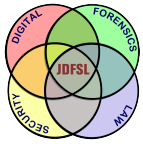

THE JOURNAL OF DIGITAL FORENSICS, SECURITY AND LAW

Journal of Digital Forensics, Security and Law

Volume 8 | Number 2 Article 3

2013

How Often Is Employee Anger An Insider Risk II? Detecting and Measuring Negative Sentiment versus Insider Risk in Digital Communications–Comparison between Human Raters and Psycholinguistic Software

Eric Shaw

Maria Payri

Michael Cohn

Ilene R. Shaw

HOW OFTEN IS EMPLOYEE ANGER AN INSIDER RISK II? DETECTING AND MEASURING NEGATIVE SENTIMENT VERSUS INSIDER RISK IN DIGITAL COMMUNICATIONS–COMPARISON BETWEEN HUMAN RATERS AND PSYCHOLINGUISTIC SOFTWARE

Eric Shaw, Ph.D.*
Suite 514
5225 Connecticut Ave. NW
Washington, DC 20015
eshaw@msn.com

Maria Payri, M. A.

Michael Cohn, M. A.

Ilene R. Shaw, R. N.

*Corresponding Author

ABSTRACT

This research uses two recently introduced observer rating scales, (Shaw et al., 2013) for the identification and measurement of negative sentiment (the Scale for Negativity in Text or SNIT) and insider risk (Scale of Indicators of Risk in Digital Communication or SIRDC) in communications to test the performance of psycholinguistic software designed to detect indicators of these risk factors. The psycholinguistic software program, WarmTouch (WT), previously used for investigations, appeared to be an effective means for locating communications scored High or Medium in negative sentiment by the SNIT or High in insider risk by the SIRDC within a randomly selected sample from the Enron archive. WT proved less effective in locating emails Low in negative sentiment on the SNIT and Low in insider risk on the SIRDC. However, WT performed extremely well in identifying communications from actual insiders randomly selected from case files and inserted in this email sample. In addition, it appeared that WT's measure of perceived Victimization was a significant supplement to using negative sentiment alone, when it came to searching for actual insiders. Previous findings (Shaw et al., 2013) indicate that this relative weakness in identifying low levels of negative sentiment may not impair WT's usefulness for identifying communications containing

significant indications of insider risk because of the very low base rate and low severity of insider risk at Low levels of negative sentiment (Shaw et al., 2013). Although many of the "false positives" acquired in the successful search for actual insiders in this experiment were shown to be true positives for other forms of insider risk, WT still produced fairly high rates of false positives that could burden analysts, as described by the search times provided. As further research and development proceeds to address this problem, we again recommend the use of WT in an integrated multi-disciplinary array of detection methods that will serve as an initial screen to narrow the search for individuals at-risk for insider activities. The implications for insider threat research, detection and prevention are discussed.

Keywords: Insider Risk, Digital Communication, Disgruntlement, Detecting, Negative Sentiment, Threat, Employee Anger, Workplace Violence, Content Analysis, Automated Psychological assessment, Psycholinguistic software

1. INTRODUCTION

Shaw et al. (2013) have summarized the long-documented association between employee disgruntlement and counter-productive work behaviors (CWB), including insider attacks and violations. They also discussed the unexplored and complicated theoretical relationship between expressed negative sentiment and insider risk and described the first published empirical data to address such basic questions as what percentage of organizational emails contain negative sentiment and what percentage of this content also contains indicators of insider risk. Using two new observational rating scales of negative sentiment (the Scale of Negativity in Text or SNIT) and insider risk (Scale of Insider Risk in Digital Content or the SIRDC) with a random sample of emails from the Enron Archive, this study found that low levels of negative sentiment were found in 20% of the sample and that Moderate and High levels of negative sentiment were extremely rare, occurring in less than 1% of communications. Less than 4% of the sampled emails displayed indicators of insider risk on the SIRDC. Emails containing High levels of insider risk comprised less than one percent of the total sample. Of the emails containing negative sentiment in the sample, only 16.3% also displayed indicators of insider risk. However, the odds of a communication containing insider risk increased with the level of negative sentiment. All of the emails found to contain insider risk indicators on the SIRDC displayed some level of negative sentiment. The authors concluded that searching communications for insider risk using mainly negative sentiment produces an extremely high false positive rate (over 84%) and that communications with low levels of negative sentiment are particularly unlikely to yield true positives. They also suggested that research efforts utilizing email collections without moderate-to-high levels of negative sentiment are unlikely to be useful due to their lack of inclusion of individuals with actual insider potential.

