JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice
eISSN : 2287-4577 pISSN : 2287-9099

Research Paper
J Inf Sci Theory Pract 7(1): 52-71, 2019
https://doi.org/10.1633/JISTaP.2019.7.1.5

Unified Psycholinguistic Framework: An Unobtrusive Psychological Analysis Approach Towards Insider Threat Prevention and Detection

Sang-Sang Tan*
Wee Kim Wee School of Communication and Information,
Nanyang Technological University, Singapore
E-mail: tans0348@ntu.edu.sg

Santhiya Duraisamy
School of Electrical and Electronic Engineering, Nanyang
Technological University, Singapore
E-mail: santhiya003@ntu.edu.sg

Jin-Cheon Na
Wee Kim Wee School of Communication and Information,
Nanyang Technological University, Singapore
E-mail: tjcna@ntu.edu.sg

ABSTRACT

An insider threat is a threat that comes from people within the organization being attacked. It can be described as a function of the motivation, opportunity, and capability of the insider. Compared to managing the dimensions of opportunity and capability, assessing one's motivation in committing malicious acts poses more challenges to organizations because it usually involves a more obtrusive process of psychological examination. The existing body of research in psycholinguistics suggests that automated text analysis of electronic communications can be an alternative for predicting and detecting insider threat through unobtrusive behavior monitoring. However, a major challenge in employing this approach is that it is difficult to minimize the risk of missing any potential threat while maintaining an acceptable false alarm rate. To deal with the trade-off between the risk of missed catches and the false alarm rate, we propose a unified psycholinguistic framework that consolidates multiple text analyzers to carry out sentiment analysis, emotion analysis, and topic modeling on electronic communications for unobtrusive psychological assessment. The user scenarios presented in this paper demonstrated how the trade-off issue can be attenuated with different text analyzers working collaboratively to provide more comprehensive summaries of users' psychological states.

Keywords: insider threat, psycholinguistics, text analysis, sentiment analysis, emotion analysis, topic modeling

Open Access

Accepted date: March 15, 2019
Received date: February 04, 2019

***Corresponding Author:** Sang-Sang Tan
PhD Candidate
Wee Kim Wee School of Communication and Information, Nanyang Technological University, 31 Nanyang Link, 637718, Singapore
E-mail: tans0348@ntu.edu.sg

1. INTRODUCTION

Compared to the advancements in outsider threat prevention and detection, the development of solutions for monitoring insider threat is still in its early stage. The discrepancy in the state of the art of outsider and insider threat mitigation does not necessarily mean that outsider threat poses greater risks to organizations. In fact, with their legitimate and privileged access to an organization's assets and their knowledge of the internal workings of the organization, it is easier for malicious insiders to target the vulnerabilities of the organization without having to overcome most of the barriers that protect the organization against outsiders. Therefore, adversarial insiders have the potential to cause more damage than outside attackers. To make matters worse, insiders are also in a better position to cover their tracks and to perpetrate crimes without being detected.

There is a growing body of literature that recognizes the importance of protecting organizations against insider threat. Gheyas and Abdallah's (2016) systematic literature review and meta-analysis revealed a clearly discernible upward trend in the number of publications related to insider threat mitigation from the year 2000. In general, many studies in this line of research including those of Chen and Malin (2011), Eberle, Graves, and Holder (2010), and Myers, Grimaila, and Mills (2009) have confined the scope of insider attacks to malicious activities that occur in the computational environment, such as data sabotage and espionage happening over organizational computing systems and networks. In the present study, however, we have taken a more general approach that views insider attacks as all types of malicious acts taken by anyone who has access to organizational resources, facilities, and information that would put an organization at risk or cause the organization to suffer any forms of loss. Examples of these malicious conducts include, but are not limited to, scenarios in which a trusted partner with legitimate access to organizational data secretly provides the data to the organization's competitor, or a former employee passes on sensitive information of an organization to his new employer. Our general view of insider attacks has an important implication: Given that we make no assumption of the context or environment in which these malicious acts might be carried out, commonly used methods that focus on tracking activities in organizational systems and networks may not be applicable. Specifically, the fact that some insider crimes might not involve unauthorized access or other anomalous conduct that can draw suspicion to an insider makes them difficult to detect through activity tracking. With respect to this difficulty, we posit that the mitigation of insider threat should also give emphasis to analyzing internal states like the personalities and emotions of individuals in addition to tracking external acts.

When it comes to modeling insider threat, there are numerous theoretical models to draw upon, the most commonly referred to being the generic set of Capability-Motivation-Opportunity (CMO) models (Schultz, 2002) that describes insider threat as a function of three dimensions: motivation, opportunity, and capability. As pointed out by Colwill (2009), while various technical and procedural solutions are available to address issues related to opportunity and capability, assessing motivation is usually more challenging. One's motivation to commit a malicious act is often affected by internal factors such as personalities and emotional states, which need to be assessed through psychological analysis. However, a direct psychological examination is not always an option considering the legal, ethical, and privacy concerns that might arise from this practice (Brown, Greitzer, & Watkins, 2013; Greitzer, Frincke, & Zabriskie, 2010; Kiser, Porter, & Vequist, 2010). Moreover, this kind of assessment is also obtrusive in nature and might be perceived as unfounded accusation and scrutiny, thus running the risk of causing human conflicts in organizations.

To be able to perform psychological analysis while minimizing the aforementioned risks, more recent attention has focused on automated text analysis of electronic communications as an alternative approach for carrying out behavior monitoring (Brown, Greitzer, & Watkins, 2013; Brown, Watkins, & Greitzer, 2013). These works originated from the field of psycholinguistics, an interdisciplinary field that studies the interrelation between psychological and linguistic aspects. Earlier research in psycholinguistics has shown that language use is correlated with psychological and emotional states (Pennebaker, Booth, & Francis, 2001; Pennebaker, Mehl, & Niederhoffer, 2003). As Brown, Greitzer, and Watkins (2013) noted, the primary advantage of using this kind of psycholinguistic approach is that "…organizations may unobtrusively monitor any and all individuals who routinely generate text with the organizations' information systems. As human analysts are excluded from the early phases of such analysis, the psycholinguistic approach may provide a means of monitoring psychosocial factors in a uniform and non-discriminatory manner" (p. 2).

Although existing studies on psycholinguistics have provided a good start, they have yet to reach the advancement needed for making a significant impact on predicting and detecting insider attacks. A major impediment to the progress of this line of research is the dilemma in balancing the risk of missed catches and the number of false alarms. On one hand, in order to minimize the risk of missing any potential threat, an optimal text analyzer should report all suspicious signs of insider threat.

On the other hand, the false alarm rate should not be too overwhelming for the output of the analysis to be actionable. This difficulty is complicated by the imperfection of text analysis methods which, in their current state of the art, cannot provide highly accurate results necessary for achieving these goals.

In the interest of finding the sweet spot in dealing with the trade-off between two seemingly contradictory states, we propose a unified psycholinguistic framework that combines multiple text analysis methods including sentiment analysis, emotion analysis, and topic modeling for unobtrusive psychological assessment. Both sentiment analysis and emotion analysis are fast-growing research areas in affective computing, a field focusing on the development of technology that enables machines to recognize and process human affect. The key difference between these two types of analyses is that the former refers to the recognition of sentiment valences (positive, neutral, or negative) whereas the latter embraces a more fine-grained analysis of human emotions (such as anger, joy, sadness, etc.). Finally, topic modeling complements our framework by facilitating the identification of significant topical patterns from the textual data.

The objective of this work is thus to develop and demonstrate the viability of a framework that combines methodologically diverse text analyzers to analyze insiders' written communications and monitor their psychological and emotional states. The proposed framework has a twofold bearing upon minimizing the risk of missed catches while maintaining a low false alarm rate. First, by taking into consideration the outputs generated by multiple text analyzers, the uncertainty in tracking potentially malicious insiders can be greatly reduced. For instance, when the outputs of all analyzers suggest that an employee has shown absolutely no sign of threat, security analysts can more confidently exclude the employee from the list of suspicious individuals. Likewise, if all analyzers indicate that an employee has a high potential of becoming a threat, it would seem reasonable to keep an extra eye on this employee or to take preventive actions that minimize the possibility of any future wrongdoings. Therefore, having multiple analyzers provides more assuring evidence to either dismiss or support further investigation. Second, each analyzer in the framework might be superior in some cases but less so in others. These text analyzers can thus complement each other in the sense that one analyzer might be able to capture the signs that other analyzers have missed. In particular, due to the fundamental technical limits embodied in different text analysis methods, different analyzers might generate contradictory results in the assessment of the same individual. With a unified framework in which different text analyzers work collaboratively across methodological divides, these contradictory results can serve as the indicator for invoking further investigation, thus reducing the risk of overlooking any signs of potential threat.

2. RELATED WORK

2.1. Conceptual Modeling of Insider Threat

Schultz (2002) pointed out that many conceptual models of insider threat can be subsumed under the broader umbrella of CMO models. Variants of CMO models include those described by Parker (1998) and Wood (2000). In general, the CMO models suggest that insider attacks happen with the presence of the following essential components (Schultz, 2002):

- Capability, which refers to the level of relevant knowledge and skill that would enable an insider to commit the crime
- Motivation, which encompasses various internal and external factors that might eventually trigger or lead to the disloyal act of an insider
- Opportunity, which depends on how easy it is for an insider to commit an attack. For example, insiders with more access rights or a system with more vulnerabilities would increase the opportunity for attack.

A recent systematic literature review by Gheyas and Abdallah (2016) categorized studies related to insider threat mitigation based on these three dimensions (or combination of dimensions). They concluded that the vast majority of studies falls into the category of opportunity. Specifically, about two-thirds of the studies examined in the systematic review used opportunity scores as the key features for insider threat detection and prediction. Drawing on an extensive range of sources, Gheyas and Abdallah (2016) summarized that most publications that concentrated on the opportunity dimension employed users' access rights and activities as indicators of opportunity. Users' access rights are defined by their system roles whereas the information on users' activities can be acquired from various types of log files such as database logs, web server logs, and error logs that provide a simple and cost-effective implementation for real-time activity tracking.

