Digital Investigation 22 (2017) 142–146

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

The use of communicated negative sentiment and victimization for locating authors at-risk for, or having committed, insider actions

Eric Shaw[*], Maria Payri, Ilene Shaw

Stroz Friedberg/Aon, Washington, DC, 20036, USA

ARTICLE INFO

Article history:
Received 30 November 2016
Received in revised form
21 June 2017
Accepted 22 June 2017
Available online 24 August 2017

ABSTRACT

This article examines the challenge of locating persons at-risk for insider actions through communications by identifying psychological states and attributions associated with past insider violations. We specifically seek to expand and replicate two earlier studies (Shaw et al., 2013a, 2013b) which examined the relationship between negative sentiment and insider risk and the utility of psycholinguistic software targeting negative sentiment and feelings of victimization for locating communications of at-risk criterion groups within an organization's communication cache. The first study attempted to replicate Shaw et al. 2013a, in which two new rating tools for negative sentiment and insider risk were applied to 1000 randomly selected Enron archive communications to determine the frequency and overlap of messages containing negative sentiment and insider risk. In the present work the data set was expanded from 1000 to 10,000 messages. The second study attempted to replicate Shaw et al. 2013b by inserting 100 communications from authors at-risk for, or having committed, insider acts into the expanded Enron sample. The software used to operationalize the search for indicators of negative sentiment and victimization located 95% of the criterion group emails and all of the targeted authors inserted into the Enron sample. In Study 3, the repetition of this search using only indicators of negative sentiment caused a significant decrement in search results, reducing the percent of located emails from 95% to 58%. This result supported earlier findings (Shaw et al. 2013a, 2013b) emphasizing the dangers of using negative sentiment alone to locate indicators of insider risk. Finally, we described the true and false positive rates obtained when an enterprise version of the psycholinguistic software was deployed to locate similar at-risk communications in an actual mail cache containing over 50 million messages from over 63,000 Senders. The implications of all three studies and this field deployment for the relationship between negative sentiment and insider risk and the ability of investigators and analyst to locate potential insiders within organizational communication caches are discussed.

Background

Insider violations in industry and government pose an increasing concern to security, law enforcement, compliance and counter-intelligence staff, as well as the leaders of these organizations who answer to their respective constituents and the public. It is extremely rare to find an insider case that has not involved the use of information systems to plan, access, steal or exfiltrate the information involved, whether it is corporate or government secrets, or employee or customer information. It is equally rare to not find evidence of the risk of impending violence committed by insiders stored in relevant information systems. Shaw and Sellers (2015) have described the Critical Pathway to Insider Risk framework, which summarizes the "path" travelled by corporate and government insiders who have committed these violations with emphasis on early manifestations of disgruntlement within the organization.

While many types of technical indicators of insider risk exist, these systems tend to focus on anomalous employee use of information systems and related behaviors (hours of attendance, copying or downloading, patterns of information accessed and sent, etc.). Unfortunately, the fact that most insiders steal or damage information to which they have authorized and regular access and the amount of information and false positives these systems produce have limited their effectiveness. These systems also do not address the underlying psychological states, personality and decision-making by these individuals that accompany increased risk.

[*] Corresponding author. 5225 Connecticut Avenue NW, Suite 514, Washington, DC, 20015, USA.
E-mail address: eshaw@msn.com (E. Shaw).

http://dx.doi.org/10.1016/j.diin.2017.06.014
1742-2876/© 2017 Elsevier Ltd. All rights reserved.

Until recently, there have been no automated systems designed to detect the negative emotions and attitudes that are frequently associated with insider actions, despite the fact that employee frustration and anger have long been associated with aggression and violence in the workplace (Glomb and Liao, 2003; Hershcovis et al., 2007; Hershcovis and Barling, 2010), as well as turnover, absenteeism, accidents on the job, alcohol consumption, and other high-risk health behaviors (O'Neil et al., 2009). Holton (2009) also found an association between anger and fraud, and Band et al. (2006) found similar links to sabotage and espionage. Occupational health researchers who study a range of counter-productive work behaviors (CWBs), from taking long lunches to workplace violence, have consistently found a strong link between negative emotions and CWBs (Sakurai and Jex, 2012; Dalal, 2005; Brief and Weiss, 2002; Schat and Kelloway, 2005). More recently, Taylor et al. (2013) examined language associated with simulated insider activities and found that assigned insiders used language indicating greater self-involvement and more negative emotion compared to themselves prior to the simulated insider assignment and compared to a control group not assigned to simulate insider activity.

