A Machine Learning Approach to Detect Insider Threats in Emails Caused by Human Behaviours

A. Michael and J.H.P. Eloff

Department of Computer Science, University of Pretoria, South Africa
e-mail: tonia.michael94@gmail.com, eloff@cs.up.ac.za

Abstract

In recent years, there has been a significant increase of human behaviour driven insider threats within organizations and these have caused massive losses and damages. Due to the fact that emails are a crucial part of the modern-day working environment, many of those insider threats exist within the organizations' email infrastructures. It is known that amidst "business-as-usual" emails sent by employees, there are non-company related mail and perhaps mail containing malicious activity and unethical behaviour. These types of insider threats are most often caused by employees who have legitimate access to an organisation's resources, such as the servers and non-public data, but abuse these privileges for various reasons such as personal gain or perhaps to inflict malicious damage on the employer. The problem is that due to the high volume and velocity of email, it is almost impossible to minimise the risk of these type of insider threat activities through techniques such as filtering and rule-based systems. The research presented in this paper aims to minimise the risk of human behaviour driven insider threats via email systems, by employing a machine learning based approach. This is done by studying and creating categories of malicious human behaviours that insiders possess, and, mapping these to phrases that would appear in email communications. A large email dataset is classified according to behavioural characteristics of employees. Machine learning algorithms are employed to identify commonly occurring insider threats and to group the occurrences into insider threat classifications.

Keywords

Cyber-security, insider threats, insider threat detection, machine learning

1 Introduction

Over the past year, according to IBM (IBM, 2019), there has been a fourfold increase in spam email causing insiders to accidentally lose data. IBM also reported that 82% of insider and privilege misuse compromises took months, if not years, to be detected. The threat caused by insiders to organisations leverages various vulnerabilities in organisations of which email infrastructures are one of the weakest points exploited, either accidentally, or, maliciously by insiders. The United States Cybersecurity Magazine (Ali, 2018) recommends that organisations focus on the monitoring of employee behaviour in an attempt to minimise the risk of insider threats.

In today's working world, emails have become the most common form of communication and businesses simply cannot function without emails. Whilst emails have proven to be an effective means of corporate communication, they have also

introduced a new platform for cyber security breaches and criminal activities to take place (Butkovic, et al., 2013).

A real-world example of this exists within the publicly available Enron email corpus, which contains the communication between Chris Germany and Victor Lamadrid (Lepinsky, 2013). The emails reveal that the two parties were planning the shutdowns of power plants and blackouts in California, such that the demand and price for power would increase to greatly increase the profits of the executives (Lepinsky, 2013). Furthermore, in the same dataset, there was also evidence found of employees manipulating balance sheet data, corruption and bribery of important officials (Sashikanth, 2015). These types of threats and incidents, caused by human behaviour, within the cybersecurity domain, are referred to as insider threats. Insider threats within organizations can be materialised by either humans or machines. For example, phishing attacks can be conducted by malicious attackers or by bots. For this study, however, only insider threats caused by humans are investigated. According to Kowalski et al. (2008), an insider threat is defined as a threat caused by a malicious current or former employee, or someone who has previously been affiliated with the organization, who has had legitimate access to the company's network, system and non-public data. Furthermore, this user exploited their access such that the confidentiality, integrity, or availability of the organization's data or systems is compromised.

The authors extend this definition of insider threats further to include threats caused by non-malicious negligent employees (Kowalski, et al., 2008). These employees are classified as insiders because they are most often tricked, via means of social engineering techniques, to leak sensitive company data. Furthermore, these employees are careless about utilizing security mechanisms or following proper security procedures. Thus, it must be noted that there are several different reasons for why employees partake in insider threats, both malicious and non-malicious. These can involve personal or financial gain, negligence, sabotage of the employer, or a need for revenge due to feelings of disgruntlement or anger towards the organization (Young, et al., 2014) (Brown, et al., 2013) (White & Panda, 2009).

It is however, difficult for organizations to detect and differentiate between normal business behaviour and malicious behaviour in email communications (Chi, et al., 2016). Furthermore, employees' email communications are usually not inspected by the employers, due to the large number of employees existing in large organizations, as well as the lack of time and the necessary software or infrastructure. In addition, some organizations do not govern the use or misuse of company email accounts.

