)	Google Trends (Fear)	SAMHSA Utilization Rate	Well-Being Index	CDC Suicide Rate
CDC Suicide Rate	-0.13	-0.17	-0.45	-0.49	-0.33	0.04	0.13	N\A
Well-Being Index	0.09	-0.37	-0.45	0.03	-0.17	-0.12	N\A	0.13
SAMHSA Utilization Rate	-0.11	0.05	-0.14	-0.05	-0.02	N\A	-0.12	0.04
Google Trends (Blame)	N\A	-0.11	0.12	0.38	-0.01	-0.11	0.09	-0.13
Google Trends (Abuse)	-0.11	N\A	0.06	-0.05	0.31	0.05	-0.37	-0.17
Google Trends (Anger)	0.12	0.06	N\A	0.56	0.47	-0.14	0.07	-0.45
Google Trends (Guilt)	0.38	-0.05	0.56	N\A	0.13	-0.05	0.03	-0.49
Google Trends (Fear)	-0.01	0.31	0.47	0.13	N\A	-0.02	-0.17	-0.33



High Correlation (Positive)	0.5	to	1.0
Moderate Correlation (Positive)	0.3	to	0.5
Low Correlation (Positive)	0.1	to	0.3
No Correlation	-0.1	to	0.1
Low Correlation (Negative)	-0.3	to	-0.1
Moderate Correlation (Negative)	-0.5	to	-0.3
High Correlation (Negative)	-1.0	to	-0.5

(Table 1 – Pearson Correlation Coefficient Result Matrix [State-Level])

In contrast, the state-level PPMC results for the SAMHSA total utilization rate data did not correlate significantly, if at all, with the adjusted search interests from the Google Trends datasets. Only two search-terms showed relationships, each had a very weak negative correlation: r (Anger) = -0.14 and r (Blame) = -0.11. No correlation was found for the 'Abuse', 'Guilt', and 'Fear' search-terms at the state-level. (Refer to Table 1). Surprisingly, 80% of the search-terms evaluated resulted in a strong or moderate positive correlation at the averaged national-level. Strong positive correlations existed with 'Guilt' (r = 0.93), 'Blame' (r = 0.87), and 'Fear' (r = 0.57), while a moderate positive correlation resulted for 'Anger' (r = 0.39). No correlation was found for the 'Abuse' search-term at the national-level. (Refer to Table 2).



National Dataset Correlation Result Set (2008-2011)								
	Google Trends (Abuse)	Google Trends (Anger)	Google Trends (Guilt)	Google Trends (Blame)	Google Trends (Fear)	Suicide Rate	SAMHSA Utilization Rate	Well-Being Index
2008	75.85	79.67	66.46	43.12	85.50	11.9	20.69	66.5
2009	73.96	84.10	65.23	52.37	80.98	12	20.85	66.3
2010	72.73	84.08	72.25	55.42	84.56	12.4	21.94	66.8
2011	77.27	82.23	70.56	65.06	87.52	12.3	22.1	66.2
Pearson Suicide Rate	**-0.20**	**0.55**	**0.95**	**0.77**	**0.38**			
Pearson SAMHS Utilization Rate	**0.08**	**0.39**	**0.93**	**0.87**	**0.57**			
Pearson Well-Being Index	**-0.73**	**0.14**	**0.45**	**-0.36**	**-0.07**			

High Correlation (Positive)	0.5	to	1.0
Moderate Correlation (Positive)	0.3	to	0.5
Low Correlation (Positive)	0.1	to	0.3
No Correlation	-0.1	to	0.1
Low Correlation (Negative)	-0.3	to	-0.1
Moderate Correlation (Negative)	-0.5	to	-0.3
High Correlation (Negative)	-1.0	to	-0.5