Ideally, a combination of detection approaches sensitive to both anomalous system behavior as well as the underlying disgruntlement that may motivate such violations might improve our ability to prevent, detect and intervene in insider risk cases. For example, information security personnel with numerous technical risk indicators might better prioritize the limited resources available to investigate these leads by starting with persons who also display signs of disgruntlement.

However, a critical but unanswered question for investigators and analysts concerned with the possible psychological precursors of insider violations is what percent of communications containing negative sentiment also contain indicators of insider risk. Are communications with negative sentiment or other alterations in language associated with insider activities a pathway for locating disgruntled at-risk individuals or a wild goose chase of false positives and unethical invasions of privacy?

Previous research on locating individuals at-risk for insider acts from their communications

Only a few previous studies have addressed the issue of what to look for in employee communications as an indicator of insider risk and whether insider risk can be effectively differentiated from negative sentiment in general. For example, Shaw et al. (2013a,b) introduced two new scales for the identification and measurement of negative sentiment and insider risk in communications by actual insiders, in order to examine the unexplored relationship between these two constructs. The inter-rater reliability and criterion validity of the Scale of Negativity in Texts (SNIT) and the Scale of Insider Risk in Digital Communications (SIRDC) were established with a random sample of emails from the Enron archive and a criterion group of established insiders, disgruntled employees, suicidal, depressed, angry, anxious, and other sampled groups. In addition, the sensitivity of the scales to changes over time as the risk of digital attack increased and transitioned to a physical attack was also examined in an actual case study. Interrater reliability for the SNIT was extremely high across groups (.944, $p \leq 0.05$) while the SIRDC produced lower, but acceptable levels of agreement (.823, $p \leq 0.05$). Both measures also significantly distinguished the criterion groups from the overall Enron sample. The scales were then used to measure the frequency of negative sentiment and insider risk indicators in the 940 random Enron email sample and the relationship between the two constructs. While low levels of negative sentiment were found in 20% of the sample, 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 sample. Of the 222 emails containing negative sentiment in the sample, only 36, or 16.3%, also displayed any indicators of insider risk. The odds of a communication containing insider risk increased with the level of negative sentiment and only low levels of insider risk were found at low levels of negative sentiment. All of the emails found to contain insider risk indicators on the SIRDC also displayed some level of negative sentiment. While this research established the relative rarity of negative sentiment and signs of insider risk, as well as the differences between the two constructs, it involved a relatively small sample size.

Shaw et al. (2013a,b) subsequently tested the effectiveness of a psycholinguistic software program[1] previously used for investigations for locating the full range of these communications. After significant testing to determine the correct combination of psychological states and attributions (linguistic indicators of anger, blame, victimization) an additional randomized sample of communications from actual insiders and persons at-risk of insider actions previously coded for their SNIT and SIRDC values were inserted into a sample from the Enron archive to determine the software's ability to locate communications high and low in negative sentiment and insider risk. The software proved less effective in locating emails Low in negative sentiment on the SNIT and Low in insider risk on the SIRDC. However, the software 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 the software'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., 2013a,b) indicate that this relative weakness in identifying Low levels of negative sentiment may not impair the software'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.

This preliminary review indicated that the software may not be effective for early identification of persons with Low levels of negative sentiment that may subsequently turn into individuals at-risk for insider activity. The low base rate for insider risk measured on the SIRDC of 16.3% for communications low in negative sentiment and the exclusively low level of insider risk contained in these emails indicates that the vast majority of these subjects present either little or no risk of insider actions. 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. Until such time, there are significant validity, resource allocation and ethical questions surrounding a focus on such individuals. The software'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 any level of risk or other sources of concern. 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, the software still produced fairly high rates of false positives that could burden analysts. An informal survey of the true positive rates of conventional insider detection software solutions focusing on technical anomalies and use of limited key words

[1] For more information on WarmTouch software (subsequently renamed Scout) see Shaw and Stroz (2004).

indicates that even with this false positive rate, this approach represents a significant improvement from current detection methods. However, these findings were based on a relatively small sample of criterion communications placed within a relatively small communications archive.