Compared to assessing the level of opportunity, assessing the level of motivation can be more challenging (Colwill, 2009), not only because of practical concerns in performing direct psychological examinations (Brown, Greitzer, & Watkins, 2013; Greitzer et al., 2010; Kiser et al., 2010), but also in terms of the difficulty in quantifying, recording, and tracking motivation systematically. The present study aims to add to this research line by demonstrating the viability of monitoring insiders' motivation using automated text analyses. In the following subsection, we review some representative works that have adopted a similar methodological approach as presented in this study. A more comprehensive survey of the large body of research in insider threat and the structural organization of existing works based on various criteria can be found in Azaria, Richardson, Kraus, and Subrahmanian (2014) and Gheyas and Abdallah (2016).

2.2. Assessing Insiders' Motivation

Numerous frameworks have been proposed and applied to tackle the problem of insider threat from the insiders' motivation dimension. Research suggests that the motivation of an insider can be assessed from various aspects including the predisposition to malicious behavior, mental disorder, personality, and emotional states (Gheyas & Abdallah, 2016). Among these aspects, psychological and emotional vulnerabilities are often considered the pivotal factors contributing to insiders' motivation. For example, anger and disgruntlement have been frequently mentioned as indicating the motivation to perpetrate malicious conduct (Greitzer, Kangas, Noonan, Brown, & Ferryman, 2013; Ho et al., 2016; Shaw & Fischer, 2005). These negative psychological and emotional states can be shaped by interrelated factors coming from the inside and the outside, including one's personality, stress level, ability to cope with criticism, issues in personal life, and corporate factors, among others (Azaria et al., 2014).

Conceptually, many studies have suggested that assessing the emotional and psychological states of insiders can be useful for detecting and preventing malicious conduct. However, in terms of methodological development, there is still ample room for improvement in this research area. Axelrad, Sticha, Brdiczka, and Shen (2013) proposed the use of a Bayesian network for scoring insiders. The proposed model incorporated five categories of variables that measure occupational stress level and personal life stress level, personality variables, attitude and affect, history of social conflict, and so forth as indicators of an insider's degree of interest, which represents the relative risk of committing an attack. Based on empirical analysis of the collected data, the initial Bayesian network model was then adjusted to produce a predictive model of insider threat. Although their framework is conceptually appealing, the collection of data from a questionnaire poses some problems. The use of this data collection method is not uncommon in existing frameworks of insider threat. For instance, in Brdiczk et al. (2012), this method was used for psychological profiling in an insider prediction model. The development and validation of their model were carried out in the setting of a popular multi-player online game called World of Warcraft. Their framework combined structural anomaly detection of abnormal patterns in social and information networks with psychological profiling of the game characters to predict which characters would turn against their social groups in the game. For psychological profiling, the researchers made use of various sources including World of Warcraft census data, gamers' personality profiles from an online questionnaire, behavioral features based on gamers' activities in the game, simple analysis of game characters' names and their guild names, and the in-game social network of each gamer. In the model proposed by Kandias, Mylonas, Virvilis, Theoharidou, and Gritzalis (2010), the researchers also suggested the use of questionnaires to determine the stress level, predisposition, and user sophistication for psychological profiling of insiders. The shortcoming of this data collection method, however, is that it only allows periodical assessment, of which the frequency depends on practical factors like costs and cooperation from employees. Such a method might also be susceptible to self-reporting bias or other human-related biases. Contrastingly, automated text analysis methods facilitate continuous monitoring and reduce human biases.

In light of evidence that suggests a correlation between psychological and emotional factors and individuals' verbal (written or spoken) behavior (Pennebaker et al., 2001, 2003), there have been several attempts to predict or detect malicious insiders from numerous forms of textual data. One of the most prominent studies was undertaken by Greitzer et al. (2013). Based on the premises that psychosocial behavior is an indicator of insider threat and that these psychosocial factors are closely associated with word usage in spoken and written language, Greitzer et al. (2013) inferred that textual contents can be a valuable source for detecting insider threat through the identification of linguistic patterns pertaining to personality traits. Drawn upon a widely accepted standard for assessing personality traits, which is often referred to as the Big Five (McCrae, 2010), Greitzer et al. (2013) highlighted three personality traits—conscientiousness, neuroticism, and agreeableness—as the major factors that offer promise for prediction and detection of insider threat. The researchers applied text analysis to an email corpus that was injected with email samples of six known criminals. Their results showed that most of the known criminals were identified as outliers with high scores on neuroticism and low scores on conscientiousness and agreeableness.

In the same vein, Taylor et al. (2013) examined the changes in language usage in electronic communications when some team members decided to turn against the team. Their data were collected from a simulated environment. The participants communicated through emails in a simulated workplace, whereby after the first stage of the study some participants were offered some incentives to start acting as malicious insiders in the team. From the text analysis of the collected emails, the researchers found that insiders showed several signs in their language usage: Compared to other co-workers, malicious insiders used more self-focused words, more negative language, and more words related to cognitive processes. Furthermore, the study also reported a deterioration in language similarity between insiders and other team members as the insiders became gradually estranged from the rest of the team over time.

Another study (Ho et al., 2016) applied linguistic analysis to conversational data collected from a multi-player gaming environment on the Google+ Hangout platform. The gaming environment simulated a betrayal scenario, in which a group member accepted an offer of incentives to betray the group. By comparing the within-group communications for those groups that did and did not have a deceptive insider (the control group), and before and after a member was compromised in a group, the study aimed to identify relevant linguistic cues such as negations, emotion-related words, words pertaining to cognitive processes, and so forth for revealing deceptive acts among the game players. The study reported that some subtle but identifiable patterns in group communications might represent an elevated risk of insider threat.

Collectively, all the studies described so far have provided evidence that text analysis can be a promising research direction towards understanding insider threat from a psychological perspective. However, these studies share a commonality; that is, all of them utilized Linguistic Inquiry and Word Count (LIWC), which is a text analysis program developed based on the works of Pennebaker and his colleagues (Pennebaker et al., 2001). Commonly used for psycholinguistic analysis, the LIWC program analyzes text by computing scores on word categories that provide insights into human social, cognitive, and affective dimensions. Although promising, existing studies remain narrow in focus, dealing mainly with text analysis using LIWC. An exception to this commonality is the framework proposed by Kandias, Stavrou, Bozovic, Mitrou, and Gritzalis (2013), that aimed to examine insiders' motivation by analyzing the content they generated and made public online. In their study, the researchers compared the performance of several machine learning techniques in classifying YouTube comments. The most accurate classifier was chosen as the model for detecting negative attitude towards authorities and enforcement of the law. Additionally, they also used a manually created, task-specific dictionary to perform classification by keyword matching. The framework proposed in the present study is similar to Kandias et al.'s framework in terms of the use of both statistical and dictionary-based methods for text analysis. However, while their framework only focused on identifying negative attitude, our framework covers a wider range of methods for a more comprehensive view that reveals the multiple facets of textual data using sentiment analysis, emotion analysis, and topic modeling.

3. METHOD

3.1. Proposed Framework

Fig. 1 shows the unified psycholinguistic framework proposed in the present study. This framework consolidates multiple

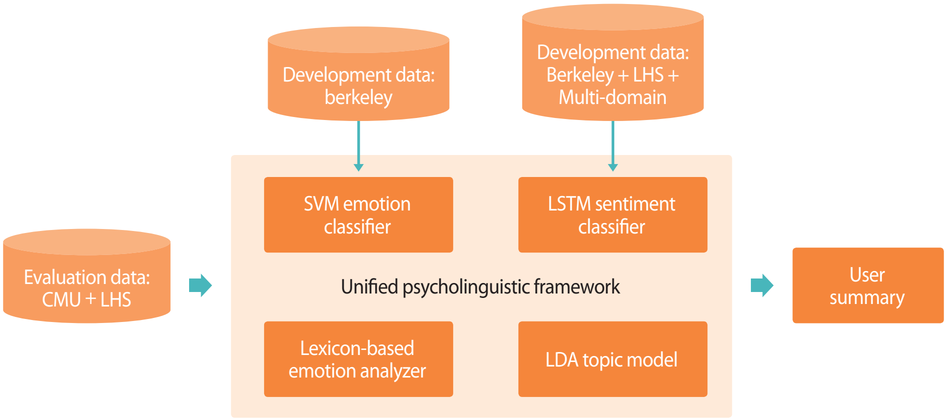


Fig. 1. The unified psycholinguistic framework. SVM, Support Vector Machine; LSTM, Long Short-Term Memory; LDA, Latent Dirichlet Allocation.

text analysis methods to generate comprehensive summaries of individuals' psychological states from their written texts. In relevance to our ultimate goal of supporting psychological assessments for prediction and detection of insider threat, we have adopted the following text analysis methods: sentiment analysis, emotion analysis, and topic modeling.

Sentiment analysis is a subfield of natural language processing that analyzes written or spoken language computationally to determine sentiment valences from textual contents. With regard to insider threat monitoring, sentiment analysis can provide an overview of whether an individual is a positive or negative person in general. In recent years, deep learning using neural networks has emerged as a powerful machine learning method for a diverse array of problems, including image processing, speech recognition, and various problems related to natural language processing. Like many machine learning methods, deep neural networks follow a data-driven approach to learn—from the training data—a function that best describes the mapping from the data to the output variable. One of the strengths of neural networks is their ability to represent a wide variety of mapping functions with very few constraints. As established by Hornik (1991) through theorem proving, with sufficient artificial neurons in the hidden layers, a multilayer neural network can be a universal approximator. The present study used Long Short Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997), a recurrent neural network architecture that has been widely adopted for deep learning in many sentiment analysis tasks.