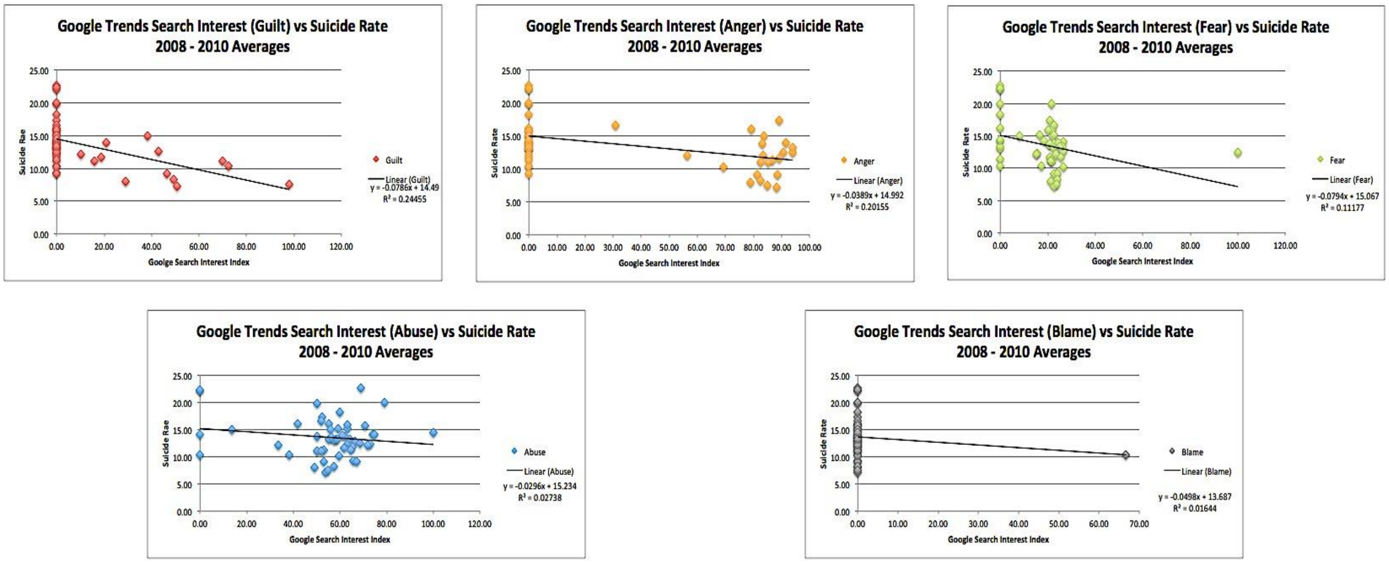
(Table 2 – National Pearson Correlation Result Matrix [National-Level])

The correlation analysis results for the Gallup-Healthways Well-Being Index yielded similar results to the SAMHSA total utilization rate. Two correlations were discovered at the state-level between the search-term Google Trends datasets and the Well-Being Index. One moderate negative correlation was revealed for ‘Abuse’ with r = -0.37 and one weak negative correlation existed for ‘Fear’ with r = -0.17. No correlation was found for the ‘Blame’, ‘Anger’ and ‘Guilt’ search-terms at the state-level. From the perspective of on a national-level, less correlation existed within this dataset in comparison with the SAMHSA total utilization rate and CDC Suicide Rate datasets. ‘Abuse’ was the only strong negative relationship with an r = -0.73, which is low in comparison to the highest national correlation coefficients within the other two datasets (CDC suicide rate (‘Guilt’) = 0.95 and SAMHSA total utilization rate (‘Guilt’) = 0.93). The search-term ‘Guilt’ showed a moderate positive correlation and ‘Blame’ resulted in a moderate negative correlation. No correlation was found for the ‘Fear’ search-term at the national-level.