The current research sought to replicate earlier findings regarding the relative distribution of negative sentiment and insider risk in a larger, randomly selected data set and test the ability of psycholinguistic software to identify communications containing insider risk in two other studies. In the first study, we attempted to challenge our previous research on the basic frequency of negative sentiment and insider risk in a larger, randomly selected data set (10,000 versus 1000 communications). Once "ground truth" in terms of the frequency of negative sentiment and insider risk was established for this larger data set, the Criterion groups utilized in Shaw et al. (2013a,b) were inserted into this larger Enron email cache. We then attempted to test the software's ability to locate Enron employee communications with negative sentiment and insider risk, as well as the inserted communications from actual insiders and at-risk groups. We also examined the relative effectiveness of measures of victimization as a supplement to negative sentiment alone, in improving the ability to locate communications from actual insiders and criterion groups.

Study one

Method

Ten thousand randomly selected emails from the Enron archive were ingested into the psycholinguistic software and after the elimination of duplicates, 9703 messages remained from the period March 1997 to November 2002—a difficult time at Enron. These messages were further reviewed by a trained coder who divided them into two groups—messages containing and not containing personal content. Messages not containing personal content included those with only corporate disclaimers, auto-generated content such as an Outlook Calendar notice, and advertising or other promotional materials. Any other email containing any personal content was included in the second group and reserved for SNIT, SIRDC coding and psycholinguistic analysis. The personal content in these emails ran from simple direct statements of fact such as "the meeting is cancelled" to more personal and intimate exchanges. Table 1 below displays the results of this division. It was interesting to note that less than half of the randomly selected messages contained personal content.

The emails containing personal content were then scored for SNIT and SIRDC values following the same procedure described in Shaw et al. (2013a,b), completed by a member of the original coding team.

Results

Table 2 shows the distribution of SNIT and SIRDC scores within the Enron emails with personal content. Of those emails coded as personal, 921 (20.27%) had negative content, and only 32 (.70%) of the personal sample were given a score for insider risk. As we can see in the last column, the software performed very well when identifying emails containing insider risk (it flagged 96.87% of these messages).

Table 1
Groups with and without personal content.

Content category	Number of emails per group	% of total
1—No Personal Content	5160	53.18
2—Contained Personal Content	4543	46.82

Table 3 shows the distribution of the 921 messages with SNIT scores across High, Medium and Low score groups. As expected, the majority of emails with negative sentiment had a low SNIT score (90.98%), while 6.95% had a medium SNIT score, and only 2.06% had a high SNIT score. The second column displays the number of these messages with SNIT scores that also contained messages with signs of insider risk on the SIRDC, accompanied by the software hit rate for each group. . As expected, all of the emails identified as containing insider risk within the Enron sample had a relatively low SIRDC score. However, with the exception of two emails, all of the communications with SIRDC scores were sent after November 2001, shortly after the Enron scandal was revealed. In fact, 30 out of 32 SIRDC emails are personal communications from employees addressed to Kenneth Lay, CEO and Chairman of Enron, and Jeffrey Skilling, who followed him in that post. As noted in previous findings, negative sentiment occurs independently of insider risk in this sample—only two of the 19 communications with High SNIT scores contained evidence of insider risk.

Also as previously found, the psycholinguistic software performed significantly better when identifying SIRDC emails that had either a medium or a high SNIT score. The one email that the software failed to identify had a low SNIT score and only two lines of text. While the phrase "You're next" was readily identified by the human coder as threatening, this early iteration of the software did not identify and score this phrase.