While sentiment analysis provides an overview of individuals' attitudes based on sentiment valences of texts, emotion analysis aims to give a more detailed view of individuals' affectual states such as angry, joyful, fear, sad, and so forth. The detection and classification of emotions have a wide range of applications, such as determining personality traits (Cherry, Mohammad, & De Bruijn, 2012) and detecting depression (Grijalva et al., 2015). With respect to the objective of the present study, we suggest that a closer look at individuals' emotions in addition to their sentiment valences can be useful for pinpointing potential threat. Specifically, emotion analysis can help to narrow down the list of possible suspects by targeting certain emotions. For instance, individuals associated with the anger emotion are probably more aggressive than individuals showing other negative emotions like sadness and fear. In the interest of exploring different approaches, we implemented two emotion analysis techniques in the present study: first, emotion classification with machine learning using Support Vector Machine (SVM) (Cortes & Vapnik, 1995); second, lexicon-based analysis using NRC emotion lexicon (Mohammad & Turney, 2013).

In addition to the analysis of sentiments and emotions, our proposed framework also includes topic modeling as one of its core components. Our topic model was built using Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003), one of the most common topic modeling methods currently in use. The purpose of incorporating topic modeling into the framework is to give security analysts an overall picture of key topics penetrating the electronic communications under surveillance. Topic modeling can come in handy in this regard as it provides a convenient way to discover topical patterns statistically from the enormous volume of textual contents.

The implementation details of the text analyzers, as well as the data used for the development and evaluation of the proposed framework, are described in the following subsections.

3.2. Data

The following datasets were used for development and evaluation of the unified psycholinguistic framework:

- CMU Enron email dataset,[1] which is a collection of corporate emails of 150 users
- LHS dataset,[2] which consists of three types of text: love letters (L), hate emails (H), and suicide notes (S)
- UC Berkeley Enron email dataset,[3] which is a subset of an Enron email collection that contains 1,702 emotionally labeled emails
- Multi-domain review dataset,[4] which is a data corpus of positive and negative online reviews, ranging over 25 different topics including health, software, automotive, magazine, baby, beauty, and electronics, among others

3.2.1. Data for the Evaluation of the Framework

The evaluation of the framework was carried out on the first two datasets, i.e. the CMU Enron email dataset and the LHS dataset. Although the most ideal way to assess the effectiveness of the proposed framework is to evaluate it against ground truths of emotions, sentiments, and topics, such an approach would require large-scale manual labeling, which is time-consuming and costly. We thus resolved to verify the results generated by our framework using reference cases obtained from the LHS dataset. To this end, two synthetic users were created from each type of texts (i.e., L, H, and S) from the LHS dataset and were injected into the CMU Enron email dataset.

[1] https://www.cs.cmu.edu/~./enron

[2] http://saifmohammad.com/WebPages

[3] http://bailando.sims.berkeley.edu/enron_email.html

[4] https://www.cs.jhu.edu/~mdredze/datasets/sentiment

For instance, the 331 love letters from the LHS dataset were split into two sets consisting of 166 and 165 documents which constitute the 'emails' sent by synthetic users '_love1' and '_love2' respectively. Likewise, synthetic users '_hate1', '_hate2', '_suicide1', and '_suicide2' were created from hate emails and suicide notes.

Prior to text analysis, all emails in the CMU dataset were cleaned to remove email headers and forwarded texts. This step is essential to ensure that we only analyzed those emails that were sent—as opposed to received or forwarded—by the Enron users to understand the users' behavior from their written texts. This cleaning step was performed automatically by a computer script, followed by a manual examination of randomly selected emails to make sure that most emails were reasonably clean. This preprocessing step was carried out only on the CMU Enron dataset; the LHS dataset required no cleaning and was used in its original form. Altogether, the CMU Enron dataset and the LHS dataset resulted in a collection of texts contributed by 156 individuals.

3.2.2. Data for the Development of the Text Analyzers

Our unified psycholinguistic framework made use of both supervised and unsupervised methods for text analysis. The lexicon-based emotion analyzer and the topic model were implemented using unsupervised methods whereas the SVM emotion classifiers and the LSTM sentiment classifier were built via supervised learning, which entailed the use of manually labeled data in the development phase.

For the development of the SVM classifiers, the UC Berkeley Enron email dataset was used for training and testing the classification models. Although this dataset contains 19 emotion labels, only a subset of the labels is relevant to the goal of the present study. From Plutchik's model of emotions (Plutchik, 1982), we chose two emotions—anger and joy—for which classification models were built. As noted in many studies (Greitzer et al., 2013; Ho et al., 2016; Shaw & Fischer, 2005), the manifestation of the anger emotion in verbal communications can be a sign of psychological stress and dissatisfactions; it is thus not surprising that this emotion is often linked to the elevated risk of insider threat. The joy emotion was chosen as a contrasting emotion to anger because the low scores of joy emotion can somehow serve as supplementary evidence of negativity in individuals. Table 1 shows the mapping of emotions from the Berkeley dataset to the two emotions we are interested in. In addition to coalescing labeled emails into these main categories, we also replaced the emotion labels at the email-level with labels at the paragraph-level because quite often, lengthy emails that have been labeled with certain emotions only contain a few paragraphs pertaining to those emotions. It is thus reasonable to carry out emotion classification at the paragraph-level instead of email-level. To this end, the emails were automatically split into paragraphs at the occurrences of ending punctuation marks and empty lines, and the paragraphs that express the labeled emotions were identified manually. After email headers were removed from the data, the resulting paragraph-level dataset contains 37,684 instances, with 81 instances and 101 instances identified as pertaining to anger and joy, respectively.

Generally speaking, deep learning techniques like LSTM usually take a fairly large amount of data to achieve satisfactory performance. Therefore, for the development of the LSTM sentiment classifier, we combined the following: the UC Berkeley Enron email dataset, the LHS dataset, and the multi-domain review dataset. Synthetic users from the LHS dataset were first injected into the Berkeley dataset to obtain a bigger corpus. We assumed that all documents generated by '_love1' and '_love2' are positive whereas the documents generated by other synthetic users are negative. However, after combining the Berkeley Enron email dataset and the LHS dataset, the resulting dataset was rather imbalanced and inadequate for an optimum classification task. We thus further increased the size of the corpus with the multi-domain review dataset. Eventually, a balanced training dataset of 4,530 documents was obtained with 1,510 documents for each sentiment class. Unlike the classification of emotions, the classification of sentiments was implemented at the email-level instead of paragraph-level. The 19 emotions in the Berkeley dataset were mapped into positive, negative, and neutral sentiments as shown in Table 2.

Table 1. Mapping of emotions from the Berkeley dataset to three emotion categories

Emotion labels in the Berkeley dataset	Classified as
Anger / agitation	Anger
Jubilation and triumph / gloating	Joy
Humour, camaraderie, admiration, gratitude, friendship / affection, sarcasm, secrecy / confidentiality, concern, competitiveness / aggressiveness, pride, shame, hope / anticipation, dislike / scorn, worry / anxiety, sadness / despair, and sympathy / support	None (i.e., classified as no emotion)

3.3. Text Analyzers

In the interest of providing a more comprehensive view for monitoring users' emotional and psychological states, we have adopted a multi-faceted approach by including diverse types of text analyzers in our framework. The SVM and lexicon-based emotion analyzers help to identify individuals showing an exceptionally high level of anger emotion or an unusually low level of joy emotion; the LSTM sentiment classifier provides a view of individuals' positivity and negativity in general. The topic model aims to shed some light on the key topics around which the communications revolve.

3.3.1. SVM Emotion Classifiers

Using the paragraph-level Enron emails, we built SVM classifiers for binary classification of anger and joy such that each classifier was responsible for classifying every email paragraph as binary 1 or 0 based on the presence or absence of the emotions. Specifically, the anger classifier would classify a paragraph as presence (binary 1) if anger were detected in the paragraph, and absence (binary 0) if the paragraph showed no sign of anger. Likewise for the joy classifier. Our classification models used linear kernels with the cost of misclassification C = 0.1.

Building an SVM classifier via supervised learning entails finding the optimal decision boundary to separate instances of one class from another. In the SVM algorithm, this optimal decision boundary is the hyperplane that has the largest distance to the closest points of all classes. The performance of an SVM classifier relies heavily on the features that constitute the multi-dimensional feature space where the search for the best fitting hyperplane takes place. The present study made use of the WEKA package contributed by Mohammad and Bravo-Marquez (2017) to generate the following features for emotion classification:

- Word and character n-grams
- Negations: adding prefixes to words occurring in negated contexts. For instance, 'I do not like you' becomes 'I do not NEG-like NEG-you'.
- Part-of-speech tags: creating a vector space model from the sequence of part-of-speech tags
- Brown clusters: mapping words to Brown word clusters to create a low-dimensional vector space model
- Lexicon features: generating lexicon-related features using various lexicons including MPQA, Bing Liu's lexicon, AFINN, NRC Word-Emotion Association Lexicon, and so forth
- Positive and negative sentiment strengths: generating strengths of sentiments using SentiStrength

With the Enron paragraph-level emails, a major hindrance we faced in emotion classification was the huge discrepancy between the numbers of class 1 and class 0 instances. In other words, the dataset was highly imbalanced. This actually projects a realistic picture of the real-world data: In general, most emails are non-emotion-related and can be regarded as 'normal' emails, and the mission of the classifiers is to detect the tiny portion of 'abnormal' emails. Under such circumstances, most classifiers tend to bias towards the majority class, resulting in an extremely low (close to zero) accuracy in identifying instances of the minority class.

To tackle this problem, we applied under-sampling and over-sampling to reduce the gap between the majority class and the minority class. Under-sampling was first applied using random resampling to shrink the majority class to 30 times the size of the minority class. This was followed by the over-sampling procedure that uses Synthetic Minority Over-sampling Technique (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) to generate synthetic

Table 2. Mapping of emotions from the Berkeley dataset to the three sentiment classes

Sentiment labels	Emotion labels in the Berkeley dataset
Positive	Jubilation, hope / anticipation, humor, camaraderie, admiration, gratitude, friendship / affection, and sympathy / support
Negative	Worry / anxiety, concern, competitiveness / aggressiveness, triumph / gloating, pride, anger / agitation, sadness / despair, shame, and dislike / scorn
Neutral	Sarcasm and secrecy / confidentiality

Table 3. Numbers of instances in both classes before and after over-sampling and under-sampling

Emotion	Before over-sampling and under-sampling		After over-sampling and under-sampling	
	Presence	Absence	Presence	Absence
Anger	81	37,603	567	2,430
Joy	101	37,583	707	3,030

samples for the minority class, resulting in an expanded size of seven times the original size of the minority class. Note that the parameters that specified the scales by which these two classes were under-sampled or over-sampled were chosen by experiments. The original sizes of both classes and their sizes after over-sampling and under-sampling are given in Table 3.