The research presented in this paper is relevant and necessary because, due to the large volumes of emails sent by employees in organizations today, insider threats within these communications, could be going undetected. As such, these are potentially placing organisations in dire risk of financial and reputational losses, disruption of operations, and harm to specific individuals (Kowalski, et al., 2008). Therefore, proper security mechanisms need to be investigated and implemented in organizations to address the problem of insider threats, and that is what this research serves to propose.

This research aims to propose an approach to assist organisations with detecting insider threats caused by employees, through identifying and classifying insider threats in corporate email communications. The contributions of research results within this paper are summarized, as shown below:

- To establish a list of the main types of insider threats and from that, to establish the human behaviours associated with these insider threats. In addition, to identify certain phrases that would be found in corporate emails, that can be mapped to these identified behaviours, causing insider threats.

- To develop an insider threat classification prototype based on the phrases identified, using machine learning techniques. This includes acquisition of a large email corpus, applying data cleaning techniques, and running machine learning algorithms.

This paper is structured as follows: Section 2 covers a study of existing work done to classify large email datasets and detecting insider threats. Section 3 covers the requirements for the prototype, and section 4 contains the high-level design of the solution. Section 5 and 6 contain the experimental results and the discussion thereof respectively, following the implementation of the prototype and the experiments conducted. Finally, section 7 contains the conclusions and scope for future work.

2 Background and related work

Insider threats have caused great damage to large organizations due to the fact that these were not detected and mitigated before they could cause harm. In 2016, a disgruntled employee responsible for the Citibank IT systems, brought 90% of the networks down due to a poor performance review obtained from management (Cluley, 2016). This employee had the technical skills, a wide range of system access, as well as a strong motivation to carry out this action. Another example is one from the Enron scandal, where after scanning the leaked email datasets, it was found that top executives John Lavoreto and Tim Belden were both aware that Enron was actively manipulating the Canadian energy market in August 2000 (Tribolet, 2016) (Cukierski, 2015).

It is thus clear that if insider threats are not found and mitigated before they cause harm to the business, the damages could be severe. This study deals with the detection of insider threats within corporate email communications. insider threats lurking within emails have certain characteristics and are initiated by employees who possess certain characteristics.

Research has been conducted where various categories of insider threats have been devised, based on human behaviours, in order to aid machine learning detection of these threats. Young et al. (2014) employ a technique using Scenario-based detectors where certain real world actions and behaviours relating to insiders are grouped according to types of insider threats. These types of insider threats are then used in clustering algorithms to sort emails into classes (Young, et al., 2014). There are three

main types of insider threats, namely Insider IT Sabotage, Insider Intellectual Property Theft and Insider Fraud (Spooner, et al., 2018) (Claycomb, et al., 2013), (Cappelli, et al., 2012), (Munshi, et al., 2012). Young et al. (2014) describe a careless user as an additional type of insider threat that does not act with malicious intent. This type of user is placed in a category called Negligence in this research (Young, et al., 2014). These categories of insider threats are summarized from related work (Spooner, et al., 2018), (Whitman, 2016) (Chi, et al., 2016), (Young, et al., 2014), (Cappelli, et al., 2012), (Nizamani, et al., 2014) (Kowalski, et al., 2008), in Table 1, compiled by the authors of the research at hand. These categories are used in the experimentation of this research.

Insider Threat Type	**Explanation**
Insider IT Sabotage	An employee has started to resent the company and becomes disgruntled for a certain reason, such as a poor performance review, and therefore desires to inflict harm on the company (Cappelli, et al., 2012). The employee chooses to misuse his/her privileged system access to cause harm (Chi, et al., 2016). This employee displays malicious behaviour. Typical behaviour of an employee of this insider threat type, may involve causing harm by destroying the company's hardware and software, as well as tampering with, or stealing the company's data (Kowalski, et al., 2008). Specific words that indicate a potentially disgruntled employee may include "wasted efforts", "not happy at work", "uncertain about the future", to name a few.
Insider Intellectual Property Theft	An employee who has worked on the creation of a system or data within the organization and feels a sense of ownership over this data. As such, the employee feels entitiled to steal his/her work and as such, will display this unethical and malicious behaviour (Young, et al., 2014). Examples of phrases associated with emails sent by this type of employee are "split the difference", "where did my money go" (Whitman, 2016)
Insider Fraud	An employee who displays unethical and malicious behaviour by engaging in illegal activities for various reasons such as to harm the organization, or for personal gain (Young, et al., 2014). Various indicators of this type of insider threat are the use of words such as "money", "share", "percent", as well as reference to "advocates" and "relations" (Nizamani, et al., 2014) These employees tend to collude with other internal employees (Spooner, et al., 2018)
Negligence	An employee of this type does not act with malicious intent but displays careless, accidental or naive behaviour in terms of following the proper security procedures (Young, et al., 2014). Such an employee would be at risk of responding to a phishing attack. Example phrases that would be expected in an email that targets a negligent employee would include emergency words and phrases such as "urgent", "as soon as possible" (Nizamani, et al., 2014)

Table 1: Insider threat categories compiled by the authors based on categories shown by Spooner et al. (2018) and Young et al. (2014).