Clearly, communications containing negative sentiment are not rare, but communications with insider risk variables overlap with these communications and are extremely rare, complicating efforts to identify at-risk individuals through their communications. While human coders can reliably distinguish emails with negative sentiment, with and without insider risk, it is not practical or desirable from a privacy perspective, to deploy human readers in a search for these rare communications. While numerous risk detection systems have been deployed to detect technical behaviors associated with insider risk (anomalous system use, technical security violations, behavior consistent with more complex insider risk models (see Shaw and Stock, 2011) there are no available computerized systems designed to detect negative sentiment and other insider risk indicators in content beyond key word identification or document proliferation trackers adapted from legal review systems. Although a number of automated systems for detecting negative and other sentiment have appeared on the market, these software packages mainly target very general consumer, employee and media sentiment, rather than CWB risk in communications. Among the multiple computer-assisted software programs designed to identify and measure qualitative content in electronic media, very few focus on the emotional states and psychological characteristics of particular individuals, and none do so for the purpose of assessing risk to others. The overwhelming majority of Qualitative Data Analyzers (QDAs) or Computer Assisted/Aided Qualitative Data Analysis Software (CAQDAS) are primarily designed for commercial concerns, reputation management, and targeted opinion monitoring. These systems are not equipped with specialized scales that include individual psychological indicators to detect insider risk. They do not identify and measure the psychological state of the author or provide data on other psychological characteristics of individuals linked to insider risk.

A system that could identify levels of negative sentiment and insider risk of concern to security, counter-intelligence, and fraud analysts and investigators could assist them in narrowing their search for suspects in these investigations. For example, the search for individuals with access, capability and/or technical or other risk indicators could be further narrowed by identifying disgruntled subjects within this suspect pool. With input from other multi-disciplinary data sources (Human Resources, Security, Financial), such a system could also help identify individuals with greater levels of disgruntlement than their peers, or increased levels of insider risk factors compared to their previous levels.

The availability of a corporate email sample reliably coded for negative sentiment and insider risk provided the opportunity to test the ability of psycholinguistic content analysis software to identify such at-risk communications. In addition to the system's ability to screen communications for signs of risk within this sample, this design also allowed for the insertion of actual insider emails (and communications from other criterion groups) into the

corpus to evaluate the system's sensitivity and false positive rate for these communications. This section provides a brief overview of the software being evaluated and its expected limitations, followed by a description of the research design utilized and the results obtained. The implications for the use of psycholinguistic software in insider investigations and detection efforts will then be discussed in light of these findings and the results from the previous article (Shaw et. al., 2013).

1.1 Background on WarmTouch (WT) Psycholinguistic Software[1]

WT was previously described in Shaw and Stroz (2004) and was designed as a tool to help investigators and analysts discover trends in communications relevant to a subject's emotional and psychological state, attitudes toward others, decision-making processes, and communication preferences. While its variables are derived, in part, from content analysis approaches to leadership analysis (Hermann, 1980; Shaw, 2003; Weintraub, 1986; Winter, Hermann, Weintraub, and Walker, 1991), personality, and threat assessment (Shaw, 2006), it has mainly been used to:

- Detect signs of disgruntlement and other insider risk factors in online communications of individuals previously identified as at-risk for insider problems.
- Assess the risk posed by the psychological state and attitudes of anonymous individuals making threats.
- Compare the psycholinguistic patterns of anonymous and known individuals to help determine the authorship and identity of unknown persons.
- Evaluate the communication preferences and dominant decision-making processes of these subjects to aid in case management.
- Describe the frequency and valence of communication between parties for the purpose of understanding relationship patterns and locating patterns indicative of risk.
- Assess the communication preferences and decision-making styles of parties of interest from their communication in order to understand, enhance or impact relationships (Shaw and Stroz, 2004).
- WT's deployment in illustrative cases is detailed in Shaw and Stroz (2004). WT has also been applied to the risk assessment of persons suspected of terrorist activities within American companies abroad, analysis, profiling and management of persons engaged in cyber

[1] WarmTouch and its use of psycholinguistic algorithms for detecting insider risk is protected by US Patents 7,058,566; 7,225,122; 7,346,492; 7,395,201; 7,526,426; 7,801,724; 7,881,924; 8,078,453.

extortion, and insiders participating in espionage, sabotage and presenting violence risk (O'Brien, 2005; Shaw and Stroz, 2004).

WT utilizes two different psycholinguistic approaches to perform these tasks. Its variables include psycholinguistic algorithms derived from the academic content analysis literature and linguistic dictionaries designed to correspond to specific psychological states or issues. In our experience the combination of these approaches makes detection of specific psychological states characterized by general arousal more likely. For example, disgruntled individuals may express their feelings of anger and victimization through the use of specific words with psycholinguistic meanings that form our algorithms (negatives as indicative of anger, the use of the word "me" as indicative of feelings of victimization). However, in some cases they may also simply threaten an anti-social act without exhibiting a marked change in these variables (e.g., "I will *kill* you," scored as instrumental aggression and a key threat word). It was therefore important in the development process that WT be able to detect both types of risk indicators. In addition, our experience indicates that there is some degree of overlap in the psycholinguistic expression of aroused states (anger, anxiety, depression), and that while the psychological algorithms used are good at detecting arousal, the psychological dictionaries are extremely useful for defining the specific psychological state involved. For example, in our case experience, angry, depressed and anxious individuals tend to show elevated levels of self-references ("I" and "me"), direct references to others, negatives, adverbial intensifiers, and negative feelings compared to controls. However, dictionary scales designed to differentiate these states are helpful in distinguishing them from each other. For example, the use of words such as "sad," "blue," and "bummed" versus "worried," "nervous," and "agitated" help us differentiate depression from anxiety.