3.3.2. LSTM Sentiment Classifier

We built the LSTM sentiment classifier using Keras, a high-level Python library that runs on top of Theano and TensorFlow to simplify the development of deep learning models. The network topology and the model parameters are described below.

- *Input layer*. Like other neural networks, an LSTM network requires numerical inputs. Therefore, text data need to be converted into numbers using word embedding—a text representation technique that maps discrete words into real-valued vectors. We chose to represent each word as a 128-dimensional vector, and the maximum length of each document in the dataset was capped at 50 words. In specific, the 4,530 documents in our dataset were converted to a set of 128 × 50 matrices.
- *Hidden layer*. The network's hidden layer contains 128 memory units. This layer takes as input the matrices generated by the word embedding representation procedure.
- *Output layer*. To tackle the three-class sentiment classification problem, the output layer of the network was designed as a dense (i.e., fully-connected) layer with three neurons and a softmax activation function to predict sentiment valences.

The network was trained for 10 epochs with a batch size of 32. Additionally, to reduce overfitting, we applied the dropout method to skip activation and weight updates for the inputs and recurrent connections at a probability of 0.2.

3.3.3. Lexicon-Based Emotion Analyzer

Lexical resources have been the key instrument in the analysis of sentiments and emotions. They provide scores, either discrete or continuous, for words and phrases that are salient indicators of sentiments and emotions. These resources can be utilized in many ways: Some studies used lexical resources as part of the rule-based approach while others incorporated lexicon-related features into the machine learning approach. But according to Mohammad (2015), the vast majority of works in emotion analysis have employed the statistical machine learning approach. One of the major obstacles in using the machine learning approach is the paucity of labeled data. As far as we know, large amounts of labeled data are only available from tweets, for which emoticons, emoji, and hashtag words such as #anger and #sadness can be used as emotion labels to produce pseudo-labeled data (Mohammad, 2012). Due to the limited amount of labeled data for emotion analysis of emails, in addition to the more widely used SVM emotion classification, we also implemented another emotion analyzer which relies merely on Mohammad and Turney's (2013) NRC emotion lexicon (version 0.92) to acquire emotion scores for the analyzed texts.

The NRC emotion lexicon provides binary values that indicate the presence or absence of Plutchik's eight emotions (Plutchik, 1982). Table 4 shows the binary values assigned to two examples of entries, 'abandonment' and 'helpful.' Using this lexicon, we followed the steps below to analyze anger and joy in the evaluation data:

- Lemmatization was first performed to preprocess the emails.
- Using the lemmatized texts, we looked up the NRC emotion lexicon to obtain a sum of scores for each emotion and for every email.
- To facilitate comparisons between individuals, we computed the final scores of anger and joy for every individual. Each final score was obtained by averaging the individual's overall score from all emails over the total number of emails written by the individual.

3.3.4. LDA Topic Model

One of the reasons for the emergence of topic modeling as a prevalent instrument for text analysis is the availability of many easy-to-use packages. In the present study, we used McCallum's topic modeling toolkit, MALLET, which provides a fast implementation of LDA (Blei et al., 2003), a method widely used for topic modeling and information summarization.

In the procedure of discovering latent topics from textual

Table 4. Examples of entries from the NRC emotion lexicon

Emotion	Abandonment	Helpful
Anger	1	0
Anticipation	0	0
Disgust	0	0
Fear	1	0
Joy	0	1
Sadness	1	0
Surprise	1	0
Trust	0	1

contents, LDA loops through all words in the text collection and assigns these words to the most probable topics. The procedure starts with a random assignment of topics and then rectifies the assignment over a large number of iterations until an optimal state is reached. In general, topic modeling considers a topic as a cluster of words that occur in some statistically meaningful ways. Given a text collection, the primary output of topic modeling is a list of keyword clusters pertaining to K topics, where K is a predetermined number that specifies how many topics are to be returned.

The present study applied LDA to extract 50 keyword clusters from all emails in the evaluation data. Labels were manually assigned to the 50 latent topics based on topic keywords and the most representative emails of each topic, i.e., emails with substantial contents related to the topics. Following the identification of these 50 most commonly discussed topics in our data, we then obtained the distribution of these key topics for each individual to perform cross-comparisons.

4. RESULTS AND DISCUSSION

Evaluations were carried out in two stages. At the first stage, text analyzers were evaluated individually to ensure that they are performing at the state-of-the-art level. At the second stage, the unified psycholinguistic framework as a whole was qualitatively assessed by its effectiveness in identifying potentially adversarial insiders through the use of various text analysis methods.

4.1. Evaluation of Individual Text Analyzers

This evaluation stage only applies to the SVM emotion classifiers and the LSTM sentiment classifier. Since these text analyzers were built via supervised learning, the learned classification models can be validated on the labeled development data. The precision (P), recall (R), and F-score (F) of emotion classification and sentiment classification are given by the equations below:

$$\mathrm{P} = \frac{\mathrm{TruePositive}}{\mathrm{TruePositive} + \mathrm{FalsePositive}} \quad (1)$$

$$\mathrm{R} = \frac{\mathrm{TruePositive}}{\mathrm{TruePositive} + \mathrm{FalseNegative}} \quad (2)$$

$$\mathrm{F} = \frac{2\mathrm{PR}}{\mathrm{P} + \mathrm{R}} \quad (3)$$

Since the evaluation data consists of highly imbalanced classes, the weighted averages of precision, recall, and F-score are used as the overall performance measures (Equations 4-6). For each of the k classes, the precision, recall, and F-score of the class are weighted by the number of instances in the class, and N is the total number of instances.

$$\mathrm{WeightedAverage(P)} = \frac{\sum_{i=1}^{k} n_i P_i}{\mathrm{N}} \quad (4)$$

$$\mathrm{WeightedAverage(R)} = \frac{\sum_{i=1}^{k} n_i R_i}{\mathrm{N}} \quad (5)$$

$$\mathrm{WeightedAverage(F)} = \frac{\sum_{i=1}^{k} n_i F_i}{\mathrm{N}} \quad (6)$$

The SVM emotion classifiers were validated with 3-fold cross-validation. Although 10-fold cross-validation is more commonly used for model validation, 3-fold cross-validation seems to be a better option in this case considering the number of class 1 instances of each emotion. In other words, it is unlikely that every partition would contain a sufficient number of class 1 instances if 10-fold cross-validation were to be used.

The validation was carried out in two experimental settings. In both settings, the classifier was trained with two-thirds of the over-sampled and under-sampled data in each round of the 3-fold cross-validation. Note that although the over-sampled minority class was used for training, the synthetic instances generated by Synthetic Minority Over-sampling Technique were excluded from the test data. In other words, the only difference between the two settings is the size of the majority class in the test data: While the first setting only tested the under-sampled majority class, the second setting tested all instances of the majority class. In the classification of highly imbalanced data, F-scores of the majority class (i.e., class 0) are usually beyond satisfactory. Therefore, our evaluation of the classifiers focuses mainly on the results obtained for the minority class (i.e., class 1), although the results for both classes are presented in Table 5 to give a complete picture of the classifiers' performance. Unless otherwise specified, the following discussion on the classifiers' performance refers to the minority class.

Overall, the results obtained with the under-sampled majority class are considered comparable to those demonstrated in existing studies (Mohammad, 2012; Mohammad, Zhu, Kiritchenko, & Martin, 2015). Nevertheless, when the emotion classifiers were validated with the full-sized majority class, the F-scores of both classifiers decreased tremendously. From the confusion matrices presented in Table 5, it can be seen that the

Table 5. Emotion classification results with 3-fold cross-validation

Emotion	Class	Classified as		Precision	Recall	F-score
		0	1			
With the under-sampled majority class						
Anger	0	2,423	7	0.985	0.997	0.991
	1	36	45	0.865	0.556	0.677
	Weighted average	-	-	0.981	0.983	0.981
Joy	0	3,015	15	0.987	0.995	0.991
	1	41	60	0.800	0.594	0.682
	Weighted average	-	-	0.981	0.982	0.981
With the full-sized majority class						
Anger	0	37,433	170	0.999	0.995	0.997
	1	33	48	0.220	0.593	0.321
	Weighted average	-	-	0.997	0.995	0.996
Joy	0	37,368	215	0.999	0.994	0.997
	1	39	62	0.224	0.614	0.328
	Weighted average	-	-	0.997	0.993	0.995

Table 6. Sentiment classification results with 10-fold cross-validation

Class	Classified as			Precision	Recall	F-score
	Neutral	Positive	Negative			
Neutral	1,374	63	73	0.896	0.910	0.903
Positive	81	1,111	318	0.752	0.736	0.744
Negative	78	303	1,129	0.743	0.748	0.745

degradation in F-scores was mainly caused by the increase in the number of false positives (i.e., false classification of class 0 as class 1). This is, however, not surprising in applications that aim to detect anomalous activities, which are rare and abnormal activities that may have serious consequences when not revealed. Apparently, the detection of potential insider threats also falls into this category of applications. Since the consequence of missing any potentially endangering individuals is detrimental, it is often necessary to have a 'skeptical' classifier to minimize the possibility of missing any suspects, even when it comes at the cost of higher false positive rate.

Table 6 shows the confusion matrix, precisions, recalls, and F-scores generated by the LSTM sentiment classifier with 10-fold cross-validation. The F-scores achieved on the prediction of the positive class (74.4%) and the negative class (74.5%) were considerably lower than the F-score achieved on the neutral class (90.3%). Nevertheless, the average F-score of the classifier (79.7%) still falls within an acceptable performance range for document-level sentiment analysis.