One of the approaches to detect insider threats in Email datasets that are too large to label manually, is by using clustering techniques (Alsmadi & Alhami, 2015). Okolica, et al. (2007) applied a clustering technique by firstly obtaining the Enron dataset and tokenizing each email into a group of words. Inflections of the same word were identified, the root word was extracted, and the multiple occurences of the same root words in one email were combined (Okolica, et al., 2007). This is known as stemming and an example of this can be shown where the words "colludes" and "colluding" are all inflections of the root word "collud". Stemming is used to improve the overall clustering accuracy and reduce the dataset size (Okolica, et al., 2007).

A frequency count was used to represent the number of times a word appeared in an email, as well as a frequency count to represent the number of times an individual's name appeared in the email body (Okolica, et al., 2007). These were sent to a tool called Author Topic (Rosen-Zvi, et al., 2004), which generated 48 topics used as categories, which are referred to as "centroids" in clustering. These categories were created based on the most frequently used words. In the research at hand, the cluster centroids that are used to group similar email data together, are predefined. These specifically relate to each of the insider threat types defined by Cappelli et al. (2012) which are shown in Table 1. Some of the topics chosen by the automated tool, in the paper from Okolica et al. (2007), such as the topic "Senior Management" were broad and as such the data in this cluster contains a lot of general business activity. Okolica et al. (2007) found that where a more focused topic such as "Research" was used, the results in the cluster were more accurate. In detecting insider threats, Okolica et al.(2007) showed that clustering was an effective means of labelling and grouping a very large email dataset.

Alsmadi and Alhami (2015) created an email clustering and classification model where five predetermined categories, namely "personal", "job", "profession", "friendship" and "others", were created for the clustering of emails from a large dataset. Methods such as tokenizing, cleaning and stemming were applied to prepare the data before clustering (Alsmadi & Alhami, 2015) (Nizamani, et al., 2014) (Mujtaba, et al., 2017). To label the large dataset, the K-means algorithm was used. Documents within the dataset were randomly selected as centroids and a similarity check took place, for each email, to cluster similar emails (Hussain & Qamar, 2014).

Alsmadi and Alhami (2015) also ran classification algorithms, after clustering took place, to enhance the experiment. This is because supervised learning involves manual intervention through the use of preclassified training data, which allows for more accurate results. Furthermore, metrics could be provided from several iterations of these classification algorithms to measure speed and accuracy. Due to this, both unsupervised clustering and supervised classification algorithms are applied to the email dataset in the experiment in the research at hand.

Mayhew et al. (2015) stated that the use of unsupervised K-means clustering, with supervised Support Vector Machine (SVM) classifiers, is the best combination for ensuring both quality, performance and efficiency. Mujtaba et al. (2017) stated that the top 3, most frequently used supervised machine learning classifiers in cross

domain big data research are SVM, Decision Trees and lastly Naïve Bayes classifiers. These three are therefore used for supervised classification in the research at hand.

Brown et al. (2013) attempted to determine whether there are prominent personality traits exposed in the emails of the users who carried out insider threats. The Enron email corpus was also consulted in this work. The five main groups of character traits that were used to measure personality include agreeableness, conscientiousness, neuroticism, extraversion and openness. Various word lists were devised for each of these categories containing email phrases that could be used by a person who possesses the given trait. Each email was scanned and if a match was found between the email and one of the word list categories, the user's score for this category was incremented. This approach of constructing and using word lists to label emails in the different categories is applied in the research at hand. This is because it provides a more automatic means to label a large dataset. The scoring of each email according to the different categories is also applied in this research.