As noted above, WT uses a number of traditional psycholinguistic content analysis coding schemes derived from work with written and spoken samples of leadership and patient verbal behavior described in the WT codebook (Shaw and Wirth-Beaumont, 2004). However, in the process of applying these measures to the online communication of general employee populations, persons in significant psychological distress, and forensic populations, we have been both supporting and revising the psycholinguistic interpretation of many of these variables. For example, in his work with patient populations and leaders, Weintraub (1989) found that the use of the word "me" at relatively high levels was indicative of passivity. This followed logically from the overwhelming tendency for "me" to be used as a pronoun denoting the receipt of action by others. Consistent with this finding, in our cases involving insiders we have found "me" to be an excellent marker of perceived victimization in the written and online communications of disgruntled and other forensic subjects. We will therefore also test the sensitivity of this measure of Victimization to

disgruntled insiders as a supplement to the less sensitive performance of negative sentiment alone, described in the earlier article (Shaw et. al., 2013).

Weintraub (1989) also did not subdivide his variable of Evaluators (judgments regarding positive or negative aspects or characteristics) and Feelings (either positive or negative expressions of emotions) into negative and positive categories. In studying individuals at-risk, versus leadership populations, we have found this subdivision useful.

Because the effectiveness of the WT psychological dictionary of terms will depend on its ability to capture the way specific subjects experience and communicate their thoughts and feelings online, they require regular revision when new subject groups are examined. For example, during a recent assignment monitoring the communication of employees suspected to be members of a potentially violent Islamic fundamentalist group, the WT dictionary was revised to include new terms in previous categories including depersonalization ("kafir," "infidel"), dangerous religiousness ("jahiliyya," "shahid"), negative evaluators ("against allah," "crusader"), negative feelings ("wrathful"), etc. Thus, prior to the deployment of WT in a new setting, updated vocabulary research is undertaken.

1.2 Use of WT in the identification of individuals at-risk for insider acts from within an email or other digital communication cache

WT was not originally designed to be used alone to locate individuals at-risk for insider acts due to their display of disgruntlement or other risk factors in their communications. Nor would we recommend its lone use for this purpose. However, we wanted to test the sensitivity and specificity of WT as a screening tool to help analyst reduce the number of communications to be reviewed in their efforts to locate potentially at-risk individuals, in combination with other risk detection approaches, such as signs of technical behavioral anomalies or variations from baseline compared to a subject's group or his own behavior (unusual downloading or copying) or consistency with more advanced models of insider risk (see summary by Shaw and Stock, 2011). For example, insider threat analysts utilizing technical risk indicators are often overwhelmed by the number of "alerts" that might indicate the presence of risk and have difficulty prioritizing their resources to investigate this feedback. However, a parallel system that could provide additional information about the likely level of disgruntlement in an individual could help prioritize these efforts.

1.3 How does WT Work?

When it comes to using WT as a screening tool to search an email or other communication cache for individuals at-risk for insider or other related problems, WT presents several options. First, WT can conduct a very broad search according to a wide variety of psycholinguistic criteria, casting a broad

net. Or, WT can search a communications cache for a very narrow selection of emails fitting very specific criteria. For example, in a search for emails containing negative sentiment, a WT analyst would select a list of psycholinguistic variables sensitive to negative emotions or beliefs, including some specialized variables sensitive to particular states applicable to insiders. Table A1 in the Appendix displays a list of WT search variables that might be selected for a broad search for negative sentiment, including some of these specialized categories.

In a search for individuals exhibiting just Disgruntlement (feelings of anger, victimization, and identification of someone to blame) a WT user might request a narrower search based on a choice of negatives, negative evaluators, and references to the term "me." Second, the WT user might decide he would like to see those communications with the highest score on insider risk variables of concern and for this purpose WT produces a list of top scorers in these categories by email address. In a third approach, the WT user must decide which segment of the distribution of subject communications displaying the selected characteristics is of interest. For example, WT will automatically supply descriptive statistics for the distribution of the variable of interest and the user may request all emails above the mean in a very broad sweep or just those emails one standard deviation above the mean—a narrower and extreme group of communications. Often through a process of trial and error, the user familiar with the communication culture of their organization will need to balance the risk of missing true positives by using a narrower search against the risks of finding numerous false positives by using a broader search. In our earlier experience with the full Enron Archive, for example, we have found that a narrow search focusing on emails one-half a standard deviation above the mean value best balances these concerns. After WT returns all of the emails fitting the selected variables (and, optionally, ranks these email in order of concern), the user has several additional options. First, the user may wish to filter out emails containing the specified variables that are not of interest due to their content. Newsletters, sports discussions groups, forwarded articles, advertisements, spam, routine technical reports on system outages, and other non-germane material that often contains negative content may be filtered by using key words from the subject line. The user can also filter by date, author, recipient, or author affiliation as indicated by the email address. We recommend doing this after WT's initial installation in order to save analysts time later. The system will then offer the user several approaches to "drill down" into the communications discovered, including the ability to examine all of the email from a particular author by time, recipients, or other WT characteristics. A particularly useful WT function in investigations is to examine the email of an identified subject to determine the frequency and emotional tone of his correspondence with others in his communications network. This view of an author's email frequency and tone toward others can

map a communication network revealing internal conflicts and coalitions that may play a role in insider investigations.