4.2. Evaluation of the Communications Using the Proposed Framework

Due to the inclusion of reference cases from the LHS dataset, at the time of evaluation we already had the prior knowledge that at least six of the 156 individuals in the evaluation data were affectively charged: two of them ('_love1' and '_love2') in a very positive way; four of them ('_hate1', '_hate2', '_suicide1', and '_ suicide2') in a very negative way. For the proposed framework to be useful in identifying insider threat, it should be able to provide summaries of the users' psychological states so that informed decisions can be made to list the hate and suicide users as potentially malicious insiders and the love users as people who are not likely to involve in any adversarial acts. To demonstrate how this goal can be achieved by the proposed framework, we adopted an outlier detection method to select a small number of anomalous individuals from the evaluation data which comprises written texts contributed by 156 individuals. We then examined the cases of these anomalous individuals to assess the credence of the results produced by our framework.

The goal of outlier detection is to identify observations that deviate from common patterns and other observations in the collected data. Arriving at the conclusion that certain observations should be categorized as outliers is a highly subjective exercise. In the present study, we used interquartile range (IQR) and established the outlier range as 3 × IQR to identify outliers based on emotion scores obtained from the lexicon-based emotion analysis method described in Section 3.3.3. The outliers suggested by IQR are presented in Table 7. The scores shown next to the users' names indicate the average number of emotion words per email. For instance, '_hate2'—who tops other users at the anger emotion ranking—used an average of 5.73 anger words per email.

Judging from the rankings of the six reference cases, it seems the lexicon-based emotion analysis was able to generate reasonable scores for outlier detection. In particular, the love users are ranked on top of other users under the joy emotion whereas the hate and suicide users have higher rankings under the anger emotion. Another interesting finding is that the six reference cases always come before the real Enron users. This observation is likely to be related to the language and communication styles used in different forms of written communications: Compared to the more personal writing styles in love letters, hate mails, and suicide notes, business and professional writing in a workplace environment often uses a more formal tone and subtler expressions of personal emotion.

From the anomalous users shortlisted by the outlier detection method, we chose three individuals to zoom in into some user scenarios that would demonstrate the usage of the unified psycholinguistic framework and support the viability of predicting and detecting potential insider threats with psycholinguistic analysis. The following users were chosen for this purpose:

- '_love2' and '_hate2'. These synthetic users serve the purpose of quick verification for the proposed framework.
- '_enron1'. This user is the Enron employee that scored highest under the anger emotion.

The user scenarios of the three users are presented with the following graphs and charts that visualize the results generated by all text analyzers in the proposed framework:

- The lexicon-based emotion timeline shows the emotion scores obtained from a simple count of emotion words per email. The scores were normalized to the range of [min, max], where min and max are the minimum and maximum count of anger-related or joy-related words—depending on which emotion is analyzed—across the three users included in the user scenarios.
- The sentiment classification timeline and proportion chart visualize the emails' class labels (positive / *pos*, neutral / *neu*, and negative / *neg*) predicted by the LSTM sentiment classifier.
- The emotion classification timeline and proportion chart visualize the emails' class labels (1 for presence, 0 for absence) predicted by the SVM emotion classifiers. Since the classifiers were trained at the paragraph-level instead of document-level, the predictions were first carried out at paragraph-level but a final class label was obtained for each email using a logical OR function, which assigned class label 1 to an email if any of its paragraphs was predicted as 1 in the classification.
- The topic distribution chart shows the counts of users' emails pertaining to the 50 topics extracted by the LDA topic model.

In the scenarios of the two synthetic users, what stands out the most is that the results produced by all text analyzers seem to agree with each other. Although it would be overstating matters to claim that they are highly similar, we can still conclude that the emotion scores generated by the lexicon-based emotion analyzer and the predictions made by the LSTM sentiment classifier and the SVM emotion classifiers show convincing similarity to a certain extent. For example, in the scenario of the love user, the lexicon-based emotion timeline (Fig. 2) depicts that the number of joyful words in the user's texts clearly surpasses the number of angry words. In agreement with this outcome, the predictions provided by the sentiment classifier and the emotion classifiers show that a large portion of the _love2 user emails are positive (Fig. 3) and joyful (Fig. 4). Likewise, the _hate2 user's emotion scores obtained from lexicon-based emotion analysis (Fig. 5) and predictions

Table 7. Outliers detected using interquartile range on lexicon-based emotion scores

	Anger	Joy
1	_hate2 (5.73)	_love1 (12.54)
2	_suicide2 (4.91)	_love2 (11.14)
3	_hate1 (4.51)	_suicide2 (7.64)
4	_suicide1 (4.1)	_suicide1 (7.1)
5	_love2 (1.49)	_hate2 (6.44)
6	_love1 (1.38)	_hate1 (4.17)
7	_enron1 (1.27)	
8	_enron2 (1.09)	
9	_enron3 (1.04)	
10	_enron4 (1.03)	
11	_enron5 (1.01)	

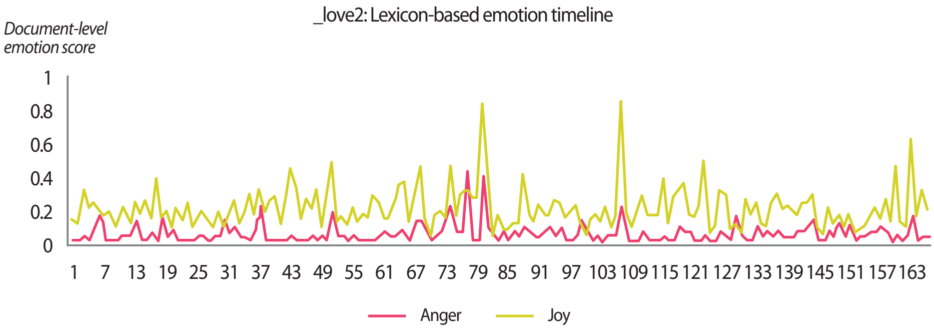


Fig. 2. Lexicon-based emotion timeline for synthetic user '_love2.'

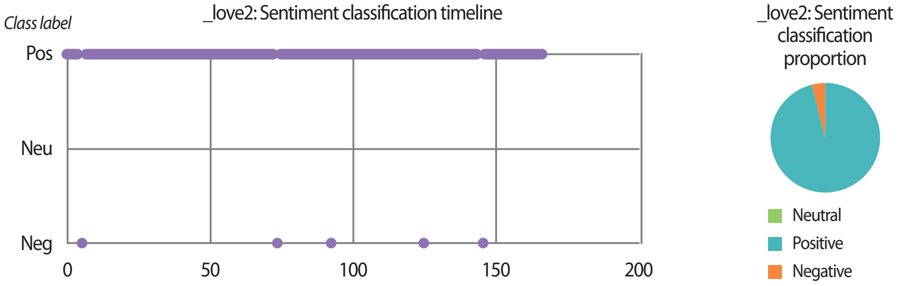


Fig. 3. Sentiment classification timeline and proportion for synthetic user '_love2.'

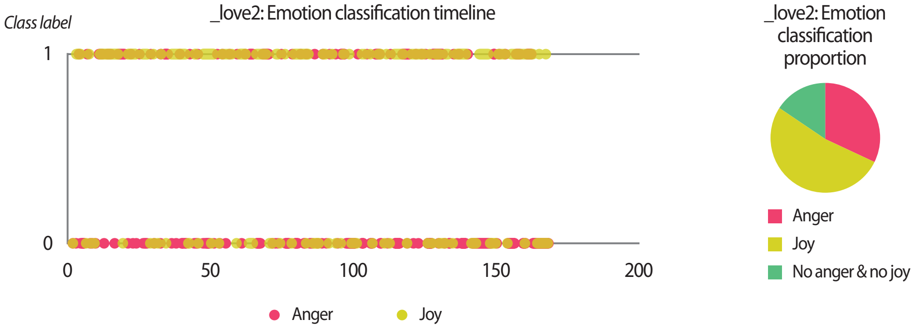


Fig. 4. Emotion classification timeline and proportion for synthetic user '_love2.'

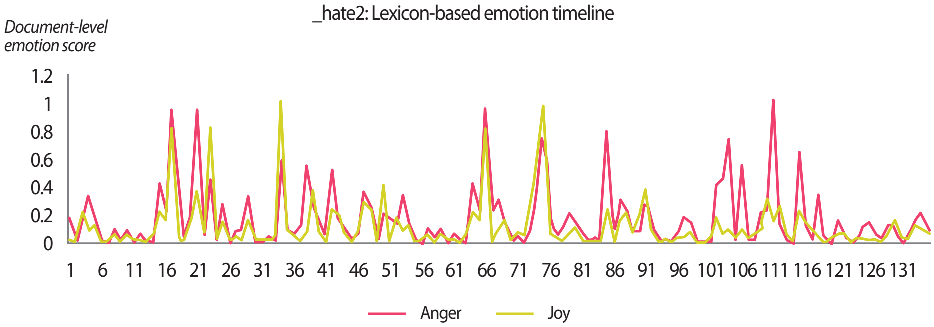


Fig. 5. Lexicon-based emotion timeline for synthetic user '_hate2.'

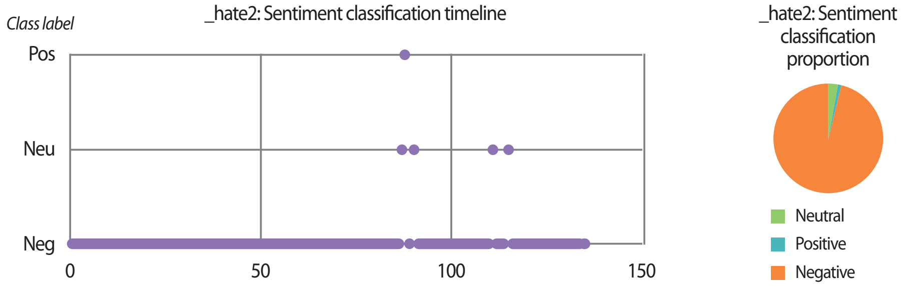


Fig. 6. Sentiment classification timeline and proportion for synthetic user '_hate2.'