In summary, various techniques applied in past research are incorporated in the approach to detect insider threats in the research at hand. Thorough data cleaning and normalizing take place, such that refined data is used in the experimentation. Past research has shown that clustering is an effective tool to group a large dataset into different categories. The centroids chosen would need to encompass a wide variety of words that similar emails could share. Word lists are also used as a classification approach. This is because a greater variety of words and phrases can be included to allow a more accurate classification. In addition, a scoring mechanism helps to improve classification accuracy, because each sentence in an email is scanned and an overall score for each insider threat category is assigned to the email, based on the word lists. Supervised machine learning is also used, taking in the labelled training set. The performance and accuracy of each machine learning classifier (SVM, Naïve Bayes and Decision Tree) is obtained to determine the approach most effective in detecting insider threats.

3 Establishing the requirements

The following have been identified as requirements for developing a prototype environment for the research at hand. These requirements are gleaned from past research. The requirements are grouped into Functional and Technical requirements:

Functional requirements:

- Detect the following types of insider threats based on human behaviours in a corporate email dataset (Spooner, et al., 2018) (Young, et al., 2014):
    - Insider IT Sabotage
    - Insider Intellectual Property Theft
    - Insider Fraud
    - Negligence

Technical requirements:

- Use a large enough email dataset pertaining to a large organization that contains a wide variety of different emails and known cases of malicious and non-malicious activities, specifically with reference to insider threats (Leber, 2018) (Alsmadi & Alhami, 2015).
- Use the body of each email as textual data in the clustering and classification process (Mujtaba, et al., 2017) (Alsmadi & Alhami, 2015).
- Use a technique of scoring to score each email in each of the categories, IT Sabotage, IT Insider Intellectual Property Theft, IT Fraud and Negligence (Young, et al., 2014) (Brown, et al., 2013) (Cappelli, et al., 2012).
- Develop a model, that includes supervised and unsupervised machine learning techniques, to classify and label email bodies according to the main types of insider threats.
- Compare the accuracy of results of these machine learning algorithms by using metrics such as accuracy and precision averages (Mujtaba, et al., 2017).

4 High level design of the prototype solution for the insider threat detection approach

The diagram in Figure 1, shown in the Unified Modelling Language, based on work done by van der Walt and Eloff (2018) contains the components of the prototype solution.

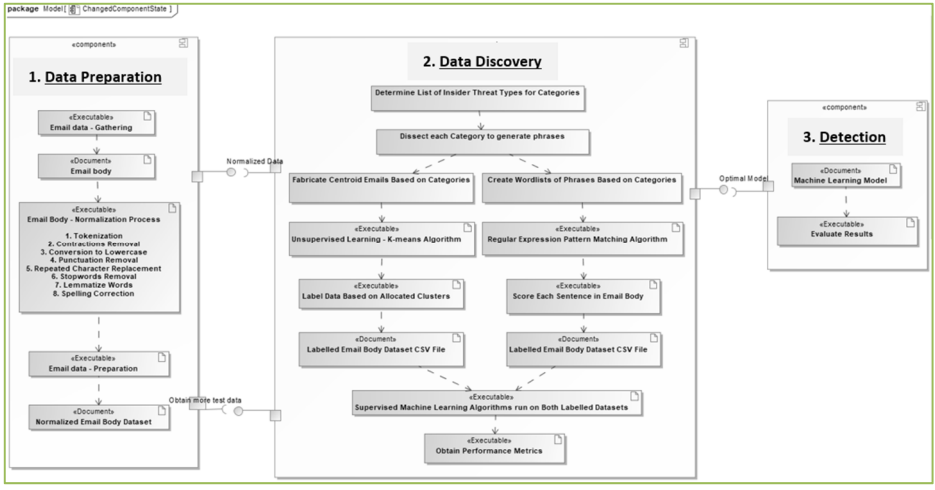

Figure 1: UML Component Diagram of the proposed prototype

A high-level discussion of the three main components in Figure 1 follows.

1. Data Preparation – The Enron email data that is obtained from Kaggle.com (Cukierski, 2015) , consists of email bodies, subjects and employee names. The email body of each email in the dataset is used in the experiment. Note that the email dataset consists of emails sent by various employee as well as emails received by these employees. To prepare the content, each email body is tokenized or broken up into individual words in an array (Okolica, et al., 2007). An example is where the phrase "the message we're trying to get across?", taken from the Enron dataset, transformed to separate words as follows; "the", "message", "we're", "trying", "to", "get", "across". Each tokenized email is then normalized according to the following steps: removal of contractions, converting to lower case, removal of punctuation, replacing repeated characters, removing stop words and correction of spelling (Mujtaba, et al., 2017). After normalization, the example email sentence used above shows as "message we try get across". This data must be cleaned for machine learning.