1.4 WT limitations

Compared to human coders, WT currently has several limitations that may limit its effectiveness as a screening tool for locating communications displaying negative sentiment and insider risk. First, the current version of WT is not able to detect many of the more subtle variables described on the SNIT that imply, rather than overtly state, negative sentiment. These categories may include sarcasm, irony, non-negative statements denoting author upset ("there should be an accounting") or implying criticism, opposition, or protest indirectly ("I've done my best for the Company"). Second, if a specific dictionary term is not in one of WT's vocabulary lists it will not be detected. Unless one of WT's non-dictionary algorithms detects an alternative risk indicator, such words may not be correctly scored. As noted above, it is therefore critical to update WT vocabulary lists for use with different populations or groups. Third, the current version of WT has a limited ability to determine whether the negativity detected is directed toward an individual or group noted in the content or a less critical topic that does not necessarily constitute a risk. For example, WT will not be able to determine whether an author is enraged at his football team's performance or a coworker mentioned in the communication. This handicap can be expected to produce "false positive" results which may or may not be of interest. For example, some analysts may be interested in such an angry display, no matter whom the target. Subsequent modifications to WT using more advanced linguistic algorithms are expected to address this concern.

While we do not expect WT or another computerized system utilizing other approaches (e.g., Computational Linguistics) to perform as well as human coders, the initial or full scale review of the content of an email cache by human coders is not a practical or desirable alternative in most settings. We therefore sought to determine WT's relative strengths and weaknesses as an initial screening tool for this task so that its limitations may be understood by users.

1.5 Research Questions

As the broad range of WT variables in Table A1 in the Appendix indicates, there are many ways to use WT as a screening tool to examine an email cache for indicators of negative sentiment and insider risk. Specific WT variables may also be better suited for different criterion groups, such as WT variables for the detection of anger, depression, and anxiety with those respective cohorts. This part of the research was designed to address the following basic questions regarding WT's performance:

- What percentage of emails categorized as High, Medium, or Low on the SNIT contained in the Enron sample did WT correctly identify?
- How many additional emails did WT select in the process of identifying the above communications that were not identified by the SNIT (false positives) that an analyst using WT would have to review?
- What percentage of the emails categorized as High or Low on the SIRDC did WT correctly identify in the Enron sample?
- How many additional emails did WT select in the process of identifying these communications that were not identified by the SIRDC (false positives) that an analyst using WT would have to review?
- What percentage of the actual communications from documented insiders did WT correctly identify?
- How many additional emails did WT select in the process of identifying the actual insider's communications that were not from insiders (false positives) that an analyst using WT would have to review?
- What percentage of these "false positives" identified in the process of searching for actual insiders would also be of interest or concern to an analyst or investigator?
- To what extent is WT's scoring correlated with the SNIT and SIRDC in time series assessments of individual variations in risk with established perpetrators, and specifically, what is the correlation between SNIT and SIRDC scores and WT measures for the online stalker described in Shaw et al. (2013)?
- Given the results described in Shaw et al. (2013) indicating that communications containing insider risk indicators are more likely to be contained in emails Moderate-to-High in negative sentiment, is WT a useful instrument for identifying this target group?
- Does WT have any feature that can supplement the false positive problem associated with using negative sentiment alone to detect insider risk?

2. RESEARCH DESIGN

This assessment of WT was conducted with 1,000 randomly selected Enron emails described in Shaw et al. (2013). In addition, ten actual insider emails randomly selected from the principal investigator's case archive were also entered into the database, raising the number of emails to 1010. To test WT's performance, two different search strategies were used—a broad search corresponding to a sweep for negative sentiment, and a narrow search corresponding to a sweep for disgruntlement. The broad search for negative sentiment was conducted for emails containing any of the WT variables associated with negative sentiment measured above the group mean. The

narrower search was conducted for emails reflecting simple disgruntlement that were one-half a standard deviation above the mean. The selection of one-half a standard deviation above the mean for this search was a compromise between the risks of false and true positive results. We did not want to limit the search by filtering-out all emails one standard deviation above the mean risking the loss of actual insiders or true positives. On the other hand we did not want to burden analysts with an excessive number of false positives by simply reviewing all emails above the mean. We were also informed in our selection of search criteria by the earlier results in Shaw et al. (2013) indicating that most insiders and at-risk individuals will be located at High or Medium versus Low levels of negative sentiment or insider risk factors.