Experiments 2 and 3: using negative sentiment with and without victimization to locate criterion group communications within 10,000 enron messages

Method

The 4543 messages containing personal content were then processed by the software by asking the system to filter out any communications which were not at least a half standard deviation above the group mean for all 4543 messages in terms of the number of:

- uses of the term "me;"
- Negatives (words such as not, never, no or containing "n't") indicative of anger or negativity;
- Negative Evaluators (words reflecting negative judgments or beliefs); and
- Negative Feelings (negative emotions).

These four psycholinguistic variables were selected to represent our search framework of Disgruntlement. "Me" was included to represent a sense of Victimization as it is very difficult for an author to use this term without being the object of action by others ("why did you do that to me?" "Are you talking to me?"). At higher rates of use, we have found this variable to be a good indicator of feelings of Victimization. The inclusion of this pronoun also tends to guarantee that the message will contain some level of personal content. Negatives, Negative Evaluators, and Negative Feelings were included to capture negative sentiment.

Results

Initially, this search filtered the original cache down from 10,100 emails to 1292 or about 13% of the total. After additional filtering using subject headers for e-commerce, newsletters, sports chat and routine reports (system outages) the emails for review were further reduced to 857 communications or 8.5% of the original cache.

Table 2
Distribution of SNIT & SIRDC Scores in 4543 Enron emails with personal content.

Emails with negative sentiment on SNIT	Emails with insider risk on SIRDC	Emails with SIRDC detected by software
921 (20.27%)	32 (0.70%)	31/32 (96.87%)

Table 3
Distribution of communications with SNIT and SIRDC scores.

SNIT Levels[a]	H, M, L Distribution of messages with SNIT scores	Number of SIRDC messages by H,M,L SNIT score and WT hit rate (SIRDC)[b]
High SNIT	19	2 (100%)
Medium SNIT	64	29 (100%)
Low SNIT	838	1 (0%)

[a] SNIT Levels: High (≥31); Medium (16–30); Low (≤15).
[b] All SIRDC emails had a low score (≤9).

Table 4
WT hit rate for criterion group emails with & without victimization by author.

Author type	Format	Number of communications	Hit rate by author	Hit rate without victimization
Online Stalker	email	17	14 of 17	4 of 17
Bruce Ivins	email	5	5 of 5	1 of 5
Depressed	chat	10	10 of 10	9 of 10
Angry	chat	10	9 of 10	9 of 10
Financial Stress	chat	10	9 of 10	4 of 10
Suicide	chat	10	10 of 10	9 of 10
Substance Abuse	chat	10	10 of 10	5 of 10
Work Stress	chat	10	10 of 10	4 of 10
Actual Insiders	email	13	13 of 13	9 of 13
Abu Jihad (Paul Hall)	email	1	1 of 1	1 of 1
Bradley Manning	chat	3	3 of 3	2 of 3
Greg Smith	op-ed	1	1 of 1	1 of 1
Total Communications		100	95%	58%

Table 4 below displays the software "hit rate" for the criterion emails by author. The system succeeded in capturing 95% of the criterion group communications and it took the analyst just under 2 h to sort through the "false positives" or emails that were not from the criterion groups. However, a large percentage of these false positive were true positives for insider risk involving significant disgruntlement among Enron employees. While the software missed five percent of the criterion emails, it did successfully locate communications from each author—a fairly successful performance for any form of psychological screening tool, which demonstrates the potential value of the psycholinguistic approach to locating individuals at-risk.

In Experiment 3, we also wanted to assess the relative contribution of the concept of victimization to the effectiveness of the search for communications reflecting insider risk. The final column in Table 4 displays the same search results when the variable "me" was omitted from the search criteria. Weintraub (1986, 1989) found "me" to be an excellent indicator of passivity in patient and leadership populations. At higher levels we have found it to be a good indicator of victimization, present in the correspondence of many insiders who feel taken advantage of by their organizations and strike back. As the final column in Table 4 indicates, there was a significant decline in search effectiveness associated with the removal of this term from the filter.