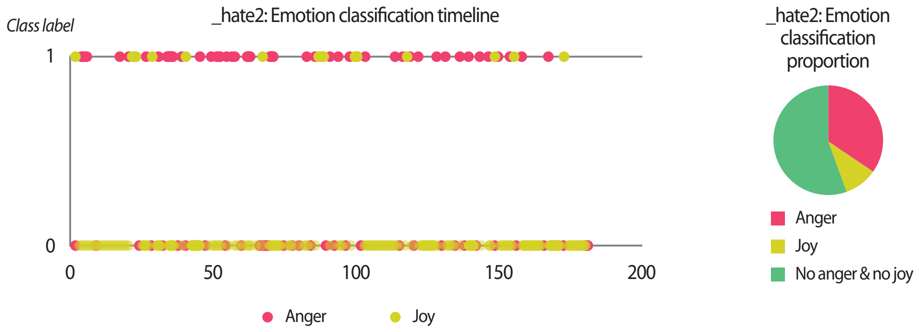


Fig. 7. Emotion classification timeline and proportion for synthetic user '_hate2.'

generated by the classifiers (Figs. 6 and 7) provide a lens into the anger and negativity in the user.

Turning now to the user scenario of the Enron user '_enron1,' we noticed that the three text analyzers for sentiment analysis and emotion analysis produced contradictory results. While the lexicon-based emotion analyzer and the LSTM sentiment classifier detected more emails that showed negative sentiments and emotions, the SVM emotion classifiers reported that the _enron1 user's emails did not contain any angry or hateful texts. Since this Enron user only produced 11 emails in his sent folder, we were able to label all emails manually to provide ground truths for verification (Table 8).

By comparing the results in Figs. 9 and 10 to the manual labels in Table 8, it can be seen that neither the sentiment classifier nor the emotion classifiers produced accurate predictions, although the predictions obtained by the sentiment classifier are slightly more accurate than the predictions generated by the emotion classifiers. However, from a closer inspection of the lexicon-based emotion timeline shown in Fig. 8, we found that there are two spikes in the anger emotion timeline that matched the manual labels in Table 8. One spike occurred on 21/2/2001 and the other spike occurred on 28/6/2001. Based on the textual contents of the two emails sent by '_enron1' on 21/2/2001 and 28/6/2001 (Fig. 11), it seems the lexicon-based emotion analysis has revealed a remarkable potential for the detection of emotions in electronic communications.

Table 8. Manually labeled sentiments and emotions for emails sent by '_enron1'

Email date	Sentiment	Anger	Joy
21/2/2001 23:09	-1	1	0
26/2/2001 5:34	0	0	0
28/2/2001 3:22	0	0	0
9/4/2001 10:06	1	0	0
1/5/2001 11:11	0	0	0
10/5/2001 1:56	0	0	0
11/5/2001 6:18	0	0	0
11/5/2001 7:03	1	0	0
15/5/2001 0:13	0	0	0
28/6/2001 1:25	-1	1	0
29/10/2001 16:38	0	0	0

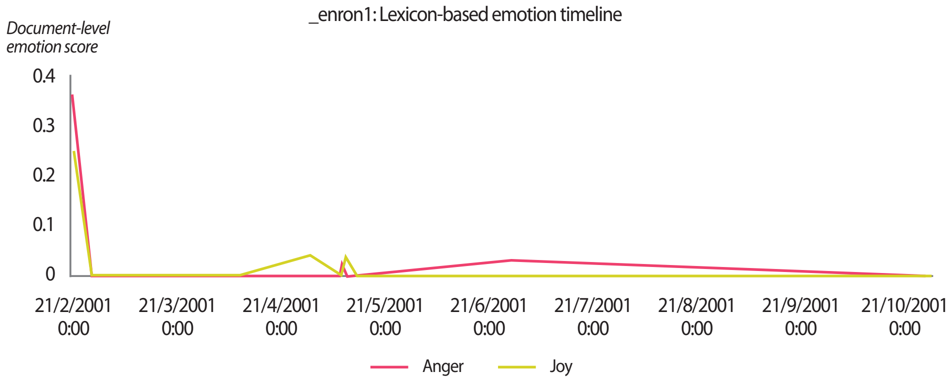


Fig. 8. Lexicon-based emotion timeline for Enron user '_enron1.'

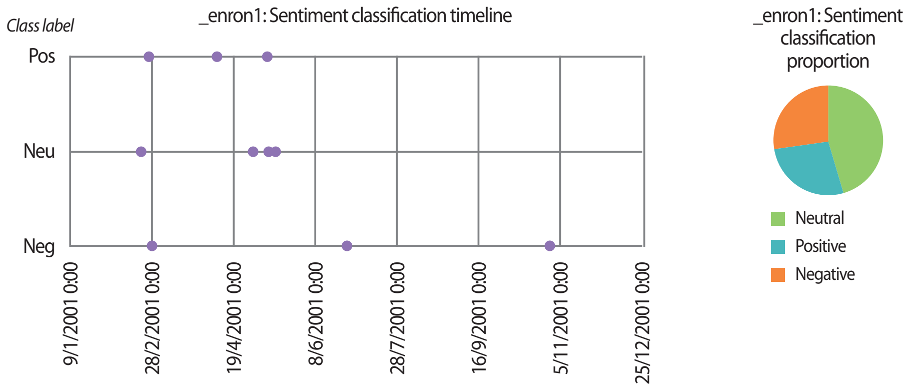


Fig. 9. Sentiment classification timeline and proportion for Enron user '_enron1.'

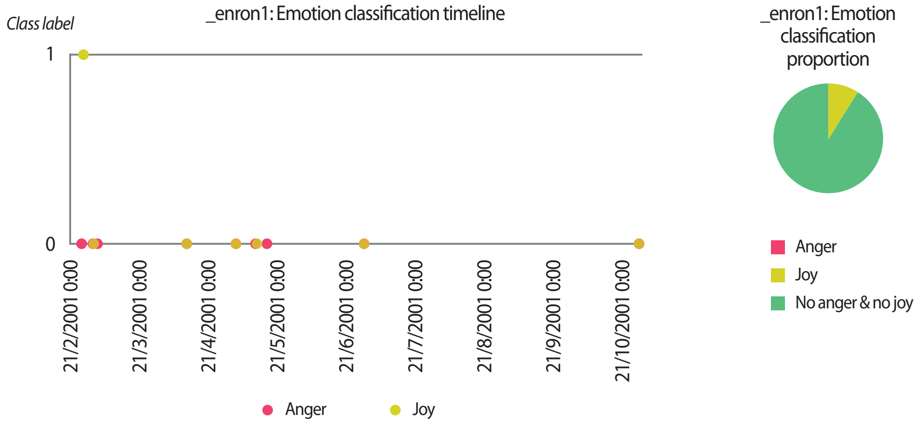
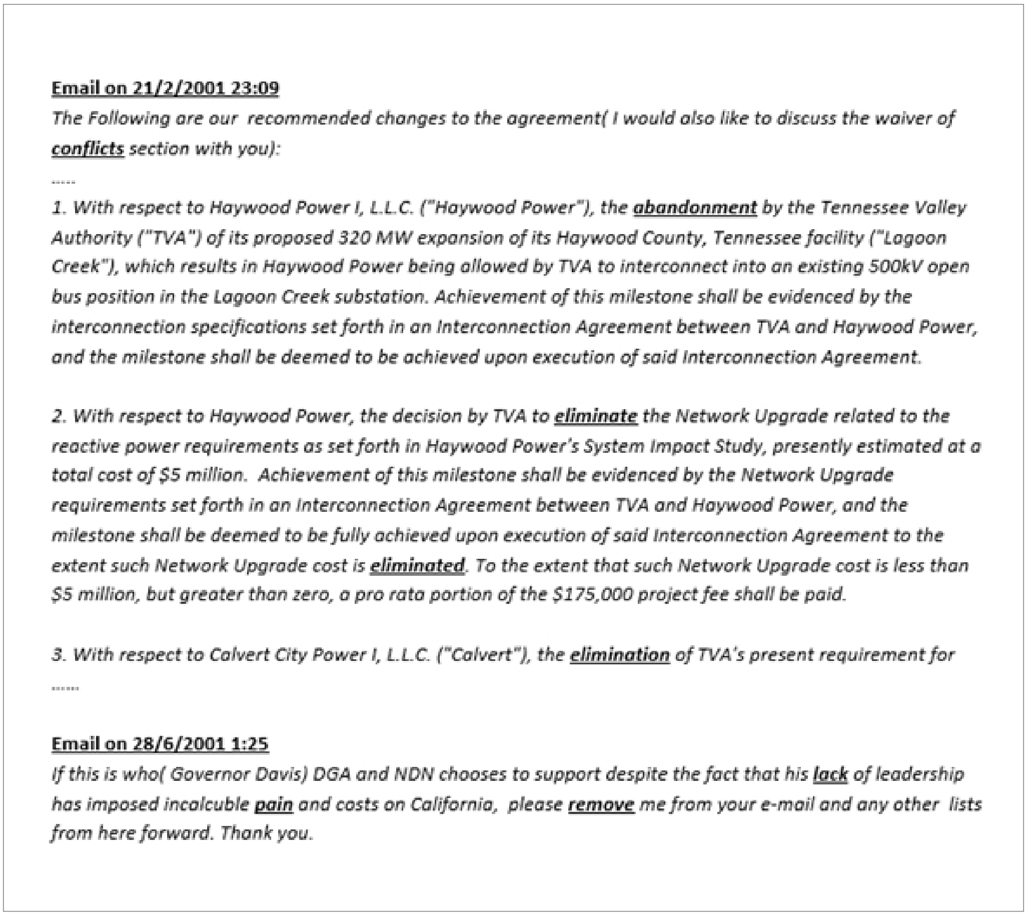


Fig. 10. Emotion classification timeline and proportion for Enron user '_enron1.'



Email on 21/2/2001 23:09

The Following are our recommended changes to the agreement(I would also like to discuss the waiver of ***conflicts*** *section with you):*

.....