2.1 Broad Search for Negative Sentiment

Table 1 below summarizes WT's performance compared to the SNIT results. As the table indicates, WT performed extremely well at locating communications either High or Medium in negative sentiment as coded by the SNIT. WT did not perform well at locating emails scored Low on negative sentiment on the SNIT. WT missed 56% of Low Sentiment emails in its Negative Sentiment search mode, and 78% of emails scored Low in negative sentiment in its Disgruntlement search mode. However, we did not expect the narrow Disgruntlement search to function well in the selection of broader negative sentiment. As noted above, communications from this Low negative sentiment group are unlikely to contain seriously disgruntled individuals at-risk for insider acts. WT also produced a seemingly significant rate of false positives in both modes, displaying its limited ability to exercise the judgments attributed to human coders discussed above. Other approaches to these searches and future testing might improve these results. We also wanted to gauge the labor burden of filtering false positives on the analyst. Therefore Table 1 also contains the timed results for these searches.

Table 2 displays WT's performance compared to human coders utilizing the SIRDC. As the table indicates, WT did extremely well at locating emails coded High in Insider Risk. WT performed less well at locating emails Low in Insider Risk as coded by the SIRDC. In the Negative Sentiment search mode, WT missed 35% of Low SIRDC emails and 59% in Disgruntled search mode. However, we did not expect a search for negative sentiment to do well in locating SIRDC subject communications.

Table 1 WT accuracy compared to SNIT: 220 of 1010 randomly selected Enron emails identified as High, Medium or Low in negative sentiment;
WT searches using sweeps for negative sentiment and disgruntlement

% Subjects Scoring High, Medium or Low on SNIT Identified By WT by Search Type	**High SNIT Score N=6**	**Medium SNIT Score N= 9**	**Low SNIT Score N=207**	**False Positives or Emails identified by WT with no SNIT Score**
WT Negative Sentiment Search Algorithm	100% 6/6	100% 9/9	43% 89/207	1010 Emails were reduced to 180 emails or 18% of total cache after search and filtering. Analyst had to "review" 155 extra emails to locate 15 High or Medium in Negative Sentiment and 89 Low in Negative Sentiment (this review took approximately 35 minutes).WT missed 116 emails Low in Negative Sentiment or 56% of the Low Negative Sentiment Group
WT Disgruntlement Search Algorithm	100% 6/6	100% 9/9	22% 46/207	1010 Emails were reduced to 67 after search and filtering or 7% of the cache. 6 of these 67 emails or 9% were "false positives." Analysts had to review 6 extra emails to locate 15 in the High or Medium SNIT group and 46 in the Low SNIT group (this took approximately 4 minutes). WT missed 78% of the Low SNIT emails.

Table 3 below displays WT's performance detecting the emails from the ten actual insiders inserted into the Enron cache. As Table 3 indicates, WT in Disgruntlement mode captured all 10 of these insiders. However, we also examined whether WT's effectiveness in detecting actual insiders would be reduced if we used only negative sentiment indicators for the search. As the second row of Table 3 indicates, the removal of the term "me," a WT measure of perceived Victimization that does not represent an overtly negative sentiment, reduced the effectiveness of the search result by 30%.

Table 2 WT accuracy compared to SIRDC: 36 of 1010 randomly selected Enron emails identified as High or Low in insider risk;
WT search using negative sentiment and disgruntlement algorithms

% Emails Scoring High or Low on SIRDC Identified by WT Search Type	**High SIRDC Score N=2**	**Low SIRDC Score N=34**	**WT Results and False Positives and Missed Communications**
Negative Sentiment Search Algorithm	100% 2/2	65% 22/34	1010 Emails reduced to 180 after Search and Filtering or 18% of cache. 155 of 180 (86%) were "false positives" so that analyst had to review 155 extra emails to locate 24/36 containing insider risk. WT missed 12 emails or 35% of emails Low in Insider Risk
Disgruntlement Search Algorithm	100% 2/2	41% 14/34	1010 Emails filtered to 67 or 7% of Cache. 51 emails or 76% of these were false positives. Analyst had to search 51 extra emails to locate 16 containing insider risk. WT missed 20 Low Insider risk emails or 59% of Low SIRDC emails

Table 3 refers to the false positive rate of WT in quotations because although analysts would have had to review up to 57 extra emails to locate the ten insiders, a significant number of these emails would not be considered false positives by analysts concerned about insider risk. For example, of the 57 emails reviewed to locate the ten insiders, 70% had SNIT scores and 21% had SIRDC scores. Examples of such "false positives" in the search for true insiders are displayed in Table A2 in the Appendix, along with their subject headers, SNIT, and SIRDC scores. Examples of these "false positives" include content concerning interpersonal and professional disgruntlement, possible fraud, and injustice attributions. Legal and security staff at several commercial and government offices where WT is being test bedded have reported that the damage and labor costs associated with insider activities has been so severe that the additional labor involved in screening for false positives is well worth the effort when true positives can be located.