Conclusions

These experiments confirmed earlier findings regarding the relatively low frequency of negative sentiment and insider risk in organizational communications and the independence of expressions of negative sentiment versus insider risk in these communications. It also confirmed earlier findings of the relative efficacy of psycholinguistic approaches in locating communications reflecting elements of disgruntlement and the importance of search methods containing both author negative sentiment and feelings of victimization in these efforts.

Our current, larger random sample from the Enron archive contained significantly fewer messages with negative sentiment than our earlier, much smaller Enron sample (9.5% versus 23.6%). We suspect this may be due to the random selection of more messages after the scandal became public as manifested in the findings above. However, the percent of messages with negative sentiment that also contained signs of insider risk was strikingly similar in these two samples with the earlier, smaller sample containing 3.4% compared to 3.8% in the current sample. Our current findings also confirmed earlier results indicating that while all messages containing insider risk contain negative sentiment, only a small percentage of communications with negative sentiment contain signs of insider risk.

This psycholinguistic approach proved an effective initial screen to assist analyst efforts to locate communications from actual insiders and individuals at-risk for insider actions. By reducing an initial email cache from 10,100 to 857 communications, the system saved an analyst significant time. Although the software "missed" five of the 100 communications contained in the criterion group cache, it identified every author through other additional writings. The third experiment in this series also highlighted the relative importance of measures of victimization for identifying "symptoms" of insider risk versus negative sentiment alone, confirming earlier findings by Shaw et al. (2013a,b), indicating that negative sentiment alone may present a relatively ineffective means for locating persons at-risk for insider actions. We have also found that subject attributions of blame further increase the effectiveness of our filters.

Table 5
SCOUT search Categories by results (N = 50,938,810 messages from 69,328 senders).

Search Category	Mean	SD	Search	# Messages Remaining	% Messages Remaining	# Senders Remaining	% Senders Remaining
Me	0.06	1.7	>7	10,999	0.02	>1000	n/a
Negatives	2.9	2.6	>2	10,270	0.02	>1000	n/a
Victimization	3.1	3	>3	2156	0.00	422	0.60
Employment	2.5	1.6	>4	383	0.00	137	0.20

It should be noted that there were still more than 725 “false positives” that required 2 h for the analyst to review. This puts the software's false positive rate[2] at .84 for the current sample of corporate emails and true insiders. On a large organizational email system this would scale to considerable analyst effort. With the use of improved statistical modelling and machine learning approaches we hope to reduce the false positive rate further along with the human labor required for determining whether the language present represents an indicator of insider risk.

Discussion

Psycholinguistics in the field: effectiveness and privacy

As of this writing we have deployed beta versions of the enterprise system at four organizations, including private and government facilities. In the process, we have encountered several practical and ethical issues that may have crossed the mind of many readers.

The first issue we encountered were questions about the software's potential for misuse in targeting innocent individuals, along with potential invasions of privacy. Table 5 below contains results for a large organization that addresses these issues. Starting with over 50 million email, chat and text messages from over 69,000 Senders the system uses a purely statistical search strategy to select communications that significantly differ from the group mean for human review. Table 5 illustrates how the search tactics are derived from the group mean and standard deviation for psycholinguistic representations of Victimization (Me and Victimization vocabulary) and negative sentiment (Negatives). The use of these filters reduced the communications to be reviewed to 2156 communications from 422 Senders. These communications still contained samples with marital and couple's conflict which without elements of threatened violence or significant financial distress are not of interest to organizational clients in these cases. To further refine the search the system was asked to eliminate communications which did not refer to employment with a frequency greater than a standard deviation over the mean for group references to Employment. This brought the number of messages actually reviewed by a trained clinician down to 383 from 137 Senders. From a privacy and ethics perspective, less than .0007% of messages from .2% of the 69,000 Senders are reviewed and these are examined solely on the basis of their statistical relationship to their peers.

This sample of communications from these 137 Senders were reviewed by a clinician trained in remote and risk assessment methods. Thirty-six percent of these 137 Senders were referred for a more complete investigation based on concerns regarding insider risk raised by the sample communications, while 38% were referred for monitoring after this review. With the addition of this human analyst in the review chain this finding indicates a 24% false positive rate for this stage of the risk assessment process.