1. With respect to Haywood Power I, L.L.C. ("Haywood Power"), the ***abandonment*** *by the Tennessee Valley Authority ("TVA") of its proposed 320 MW expansion of its Haywood County, Tennessee facility ("Lagoon Creek"), which results in Haywood Power being allowed by TVA to interconnect into an existing 500kV open bus position in the Lagoon Creek substation. Achievement of this milestone shall be evidenced by the interconnection specifications set forth in an Interconnection Agreement between TVA and Haywood Power, and the milestone shall be deemed to be achieved upon execution of said Interconnection Agreement.*

2. With respect to Haywood Power, the decision by TVA to ***eliminate*** *the Network Upgrade related to the reactive power requirements as set forth in Haywood Power's System Impact Study, presently estimated at a total cost of $5 million. Achievement of this milestone shall be evidenced by the Network Upgrade requirements set forth in an Interconnection Agreement between TVA and Haywood Power, and the milestone shall be deemed to be fully achieved upon execution of said Interconnection Agreement to the extent such Network Upgrade cost is* ***eliminated****. To the extent that such Network Upgrade cost is less than $5 million, but greater than zero, a pro rata portion of the $175,000 project fee shall be paid.*

3. With respect to Calvert City Power I, L.L.C. ("Calvert"), the ***elimination*** *of TVA's present requirement for*

....

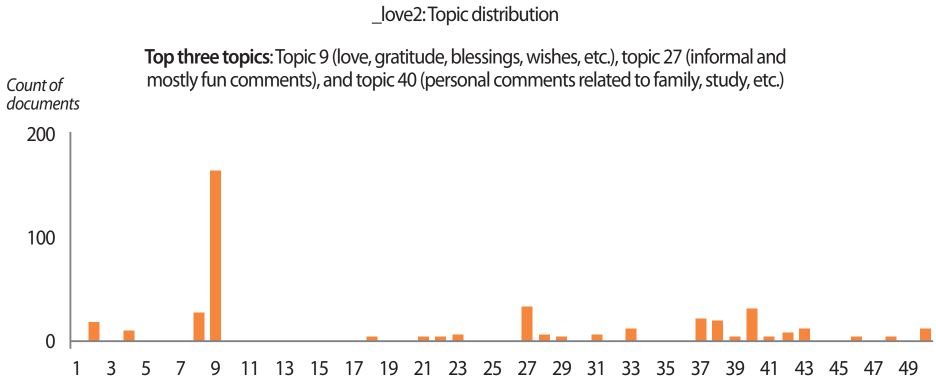
Email on 28/6/2001 1:25

If this is who(Governor Davis) DGA and NDN chooses to support despite the fact that his ***lack*** *of leadership has imposed incalcuble* ***pain*** *and costs on California, please* ***remove*** *me from your e-mail and any other lists from here forward. Thank you.*

Fig. 11. Emails sent by '_enron1' on 21/2/2001 and 28/6/2001. Words that signify anger and negativity are underlined.

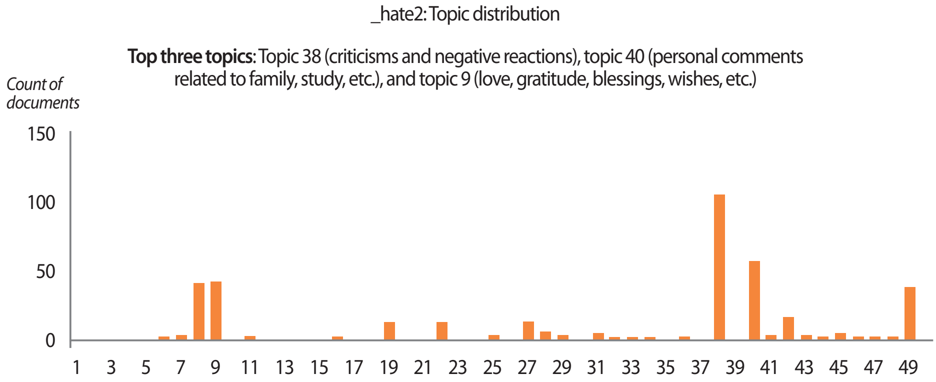
In addition to understanding individuals' psychological states from sentiment analysis and emotion analysis, we also extracted 50 key topics from the data corpus using the LDA topic model. The list of keywords composing the 50 topics is given in Appendix A. For each user scenario, the topic distribution and the top three topics discussed by the user are presented in the topic distribution charts (Figs. 12-14). The top topics revealed that the emails written by '_enron1' were mainly business-related, covering topics on payments and charges, regulatory concerns, secretarial communications, and so forth. On the other hand, the synthetic users tend to discuss matters revolving around more personal themes like blessings and wishes, fun comments, and criticisms. Another interesting finding from topic modeling is that key topics also reflect the sentiments and emotions of users. For instance, the top topic of '_hate2' (i.e., criticisms and negative reactions) shows that this user had a tendency to criticize and react negatively. Likewise, the positive behavioral patterns of '_love2' can be easily spotted from the user's top topics. These results suggest that topic modeling can be a sensible supplementary technique for assessing individuals' psychological states from their verbal communications.

Taken together, the user scenarios presented so far demonstrated how the unified psycholinguistic framework can keep the false alarm rate at a manageable level without compromising the detection of potential insider threats. The dilemma has been addressed in two ways: First, as seen from the scenarios of the synthetic users, the uncertainty in the insider tracking process can be reduced considerably when multiple text analyzers agree with each other; second, as demonstrated by the scenario of the Enron user, multiple text analyzers might complement each other and produce contradictory results in some cases. This scenario can be taken as an indicator for invoking a follow-up investigation by a human analyst to minimize the risk of missed catches.


_love2: Topic distribution

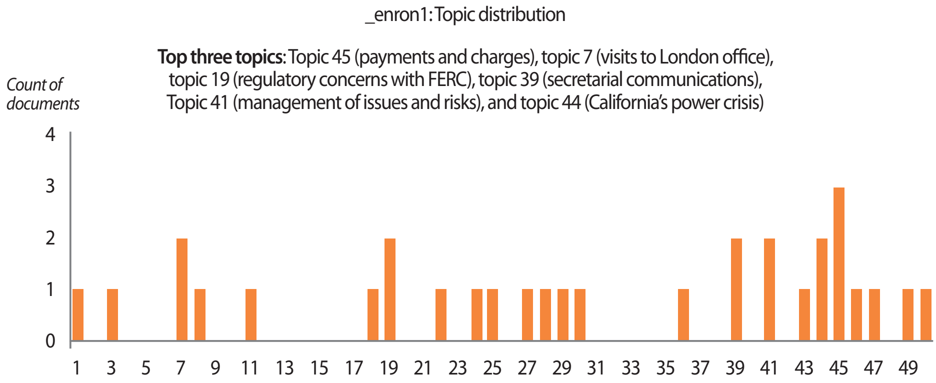
Top three topics: Topic 9 (love, gratitude, blessings, wishes, etc.), topic 27 (informal and mostly fun comments), and topic 40 (personal comments related to family, study, etc.)

Fig. 12. Topic distribution and the top three topics for synthetic user '_love2.'


_hate2: Topic distribution

Top three topics: Topic 38 (criticisms and negative reactions), topic 40 (personal comments related to family, study, etc.), and topic 9 (love, gratitude, blessings, wishes, etc.)

Fig. 13. Topic distribution and the top three topics for synthetic user '_hate2.'


_enron1: Topic distribution

Top three topics: Topic 45 (payments and charges), topic 7 (visits to London office), topic 19 (regulatory concerns with FERC), topic 39 (secretarial communications), Topic 41 (management of issues and risks), and topic 44 (California's power crisis)

Fig. 14. Topic distribution and the top three topics for Enron user '_enron1.'

5. CONCLUSION

The present study was undertaken to predict and detect insider threat by monitoring electronic communications for identifying individuals with troubling psychological patterns. To that end, we combined several text analysis methods—lexicon-based emotion analysis, LSTM sentiment classification, SVM emotion classification, and LDA topic modeling—to form a unified psycholinguistic framework. This is the first study that examined the use of multiple text analysis methods for psycholinguistic assessment in insider threat mitigation. The user scenarios presented in this paper demonstrated how the issue of the trade-off between the risk of missed catches and the false alarm rate can be attenuated. Overall, the text analyzers in our framework achieved acceptable performance. Further improvement is possible but is limited by some known constraints, such as highly imbalanced classes and the paucity of labeled data. In terms of directions for future research, considerably more work will need to be done to overcome these constraints and to achieve better accuracy in sentiment classification and emotion classification. Another natural progression of this work is to carry out the evaluation of the framework on data containing real or simulated insider threat.

REFERENCES

Axelrad, E. T., Sticha, P. J., Brdiczka, O., & Shen, J. (2013). A Bayesian network model for predicting insider threats. In *Proceedings of the 2013 IEEE Security and Privacy Workshops* (pp. 82-89). Piscataway: IEEE.

Azaria, A., Richardson, A., Kraus, S., & Subrahmanian, V. S. (2014). Behavioral analysis of insider threat: A survey and bootstrapped prediction in imbalanced data. In *IEEE Transactions on Computational Social Systems* (pp. 135-155). Piscataway: IEEE.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Brdiczka, O., Liu, J., Price, B., Shen, J., Patil, A., Chow, R., . . . Ducheneaut, N. (2012). Proactive insider threat detection through graph learning and psychological context. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy Workshops* (pp. 142-149). Piscataway: IEEE.

Brown, C. R., Greitzer, F. L., & Watkins, A. (2013). Toward the development of a psycholinguistic-based measure of insider threat risk focusing on core word categories used in social media. In *AMCIS 2013 Proceedings* (pp. 3596-3603). Atlanta: Association for Information Systems.

Brown, C. R., Watkins, A., & Greitzer, F. L. (2013). Predicting insider threat risks through linguistic analysis of electronic communication. In *Proceedings of the 46th Hawaii International Conference on System Sciences* (pp. 1849-1858). Piscataway: IEEE.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357.

Chen, Y., & Malin, B. (2011). Detection of anomalous insiders in collaborative environments via relational analysis of access logs. In *Proceedings of the First ACM Conference on Data and Application Security and Privacy* (pp. 63-74). New York: ACM.