WT Search Type	10 Insider Emails	WT Results and "False Positives"
Search for Disgruntlement With Victimization	10/10 (100%)	1010 Emails reduced to 67 after search and filtering or less than 7% of cache. Analyst had to search 57 extra emails to locate 10 True Positives (which took approximately 12 minutes).
Search for Disgruntlement Without Victimization	7/10 (70%)	1010 Emails reduced 112 or 11% of the cache and then 66 with filtering. Analysts had to review 59 additional emails to locate 7 true positives. WT missed 3 or 30% of true positives without Victimization.

Table 3 Percentage of true insider emails located by WT by search type

2.2 WT Correlation with SNIT and SIRDC in Time Series Evaluation of the Online Stalker

Although WT may be relatively less sensitive to Low levels of negative sentiment and insider risk, we examined whether WT measures correlated with SNIT and SIRDC results over time across both High and Low levels when tracking the emotional state and risk of an actual insider. While WT's relative insensitivity to Low levels of negative sentiment and insider risk currently limits its usefulness to detect these groups, WT could still be used to monitor previously identified subjects at any level if its measures parallel those of the SNIT and SIRDC.

Figure 1 below displays the original SNIT and SIRDC values for the 17 stalker emails described in Shaw et al. (2013) with the WT Disgruntlement measure used in this research for the narrow search. For the purpose of comparison in this case study, the WT, SNIT, and SIRDC values have been normalized per 100 words to better control for the wide variations in the length of the stalker's emails. As Figure 1 portrays, the WT variable of Disgruntlement (composed of Negative Evaluators, Negatives, and Me) tracked closely with the SNIT and SIRDC over the 17 emails, peaking just prior to the physical attack and subsiding afterwards. Disgruntlement was highly correlated with both the SNIT ($r=.867$, $p<.01$, bilateral) and the SIRDC ($r=.628$, $p<.05$, bilateral) across the 17 emails.

It was interesting to note WT's lack of correlation with the SNIT and SIRDC on Stalker email number 9, which read simply "Your credibility is gone here. You should follow." While human coders easily picked up the negativity in

these statements, the search version WT variables used for this research did not include the WT variable Direct References ("you'), which would have scored both these phrases. However, WT is not yet programmed to detect the negativity communicated by the negative reference to "credibility" as "gone" or to "you" and "should leave." Further analysis is underway to understand the relative divergence between the WT and SNIT and SIRDC scores. In particular we hope to explore whether WT's Disgruntlement algorithm is more sensitive to some forms of insider risk than the SNIT or SIRDC, whether these divergences are false positives (WT is scoring terms that do not necessarily indicate risk), or whether WT is missing true positive indicators as in the examples above.

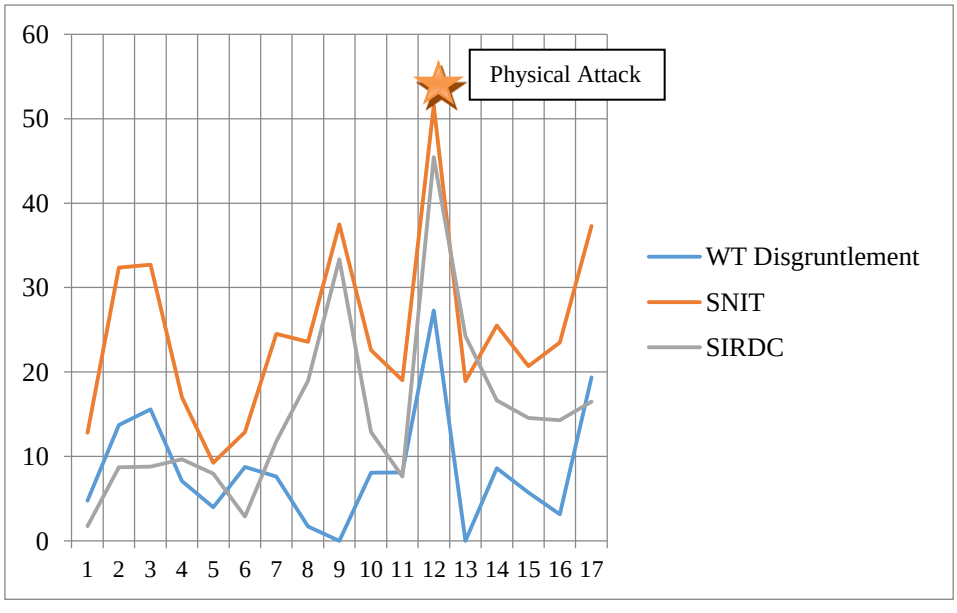

Figure 1 Comparison of Normalized SNIT, SIRDC and WT Disgruntlement Scores across 17 Stalker Emails

3. CONCLUSIONS

WT appears to be an effective means for locating communications scored High or Medium in negative sentiment by the SNIT or High in insider risk by the SIRDC. WT proved less effective in locating emails Low in negative sentiment on the SNIT and Low in insider risk on the SIRDC. However, WT performed extremely well in identifying communications from actual insiders randomly selected from case files. In addition, it appeared that WT's measure of perceived Victimization was a significant supplement to using negative sentiment alone, when it came to identifying actual insiders. Finally, WT's

correlation with the SNIT and SIRDC over time in the online stalker case indicates that its sensitivity to changes in psychological variables is closely correlated with these human ratings and that it is also sensitive to changes over time associated with violence risk.