5.2 Regression Analysis

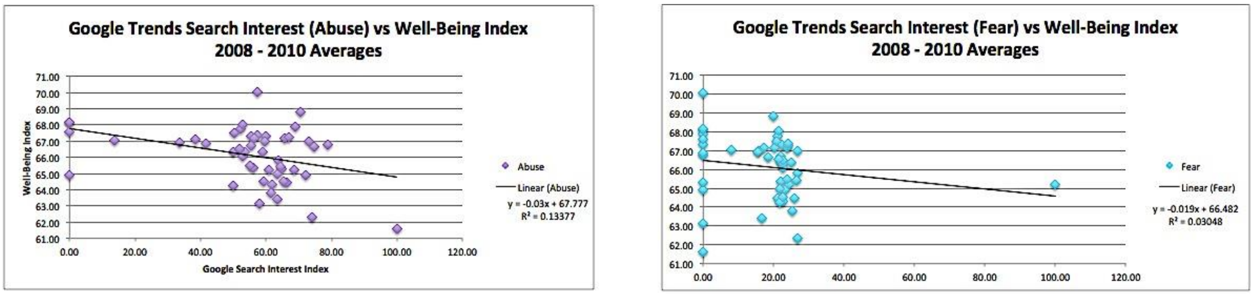
The result sets from the correlation analysis where a moderate to strong correlation was presented at a state-level was used as the basis for a simple linear regression analysis. The goal of this analysis was to model the relationship revealed in the previous analysis to provide a better understanding for result set interpretations. For this simple linear regression analysis, each correlated pair was graphed to a scatterplot to show the form of the regression, direction, trend-line, equation of the line, and goodness of fit for the model (R^2). The closer to 0 the less ‘good’ the model fits the data, while the closer to ‘1’ the better the data fits to the model.

The linear regression analysis was performed to each of the search-terms for the Google Trends datasets and compared to the CDC suicide rates dataset. Each search-term showed a level of correlation in the previous analysis. However, 40% of the search-terms showed a low level of “goodness to fit” for the model: R^2(‘Guilt’) = 0.24 R^2(‘Anger’) = 0.20, while the remaining 60% hardly showed a “good fit”. Each correlation scatter plot for Google Trends vs. suicide rate did show a decreasing, negative form and direction. The scatterplot graph for Google Trends (‘Blame’) showed very little form, having much of the plots around the same grouping (refer to charts presented in figure 7).



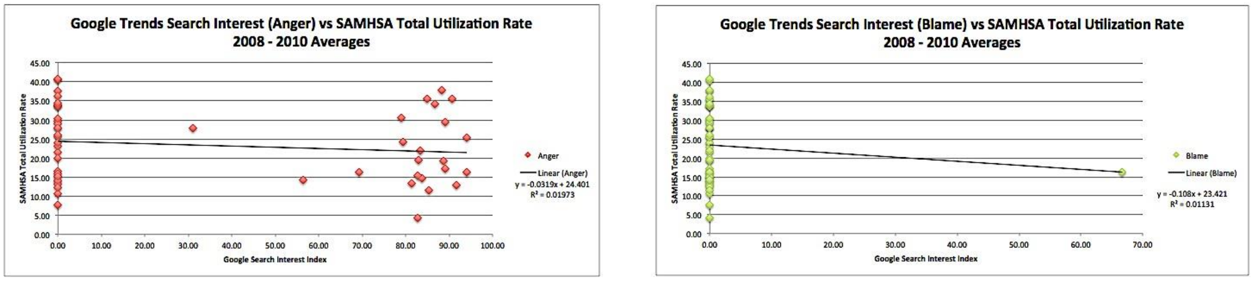
(Figure 7 – Suicide Rates vs. Correlated Google Trends Regression Analysis)

A simple linear regression analysis was performed for two out of the five search-terms for the Google Trends datasets (‘Abuse’ & ‘Fear’) and compared to the Gallup-Healthways Well-Being Index. These were chosen because they were the only two search-terms to show any significant correlation in the previous analysis. The scatter plot for the ‘Abuse’ search-term Google Trends data illustrated the moderate negative or indirect correlation that exists between the dependent (y or predicted) and independent (x or predictor) variables with the greater ‘good fit’ value of the two search-terms: R^2 (‘Abuse’) = 0.13 and R^2(‘Blame’) = 0.03 (refer to Figure 8).