It should be noted that we consider this psycholinguistic system to be part of an array of tools to screen and detect individuals at-risk within an organization. Other methods, such as anomalous network and interpersonal behavior are also necessary information sources. But the studies presented here allow us to be optimistic about the effectiveness of psycholinguistic assessment and its almost human sensitivity. While computerized psycholinguistic approaches still struggle to detect sarcasm and irony, our current research focusses on using the large samples of established insider communications we have gathered and advanced statistical and machine learning methods to supplement this initial approach.

[2] FP/FP + TN Where FP represent the total number of false positives and TN the number of true negatives.

References

Band, S., Cappelli, D., Fischer, L., Moore, A., Shaw, E., Trezciek, R., 2006. Comparing Insider it Sabotage and Espionage: a Model Based Approach, Technical Report. Software Engineering Institute, Carnegie Mellon.

Brief, A.P., Weiss, H.M., 2002. Organizational behavior: affect in the workplace. Annu. Rev. Psychol. 53, 279–307.

Dalal, R.S., 2005. A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. J. Appl. Psychol. 90, 1241–1255.

Glomb, T.M., Liao, H., 2003. Interpersonal aggression in work groups: social influence, reciprocal, and individual effects. Acad. Manag. J. 46, 486–496.

Hershcovis, M.S., Barling, J., 2010. Towards a multi-foci approach to workplace aggression: a meta-analytic review of outcomes from different perpetrators. J. Organ. Behav. 31, 24–44. http://dx.doi.org/10.1002/job.621.

Hershcovis, M.S., Turner, N., Barling, J., Inness, M., LeBlanc, M.M., Arnold, K.A., et al., 2007. Predicting workplace aggression: a meta-analysis. J. Appl. Psychol. 92, 228–238.

Holton, C., 2009. Identifying disgruntled employee systems fraud risk through text mining: a simple solution for a multi-billion dollar problem. Decis. Support Syst. 4, 853–864.

O'Neil, O.A., Vandenberg, R.J., DeJoy, D.M., Wilson, M.G., 2009. Exploring relationships among Anger, perceived organizational support, and workplace outcomes. J. Occup. Health Psychol. 3, 318–333.

Sakurai, K., Jex, S.M., 2012. Coworker incivility and incivility targets work effort and counterproductive work behaviors: the moderating role of supervisor social support. J. Occup. Health Psychol. 17, 150–161.

Schat, A.C.H., Kelloway, E.K., 2005. ‘Workplace aggression’. In: Barling, J., Kelloway, K., Frone, M. (Eds.), Handbook of Work Stress. Sage Publications, pp. 189–218.

Shaw, Eric, Sellers, Laura, June 2015. Application of the Critical-path Method to Evaluate Insider Risks, Studies in Intelligence, vol. 59. Central Intelligence Agency, Washington, DC. No. 2.

Shaw, E., Stroz, E., 2004. WarmTouch software: assessing friend, foe and relationship. In: Parker, T. (Ed.), Cyber Adversary Characterization: Auditing the Hacker Mind. Syngress Publications, Rockland, MA.

Shaw, E., Payri, M., Cohn, M., Shaw, I.R., 2013a. How often is employee anger an insider risk I? Detecting and measuring negative sentiment versus insider risk in digital communications (Part 1 of 2). J. Digital Forensics Secur. Law 8 (1), 39–71.

Shaw, E., Payri, M., Cohn, M., Shaw, I.R., 2013b. 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 (Part 2 of 2). J. Digital Forensics Secur. Law 8 (2), 73–92.

Taylor, P.J., Dando, C.J., Ormerod, T.C., Ball, L.J., Jenkins, M.C., Sandham, A., Menacere, T., 2013. Detecting insider threats through language change. Law Human Behav 37. http://dx.doi.org/10.1037/lhb0000032.

Weintraub, W., 1986. Personality profiles of American presidents as revealed in their public statements: the presidential news conference on Jimmy Carter and Ronald Reagan. Polit. Psychol. 7 (2), 285–295.

Weintraub, W., 1989. Verbal Behavior in Everyday Life. Springer Publishing Company, Inc.