2. Data Discovery – Both supervised and unsupervised machine learning algorithms are utilized in the process of detecting insider threats. The aim is to label the dataset based on the main types of insider threats based on human behaviours shown in Table 1 (Cappelli, et al., 2012). A Regular Expression Pattern Matching classifier is applied, comparing the words in the email body to those in the defined word lists. For example, one email contains the phrase "I cannot keep working under these orders" and the algorithm matches this with a similar phrase in the Sabotage word list. Each sentence is checked against the word lists and if a match is found, a score for the associated category is incremented. The email in the example above causes the score in the Sabotage category of the email's scoring dictionary to be incremented (refer to Table 2 for the dictionary layout). The category with the highest acceptable score for a given email is chosen as the label for the email. This is stored as a new labelled dataset in a CSV file. An unsupervised K-means clustering algorithm is also used in this step to cluster the data. Based on the clusters, the data is labelled. This is stored as dataset file 2. Both labelled datasets are then used to train the supervised machine learning algorithms. Metrics are obtained to measure performance of the machine learning algorithms for each of the CSV files.

3. Detection – Due to the fact that an email dataset is fairly large, the processes of labelling the email dataset shown in the Data Discovery step are automated. There is a program that runs the machine learning classifiers sequentially, using the labelled datasets as training data, so that the classifiers are able to identify insider threats in a testing dataset of unclassified emails (Alsmadi & Alhami, 2015).

5 Experimentation results

The first two components shown in Figure 1 in section 4, namely Data Preparation and Data Discovery were executed with a test set of email data, as a proof of concept for this approach. A total of 10000 email records, from a total set of 500000, were

extracted from the Enron email dataset in CSV format, to be used for the research at hand (Cukierski, 2015). This CSV data was used in a Python environment (Python, 2019) and read in with the "pandas" library (Pandas, 2019). This section briefly displays the results obtained in the clustering and classification tasks of the Data Discovery component in Figure 1.

5.1 Data Discovery: Regular Expression Pattern Matching Algorithm

With this classification technique, all wordlists were compared with each normalized email to obtain matches ensuring that scoring could take place to determine which labels should be assigned. The results of the scoring function are shown in section 5.2.

5.2 Data Discovery: Score Each Sentence in Email Body

The scoring dictionaries of the highest scored emails are shown below in Table 2, extracted from the log file printed during the execution of the Regular Expression Pattern Matching classifier. Each individual category was incremented when a match was found within each sentence of an email. Each category was then divided by the total number of sentences in the given email. This yielded a value between 0 and 1. If the value was above a given threshold, in this case 0.5, the email was allocated that category as the label in a new CSV file. A snippet of the CSV file labelled with this approach is shown in Figure 4, Section 5.5.

Rank	Dictionary Scores
1	{"number sentences": 120, "total score": 25}, {"fraud": 19, "ipthief": 0, "negligence": 4, "non_malicious": 0, "sabotage": 0}
2	{"number sentences": 102, "total score": 23}, {"fraud": 17, "ipthief": 0, "negligence": 5, "non_malicious": 0, "sabotage": 0}
3	{"number sentences": 154, "total score": 22}, {"fraud": 8, "ipthief": 0, "negligence": 11, "non_malicious": 0, "sabotage": 0}
4	{"number sentences": 87, "total score": 19}, {"fraud": 0, "ipthief": 0, "negligence": 17, "non_malicious": 0, "sabotage": 0}
5	{"number sentences": 52, "total score": 17}, {"fraud": 9, "ipthief": 0, "negligence": 5, "non_malicious": 0, "sabotage": 0}

Table 2: Results in JSON format, showing the dictionaries containing scores for each insider threat type of the top 5 highest scored emails in the dataset.

The aforementioned was one of two approaches used in the prototype to label the data. Section 5.3 contains the second approach.

5.3 Data Discovery: Unsupervised Learning – K-means Algorithm

The K-means classifier was run to create clusters of the email dataset. The centroids that were fabricated were structured as emails to contain specific phrases linking to each insider threat type (as shown in Figure 2). There was one centroid created per insider threat type. The K-means classifier used the fabricated centroids to create clusters of the normalized email dataset.