3.1 Implications for Detecting Communications Containing Insider Risk

While we have not yet identified a single target or criterion group for searches to identify individuals at-risk for insider issues, there has been some progress in this regard. Based on the preliminary findings from Shaw et al. (2013), individuals likely to be at-risk for insider actions may present psycholinguistic characteristics such as High or Medium levels of negative sentiment, as measured on the SNIT; High levels of insider risk indicators, as measured on the SIRDC; and some level of perceived victimization or mistreatment. As noted in Shaw et al. (2013), the frequency of insider risk subsides as negative sentiment declines.

WT was not designed as a single tool for identifying disgruntled individuals at risk for insider actions. Rather, it was designed to be used in conjunction with other tools (such as technical anomaly detection measures) and more complex models of insider risk, to serve as an initial screen to help narrow a field of subjects to reduce the false positive rate associated with locating individuals at-risk for insider violations. In this assessment of WT for this purpose, the system performed extremely well in identifying communications determined by human coders to be High and Medium in negative sentiment and High in insider risk factors. It also performed extremely well in identifying communications from established insiders. However, WT does not appear to be an effective means to identify and measure the full range of negative sentiment—it does not perform well compared to human coders at the Low range of negative sentiment.

However, results from Shaw et al. (2013) and this research indicate that this relative weakness may not impair WT's usefulness for identifying communications containing significant indications of insider risk because of the very low base rate of insider risk at Low levels of negative sentiment. WT may not be effective for early identification of persons with Low levels of negative sentiment that may turn into individuals at-risk for insider activity. However, the low base rate of 16.3% for communications with negative sentiment that also contain insider risk and the exclusively low insider risk scores within this group, indicate that the vast majority of these subjects present either little or no risk of insider actions. It may be unethical to target these individuals without statistical evidence that they have such risk factors. Further time series research will be necessary to determine whether this group Low in negative sentiment and insider risk ever converts to more concerning risk levels. Or, whether this low level of negative sentiment is a common

manifestation of discontent present in many groups. However, WT's relative lack of sensitivity to lower levels of negative sentiment and insider risk in search mode would not limit its use in monitoring previously identified individuals with higher levels of these risk factors or other sources of concern regarding risk. This appeared to be true in the case study of the online stalker presented above, where WT's measure of disgruntlement followed, and was correlated with, observer ratings using the SNIT and SIRDC at both high and low levels of negative sentiment and insider risk.

Although many of the "false positives" acquired in the successful search for actual insiders in this experiment were shown to be true positives for other forms of insider risk, WT still produced fairly high rates of false positives that could burden analysts. We have tried to communicate the extent of this potential burden by including data on the actual time taken by our users to filter false positives. As further research and development proceeds to address this problem, we again recommend the use of WT in an integrated multi-disciplinary array of detection methods that will serve as an initial screen to narrow the search for at-risk individuals using other forms of behavior associated with this target group.

ACKNLOWLEDGEMENTS

****This work was supported by a grant from the Department of Defense; however, the findings, conclusions and opinions expressed are solely those of the authors. We also want to acknowledge the support of readers Drs. Marcus Rogers and Stephen Band.****

REFERENCES

Hermann, M.G. (1980). Explaining foreign policy behavior using the personal characteristics of political leaders. *International Studies Quarterly*, *24*: 7-46.

O'Brien, T. (2005). The rise of digital thugs. *New York Times*, August 7, 2005, Business section, p. 1.

Shaw, E. D., Payri, M., Cohn, M., and Shaw, I. R. (2013). How often is employee anger an insider risk I? Detecting and measuring negative sentiment versus insider risk in digital communications. *Journal of Digital Forensics, Security and Law*, *8*(1): 39-71.

Shaw, E. D. (2006). The role of behavioral research and profiling in malicious cyber insider investigations. *The International Journal of Digital Forensics and Incident Response*, *3*: 20-31.

Shaw, E. D., and Stock, H. (2011). Behavioral risk indicators of malicious insider theft of intellectual property: Misreading the writing on the wall. Symantec Corporation, White Paper. Retrieved from

http://investor.symantec.com/phoenix.zhtml?c=89422&p=irol-newsArticle on December 7, 2011.

Shaw, E. D., and Stroz, E. (2004). WarmTouch software: Assessing friend, foe and, relationship. In T. Parker (Ed), *Cyber adversary characterization: Auditing the hacker mind.* Maryland Heights, MO: Syngress Publications.

Shaw, E. D., and Wirth-Beaumont, E. (2004). *WarmTouch codebook and algorithms*. Shaw Stroz LLC.