(Figure 8 – Well-Being Index vs. Correlated Google Trends Regression Analysis)

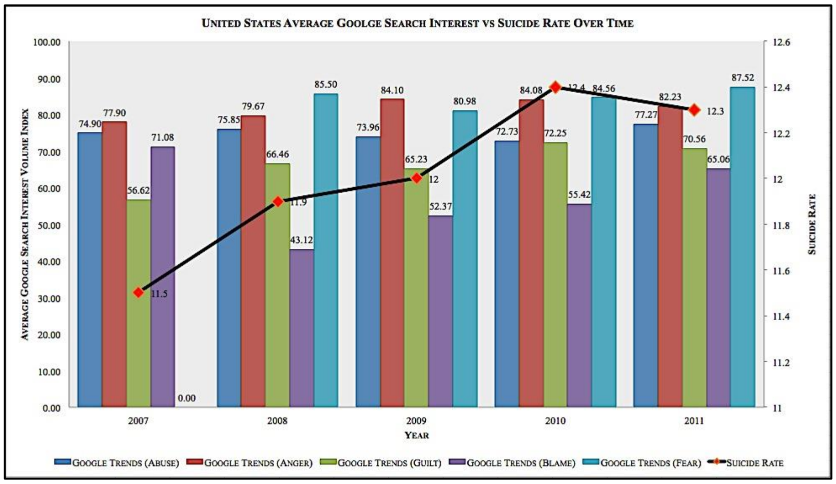
Out of all the dependent datasets the SAMHSA total utilization rates showed the least correlation with the five Google Trend search-terms datasets. Only two search-terms, ‘Anger’ and ‘Blame’, showed correlation, each is considered weak negative correlations. This was illustrated in more detail by the linear regression analysis. In both instances the trend-line demonstrated a weak fit for the model (R^2 (‘Anger’) = 0.02 and R^2(‘Blame’) = 0.01). Each scatterplot also consisted of a shallow slope or change between datasets (refer to Figure 9).



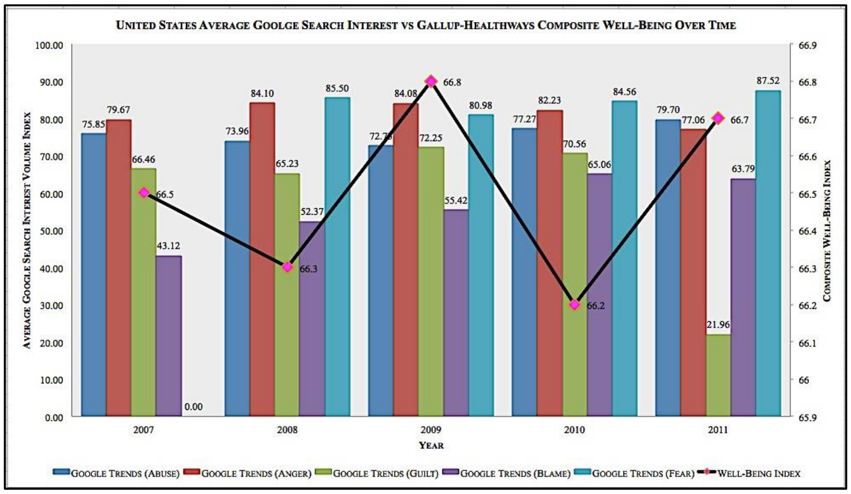
(Figure 9 – Correlated Google Trends vs. SAMHSA Utilization Rate Regression Analysis)