Cherry, C., Mohammad, S. M., & De Bruijn, B. (2012). Binary classifiers and latent sequence models for emotion detection in suicide notes. *Biomedical Informatics Insights*, 5(Suppl 1), 147-154.

Colwill, C. (2009). Human factors in information security: The insider threat–Who can you trust these days?. *Information Security Technical Report*, 14(4), 186-196.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.

Eberle, W., Graves, J., & Holder, L. (2010). Insider threat detection using a graph-based approach. *Journal of Applied Security Research*, 6(1), 32-81.

Gheyas, I. A., & Abdallah, A. E. (2016). Detection and prediction of insider threats to cyber security: A systematic literature review and meta-analysis. *Big Data Analytics*, 1(1), 6.

Greitzer, F. L., Frincke, D. A., & Zabriskie, M. (2010). Social/ethical issues in predictive insider threat monitoring. In M. J. Dark (Ed.), *Information assurance and security ethics in complex systems: Interdisciplinary perspectives* (pp. 1100-1129). Hershey: IGI Global.

Greitzer, F. L., Kangas, L. J., Noonan, C. F., Brown, C. R., & Ferryman, T. (2013). Psychosocial modeling of insider threat risk based on behavioral and word use analysis. *e-Service Journal*, 9(1), 106-138.

Grijalva, E., Newman, D. A., Tay, L., Donnellan, M. B., Harms, P. D., Robins, R. W., & Yan, T. (2015). Gender differences in narcissism: A meta-analytic review. *Psychological Bulletin*, 141(2), 261-310.

Ho, S. M., Hancock, J. T., Booth, C., Burmester, M., Liu, X., & Timmarajus, S. S. (2016). Demystifying insider threat: Language-action cues in group dynamics. In *Proceedings of the 49th Hawaii International Conference on System Sciences* (pp. 2729-2738). Piscataway: IEEE.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9, 1735-1780.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2), 251-257.

Kandias, M., Mylonas, A., Virvilis, N., Theoharidou, M., & Gritzalis, D. (2010). An insider threat prediction model. In S. Katsikas, J. Lopez, & M. Soriano (Eds.), *Lecture notes in computer science: Vol. 6264. Trust, privacy and security in digital business* (pp. 26-37). Berlin: Springer.

Kandias, M., Stavrou, V., Bozovic, N., Mitrou, L., & Gritzalis, D. (2013). Can we trust this user? Predicting insider's attitude via YouTube usage profiling. In *Proceedings of the 2013 IEEE 10th International Conference on Ubiquitous Intelligence and Computing and 2013 IEEE 10th International Conference on Autonomic and Trusted Computing* (pp. 347-354). Piscataway: IEEE.

Kiser, A. I., Porter, T., & Vequist, D. (2010). Employee monitoring and ethics: Can they co-exist?. *International Journal of Digital Literacy and Digital Competence*, 1(4), 30-45.

McCrae, R. R. (2010). The place of the FFM in personality psychology. *Psychological Inquiry*, 21(1), 57-64.

Mohammad, S. M. (2012). #Emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics* (pp. 246-255). Stroudsburg: Association for Computational Linguistics.

Mohammad, S. M. (2015). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In H. L. Meiselman (Ed.), *Emotion measurement* (pp. 201-237). Duxford: Woodhead Publishing.

Mohammad, S. M., & Bravo-Marquez, F. (2017). Emotion intensities in tweets. In *Proceedings of the Sixth Joint Conference on Lexical and Computational Semantics* (pp. 65-77). Stroudsburg: Association for Computational Linguistics.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3), 436-465.

Mohammad, S. M., Zhu, X., Kiritchenko, S., & Martin, J. (2015). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4), 480-499.

Myers, J., Grimaila, M. R., & Mills, R. F. (2009). Towards insider threat detection using web server logs. In *Proceedings of the 5th Annual Workshop on Cyber Security and Information Intelligence Research: Cyber Security and Information Intelligence Challenges and Strategies* (p. 54). New York: ACM.

Parker, D. B. (1998). *Fighting computer crime: A new framework for protecting information*. New York: Wiley.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2001). *Linguistic inquiry and word count: LIWC 2001*. Retrieved February 5, 2019 from http://www.depts.ttu.edu/psy/lusi/files/LIWCmanual.pdf.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1), 547-577.

Plutchik, R. (1982). A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5), 529-553.

Schultz, E. E. (2002). A framework for understanding and predicting insider attacks. *Computers & Security*, 21(6), 526-531.

Shaw, E. D., & Fischer, L. F. (2005). *Ten tales of betrayal: The threat to corporate infrastructure by information technology insiders analysis and observations*. Retrieved February 5, 2019 from http://www.dtic.mil/dtic/tr/fulltext/u2/a441293.pdf.

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 and Human Behavior*, 37(4), 267-275.

Wood, B. (2000). An insider threat model for adversary simulation. In *Proceedings of the Workshop on Mitigating the Insider Threat to Information Systems* (pp. 41-48). Arlington: RAND.

APPENDIX A

Keywords for the 50 Topics Generated from Topic Modeling

Topic ID	Keywords
1	power project system plant gas fuel approved transmission development cost production process run load opportunity request gathering capacity proposal approval
2	hope great happy management game home tickets tonight fun nice night sat weekend guys party play christmas birthday stay hey
3	kay ge stuff don email ben working lisa lee docs ena sheila change llc closing emails reminder ready equipment advise
4	time free weeks couple feel wanted told lunch interested guys week meet called today forward taking thought place ago asked
5	mark enron greg mary david works delainey john taylor counsel ed andy liz tim general group fletch president whalley services
6	deal deals volume month jan ces sitara desk storage feb oct nov created booked volumes dth sale entered book changed
7	mail london message houston df office voice received check leave pls trip flight calendar hotel address number wed messages travel
8	people money told years big story made lot things find country worth talking didn making world times hard past running
9	love life day feel man time heart dear ve make world god true things mind happy baby met words kind
10	enron corp houston north america eb texas street smith legal debra perlingiere department sara shackleton fax phone gisb tx ph
11	business review employees global enron process ena performance prc employee rick year focus feedback commercial meetings level training committee management
12	phone fax sara st susan carol confirm cell clair lawyer send spoke ss confirms lawyers confirmation handling suzanne leslie reach
13	trading trade financial counterparty book products eol physical credit master trades online canada counterparties product swap power transactions books legal
14	meeting monday friday pm thursday afternoon tuesday attend wednesday morning meet tomorrow office schedule week scheduled set noon attending unable
15	call give jeff discuss ll conference tomorrow today folks number set chance df thoughts heather asap apologies answer srs voicemail
16	report desk data information open phillip position west positions spreadsheet reports update items run mat find format track summary var
17	attached comments draft agreement questions form version forward revised request discussed latest final document approval hesitate prepare proposed agreements attaching
18	sally office team week meeting plan operations group brent patti work key james memo meetings join working role review calgary
19	information ferc options option case provide policy specific terms part concerns aware regulatory made related confidential additional including concern reference
20	mid kate dec changed columbia pst deals deal ees broker morgan bpa avista epmi stephanie mc pget cob sp enpower
21	talk today chris yesterday ben tomorrow morning talked wanted didn robin matt hey lets brian joe chicago ya don working
22	make access put bill note group page set read line idea computer change work picture suggestions direct time link mind
23	day time date days june july december april march plan october end vacation august september january schedule november dates back
24	person left kevin check info message michael city eric gary rob tx pass jason east portland julie leave asked ckm
25	agreement credit language ena section party master isda guaranty transaction agreements parties termination contract paragraph assignment respect dated transactions law
26	gas daily index price basis el socal paso mmbtu east pg day pool volumes west hpl curve point ena flow
27	don ve didn guy thing remember ll back guys bad pretty doesn couple stuff thought half figure worry couldn finally
28	year program game team play big center ut university end top students early recruiting round school turn football national won
29	market price prices million year power term demand costs supply cost based long increase high percent buy summer sales cap
30	fyi jim file kim info tom michelle fine notes dg files update lynn questions linda jennifer coordinate harry alan fred
31	issues issue make order problem agree response understand problems decision point future clear discuss made credit end comment suggest line
32	contract rate capacity tw service contracts firm point delivery pipeline term df tariff ena rates order release fuel transportation points
33	good great hope sounds job things work hear glad pretty luck guys thought hey talk interesting summer care trip nice
34	send copy letter review sign signed documents print copies received executed sending lynn elizabeth marie original document attachment attached signature
35	john dave forward paul rick resume dan interest manager interested interview eric congratulations frank peter director robert follow michelle directly
36	buy mw bid short hours offer blackberry sale wireless handheld sell peak hour bill day real schedule power purchase show
37	work week time make back move lot start working good things hard long moving weather ready doesn rest busy happen
38	don people question sense makes answer fact make site read opinion doesn case find problem long situation isn wrong website
39	vince jeff presentation steve ken mr shirley assistant lay fyi skilling invite kaminski invitation stinson speak karen join sherri george
40	school work years family class part visit hope live dr remember care day children miss find write don parents pictures
41	group risk involved model project management support working provide current experience manage issues position area process structure large additional including
42	room house place bring car front church side black water small red hit stop white walk door parking boat put
43	ll back don ve heard haven today check email guess fine hear waiting chance wait touch password checked talk chat
44	california state power energy davis utilities edison utility electricity contracts commission puc customers governor dwr billion public plan pg bill
45	pay amount payment cash contract notice tax paid due account charge made period purchase event charges money receive fee days
46	facility test site unit prior permit units transwestern completed required submitted request issue air construction additional mexico system activities agency
47	mike list email contact send address bob add scott names distribution asked update stacey jay message david missing forwarded richard
48	number pl numbers questions change sheet put correct call problem find stan errol note today louise added cindy verify worksheet
49	enron company energy services business trading gas marketing markets stock companies natural capital resources corporation interest europe officer york board
50	night weekend dinner home saturday sunday town house dad tonight mom leaving leave kids fun friday coming friend austin plans