Shaw, E. D. (2003). Saddam Hussein: Political psychological profiling results relevant to his possession, use and possible transfer of weapons of mass destruction (WMD) to terrorist groups. *Studies in Conflict and Terrorism, 26*: 347-364.

Weintraub, W. (1986). Personality profiles of American presidents as revealed in their public statements: The presidential news conferences of Jimmy Carter and Ronald Reagan. *Political Psychology, 7*: 285-295.

Weintraub, W. (1989). *Verbal Behavior in Everyday Life*. Springer.

Winter, D. G, Hermann, M. G, Weintraub W, and Walker, S. G. (1991). The personalities of Bush and Gorbachev measured at a distance: Procedures, portraits and policy. *Political Psychology, 2*: 215-245.

APPENDIX

Table 1 Description of WT Psycholinguistic Search Variables

Variable	Variable Description and Examples
Negative Evaluators	Includes negative or pejorative judgment words, not an affect or a negation; describes characteristics of a person, group, action, or situation– Stupid; abuse; psycho, bad, poor, loser
Negative Feelings	Affectively negative response or experience–Aggravate; cranky; rage
Negatives	A term that specifically opposes, negates or contradicts– No, not, never, words with n't suffix (won't) contradict
Adverbial Intensifiers	Adverbs or other modifiers that increase the power of a statement–So, very, really, too
Anger (algorithm)	A combination of negative terms such as negative evaluators and feelings along with personal pronouns corrected by content that modifies or ameliorates the negative statement
Anger (vocabulary)	A list of terms associated with the expression of anger–Hate, angry, dislike, pissed, irritated
Anxiety (algorithm)	A combination of retractors + qualifiers + neg feelings + explainers
Anxiety (vocabulary)	A list of terms associated with anxiety-Worried, concerned, afraid
Depression (algorithm)	A combination of I + me + negative feelings + negative evaluators + adverbial intensifiers
Depressed (vocabulary)	A list of words associated with depression--Sad, blue, bummed, depressed, guilty, despondent, crushed, heart-broken
Instrumental Aggression	Terms referring to weapons or violent means, shows instrumental intention to harm; distinguished from anger in that anger involves emotion, rather than intention with method
Victimization	words describing an author's belief that they have been the subject of unfair, unjust, biased or otherwise prejudicial or persecutory actions
Trapped	words describing an author's psychological state in which he or she believes their options are very narrow and they have little or no choice but to act
Sexuality	references body parts, sexuality or sexual acts that also tend to reduce persons to objects
Dehumanization	Includes disparaging remarks, disparaging names or slurs that treat target as object, inhuman, making attack easier
Disgruntlement (algorithm)	A combination of negatives for anger, me for victimization and direct references indicating the possible presence of someone to blame
Disgruntlement (Long) (algorithm)	Disgruntlement above plus additional variables that capture added personalization (I), added anger (negative evaluators and feelings)

Table 2 Examples of "False Positives" in WT Search for True Insiders by Subject Line, excerpt, SNIT and SIRDC Score

Subject Line	Excerpt	SNIT Score	SIRDC Score
☹	I'm so depressed about the whole thing. Rayfael feels bad too, but it's important he not be near death anytime we're together…He feels like the bad guy. And I feel like a bad father!!!	54	5.67
Organizational Announcement	It is with great regret that I announce that Jeff Skilling is leaving Enron…I regret his decision…our stock price has suffered substantially…	13	2.0
Demand Kenneth Lay donate proceeds from Enron Stock Sales	…while you netted well over $100 million many of Enron's employees were financially devastated when the company declared bankruptcy and their retirement plans were wiped out	26	12.33
No subject	Looks like I'll be rotating out of this job just in time. Jeff said Louise is just like Shankman. Yikes…We've decided not to tell Binh anything about her. Lavo stresses her out…	17	1.33
ALSO URGENT AND CONFIDENTIAL	Diomedes is going to jump all over me...the implications are much broader and we don't want to come off looking dishonest to our own people. And, believe me the rumors are out there. I just don't want to make it worse…	32	6.67
Re: Directions	Damn'it Jeff. I don't have time…The panic is my waste of time trying to get things organized. That's the panic…	34.67	0.0
Re: Hey	You must be having a bad day!!!But you know it's not my fault…	15.0	0.0
Re: Pacific Virgo	I am concerned about this expert's conclusions, the most troubling of which is that our putting the ship on notice of the end use of the cargo would have been of so little importance	10.0	2.67
Re: PHC Statement	I DO NOT SEE HOW WE CAN SUPPORT PRICE CAPS… If we could not support them on the wholesale side, how can we support on the retail side? Doesn't it kill the ESP business? WE MUST STICK WITH THE PRINCIPLES WE FOUGHT FOR ALL THESE YEARS—SOME THINGS ARE JUST WORTH FIGHTING FOR…the result of the knee-jerk, weekly about-face by the CPUC which cannot possible be considered to be reasoned…	84	5.0