5.3 Trend Analysis

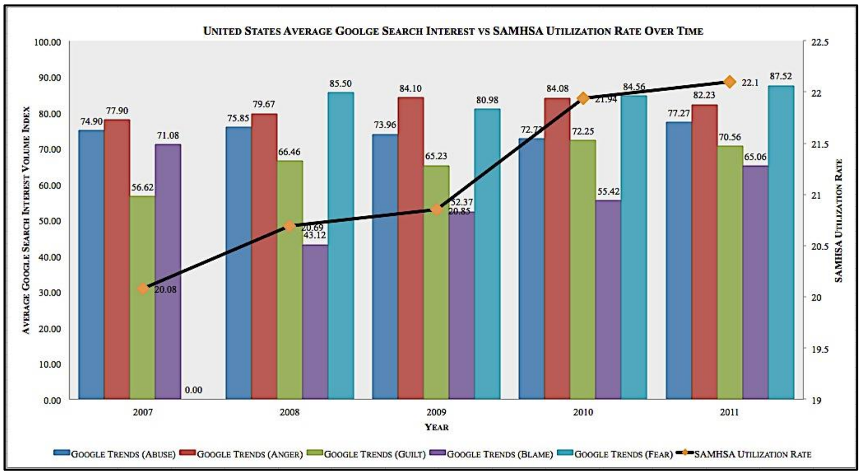
A trend analysis was conducted through the use of a combination graph to reveal any underlying data patterns between the independent and dependent datasets. In the combination graph, each averaged search-term Google Trends search interest was represented overtime as a series on the column-graph, while the dependent variable (CDC suicide rates, SAMHSA total utilization rate, or Gallup-Healthways Well-Being Index) was represented as a line-graph. The only underlying pattern that existed between all the trend analyses was between the Google Search Interest and Suicide Rate over time. Excluding the year 2007, the search interest increases and decreases between 2008 and 2011 aligned with the suicide rate behaviors. The Google Search Interest SAMHSA total utilization rate and Well-Being Index both showed no consistency overtime (refer to charts 1 to 3).



(Chart 1 – Google Search Interest vs. Suicide Rate Trend Analysis)



(Chart 2 - Google Search Interest vs. Well-Being Index Trend Analysis)



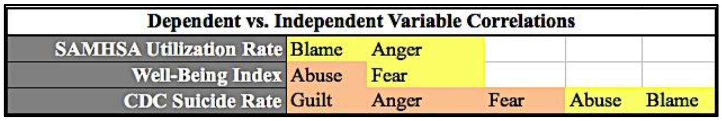
(Chart 3 - Google Search Interest vs. Total Utilization Rate Trend Analysis)

Chapter 6. CONCLUSION

6.1 Discussion

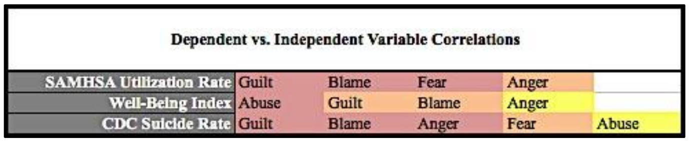
The purpose of this study has been to evaluate and determine if correlation exists between web analytic tools such as Google Trends and static national mental illness statistics, with an emphasis on suicidal behavior. Various studies have been conducted in the past to show a correlation between national mental illness statistics and social networking; however this study dug deeper into the use of search engine queries.

Going into this study the first hypothesis made was the belief that a moderate to strong correlation would exist between suicide rates and each of the Google Trend search-terms. The results of the correlation analysis of these datasets did demonstrate significant correlation for each search-term on both a state and national level. However, no state-level strong correlations were revealed and only 60% of the words had a moderate correlation. There was a pattern between the linear regression analysis and the correlation analysis of the moderately correlating search-terms and suicide rates. In the linear regression, two out of the three moderate negative correlating search-terms had similar "goodness-fit", R^2 values. At a national-level 80% of the search-terms used for the Google Trends datasets revealed a moderate to strong correlation (refer to table 3 & 4). The knowledge gained from both the correlation and regression analysis at the state and national level could be interrupted to mean that lower suicide rates were associated with lower search-interest for each of the search-terms, and this data could be applied to predictive analytics for further statistical testing.



Dependent vs. Independent Variable Correlations					
SAMHSA Utilization Rate	Blame	Anger			
Well-Being Index	Abuse	Fear			
CDC Suicide Rate	Guilt	Anger	Fear	Abuse	Blame

(Table 3 – Dependent vs. Independent Variable Correlations)



Dependent vs. Independent Variable Correlations					
SAMHSA Utilization Rate	Guilt	Blame	Fear	Anger	
Well-Being Index	Abuse	Guilt	Blame	Anger	
CDC Suicide Rate	Guilt	Blame	Anger	Fear	Abuse

(Table 4 – National Dependent vs. Independent Variable Correlations)

The results for the SAMHSA total utilization rate dataset in comparison to the Google Trends datasets were consisted with the stated expected hypothesis. At a state-level the utilization rates showed the weakest overall search-term correlations in comparison to the suicide rates and Well-Being index datasets. This could indicate that negative affect terms might be searched to a degree of frequency, but would not indicate a link to individuals who seek mental health services. This could also be due to the anti-social behavior some individuals with mental illness demonstrate. In other words, those experiencing depression may not seek mental health services and try to self-treat which could cause an increase in related online search queries.

When examining the results of each analysis for the Gallup-Healthways Well-Being Index, correlation and patterns were revealed on both a state and national level. While these statistical findings were higher than those of the SAMHSA total utilization rate results, they were not as high or consistent as those of the suicide rate result sets. This could be in part because of the wide spectrum of domains considered during the calculation of the Well-Being Index, having only 30% of the index representing health related domains. This measure did provide a more rounded interruption of the datasets and could be used as more of a 'bridging' factor between the suicide rate and utilization rate datasets.

While this study yielded interesting results, some of which aligned with the initial thoughts and hypotheses, internal and external limitations imposed on these results could have impacted the results to an extent. For instance, had the time range overlap been consistent and longer between all the datasets and data granularities, these tests would have given a stronger interpretation and finding. Bias could have also been implicated into the results through the use of averaging techniques. These averages could have inflated the correlation and trend analysis results, as evidence in the national-level result sets.

6.2 Conclusions

Data mining and analysis of online activity has aided in extending the information and knowledge base for mental illness for public health. In this study, the potential of using online search trending as an additional data source for public health surveillance, early intervention, and prevention was demonstrated through various statistical analyses. Each statistical test focused on identifying relationships between this data source and traditional, government-provided statistics, such as suicide rates.

The results provided through the use of technologies like Google Trends in this study revealed that search engine trending could be leveraged in the future to alert and monitor public mental health and suicidal behavior. This found correlation could contribute to the continuous search for non-traditional mental health data capture methods. One of the ongoing struggles with the early detection and prevention of suicidal behavior and depression-like mental illnesses is that dependency of the individual reach out to others directly. Often times the nature of these behaviors make the possibility of these persons reaching out decline. By leveraging indirect communications through the analysis of social media, text-messaging, and online search trending a different perspective provided, thus lessening the dependency on individual's interactions with health personnel.

While more emphasis was given to suicide-based illness and behaviors, other mental illnesses could also implement similar methodologies to determine correlation in future studies. The data capture processes and struggles experienced throughout this study indirectly highlighted the continuing need for national and state level electronic data capture usage with in the health industry. As the United States moves closer to adopting electronic medical record usage, the ability to access more current datasets in the future would lead to a more in-depth exploration into mental health surveillance, as well as encourage further investigation to the roles online resources could play.

REFERENCES

Baer, J. (n.d.). 11 Shocking New Social Media Statistics in America [Web log post]. Retrieved June, 2013, from http://www.convinceandconvert.com/the-social-habit/11-shocking-new-social-media-statistics-in-america/

Cherry, C., Mohammad, S. M., & Bruijn, B. D. (2012, January 30). Binary Classifiers and Latent Sequence Models for Emotion Detection in Suicide Notes. *National Center for Biotechnology Information*. Retrieved July 19, 2013, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3409480/

De Choudhury, M., Counts, S., & Horvitz, E. (2013, May 4). *Social Media as a Measurement Tool of Depression in Populations*. Retrieved July 29, 2013, from http://research.microsoft.com/en-us/um/people/horvitz/depression_populations_websci_2013.pdf

"The Durkheim Project" Will Analyze Opt-In Data From Veterans' Social Media And Mobile Content -- Seeking Real-Time Predictive Analytics for Suicide Risk. (2012). *PR Newswire*. Retrieved July 26, 2013, from http://www.prnewswire.com/news-releases/the-durkheim-project-will-analyze-opt-in-data-from-veterans-social-media-and-mobile-content----seeking-real-time-predictive-analytics-for-suicide-risk-213922041.html

Gallup-Healthways Well-Being Index Findings (Rep.). (n.d.). Retrieved May 30, 2013, from http://www.well-beingindex.com/findings.asp

Google Annual Search Statistics (Rep.). (2013, June 18). Retrieved July, 2013, from Google Official History, Comscore website: http://www.statisticbrain.com/google-searches/

Google Trends [Computer software]. (n.d.). Retrieved June/July, 2013, from http://www.google.com/trends

How People Spend Their Time Online [Infographic] [Web log post]. (2012, February 2). Retrieved August, 2013, from http://www.go-gulf.com/blog/online-time/

Our Project. (n.d.). *Durkheim Project*. Retrieved June, 2013, from http://durkheimproject.org/our-project/

Shalin, D. (2004, November 5). Suicide Trends and Prevention in Nevada. *Suicide Trends*. Retrieved August, 2013, from http://cdclv.unlv.edu/healthnv/suicide.html

State of the States [Computer software]. (n.d.). Retrieved June 5, 2013, from http://www.gallup.com/poll/125066/State-States.aspx

Substance Abuse & Mental Health Services Administration. (n.d.). *SAMHSA Uniform Reporting System (URS) Output Tables: 2007*. Retrieved June 10, 2013, from http://www.samhsa.gov/dataoutcomes/urs/urs2007.aspx

Substance Abuse & Mental Health Services Administration. (n.d.). *SAMHSA Uniform Reporting System (URS) Output Tables: 2008*. Retrieved June 10, 2013, from http://www.samhsa.gov/dataoutcomes/urs/urs2008.aspx

Substance Abuse & Mental Health Services Administration. (n.d.). *SAMHSA Uniform Reporting System (URS) Output Tables: 2009*. Retrieved June 10, 2013, from http://www.samhsa.gov/dataoutcomes/urs/urs2009.aspx

Substance Abuse & Mental Health Services Administration. (n.d.). *SAMHSA Uniform Reporting System (URS) Output Tables: 2010*. Retrieved June 10, 2013, from http://www.samhsa.gov/dataoutcomes/urs/urs2010.aspx

Substance Abuse & Mental Health Services Administration. (n.d.). *SAMHSA Uniform Reporting System (URS) Output Tables: 2011*. Retrieved June 10, 2013, from http://www.samhsa.gov/dataoutcomes/urs/urs2011.aspx

Xu, J., M.D., Kochanek,, K. D., M.A., Murphy, S. L., B.S., & Tejada-Vera, B., B.S (2010). Deaths: Final Data for 2007. *National Vital Statistics Reports*, *58*(19). Retrieved June 05, 2013, from http://www.cdc.gov/nchs/data/nvsr/nvsr58/nvsr58_19.pdf

Xu, J., M.D., Kochanek,, K. D., M.A., Murphy, S. L., B.S., & Miniño, A. N., M.P.H. (2011). Deaths: Final Data for 2008. *National Vital Statistics Reports*, *59(*10). Retrieved June 05, 2013, from http://www.cdc.gov/nchs/data/nvsr/nvsr59/nvsr59_10.pdf

Xu, J., M.D., Kochanek,, K. D., M.A., Murphy, S. L., B.S., Miniño, A. N., M.P.H., & Kung H., Ph.D. (2011). Deaths: Final Data for 2009. *National Vital Statistics Reports*, *60*(3). Retrieved June 05, 2013, from http://www.cdc.gov/nchs/data/nvsr/nvsr60/nvsr60_03.pdf

Xu, J., M.D., Kochanek,, K. D., M.A., & Murphy, S. L. (2013). Deaths: Final Data for 2010. *National Vital Statistics Reports*, *61*(4). Retrieved June 05, 2013, from http://www.cdc.gov/nchs/data/nvsr/nvsr61/nvsr61_04.pdf

Xu, J., M.D. & Hoyert, L, D., Ph.D. (2013). Deaths: Preliminary Data for 2011. *National Vital Statistics Reports*, *61*(4). Retrieved June 05, 2013, from http://www.cdc.gov/nchs/data/nvsr/nvsr61/nvsr61_06.pdf