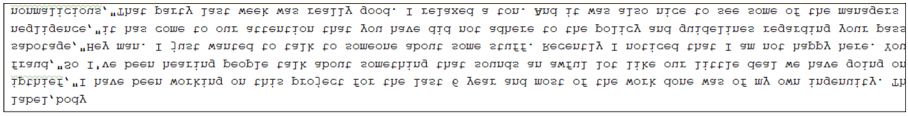

Figure 1: Fabricated centroids for K-means algorithm

5.4 Data Discovery: Label Data Based on Allocated Clusters

Insider Threat Type	Centroid (Fragment of Actual Centroid)	Closest Email (Fragment of Actual Email)	Similarity (%)
Insider Intellectual Property Theft	**Centroid 1:** "not entirely will split difference. submit enquiry finance find someone embezzle account payment product. want know money go entitle 75 profit generate trade formula improvement please advise"	"gary list goal send financial trade group . please review make change email. please take minute make change return asap. financial trade create 110 million gross margin 80 million.1 successful equity trade group area can 20 take advantage en ron 0 1 network competitive advantage establish trade business start up london tokyo thanks"	34.69
Insider Fraud	**Centroid 2:** "send communication house email cold result possible lawsuit . let get stuff way soon possible legal step start ask question . strong lawyer hand would rather not involve matter . know not exactly play rule neither . want keep manipulate energy price suggest keep lid . not let loose tongue cause legal conflict . play card right may not illegal eye stakeholder ignorant"	"last friday patricia arch mario arch daughter full time asian employee express distort view press generate not abide annex not speak behalf asian not mean shoot messenger cold not refrain express opposite opinion . two problem settle contract never attempt arbitrate solve bilateral contract dispute . lack clarity point sole frustrate attempt solve conflict last 6 month"	41.26

Table 3: A snippet of two of the centroids, a snippet of their closest emails from the dataset and their cosine similarity values (Alsmadi & Alhami, 2015).

Each email was then provided a label according to the cluster it was most similar to. A snippet of the resulting file is shown in figure 3 in section 5.5. The main premise of the K-means algorithm is to calculate the cosine similarity between the centroids and fabricated emails (Alsmadi & Alhami, 2015). Table 3 shows two of the centroids, their most similar emails in the dataset, as well as the cosine similarity value.

5.5 Data Discovery: Labelled Email Body Dataset CSV File

A snippet of the dataset that is labelled using the K-means clustering classifier, is shown in Figure 3.

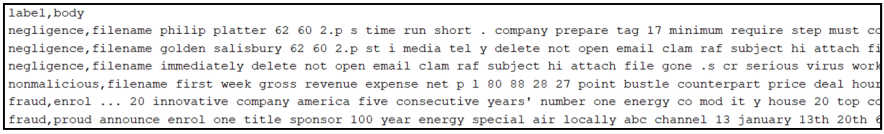

Figure 3: Snippet of dataset labelled with the K-means Clustering Algorithm

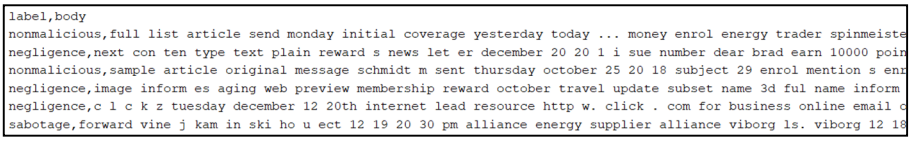

Figure 4: Snippet of dataset labelled with the Regular Expression Pattern Matching Algorithm

5.6 Data Discovery: Supervised Machine Learning Algorithms run on Both Labelled Datasets

The normalized and labelled email body dataset CSV file, that was labelled using the Regular Expression Pattern Matching Classifier, and the dataset labelled with the K-means classifier, was then used as training data into the machine learning algorithms. The supervised machine learning classifiers, SVM Naïve Bayes and Decision Tree were used in this experiment (Mayhew, et al., 2015). Each tokenized, normalized email body was vectorized, to convert the text into a number vector, to be used as input for the machine learning algorithms.

5.7 Data Discovery: Obtain Performance Metrics

The machine learning algorithms were executed for 30 iterations, and their results are summarized in the table below. It is clear that the results are closely related in terms of accuracy and performance.

	CSV 1: Regular Expression Pattern Matching labelled file		**CSV 2: K-Means Cluster labelled file**	
Algorithm	**Accuracy**	**Precision**	**Accuracy**	**Precision**
SVM	0.68	0.77	0.45	0.56
Decision Tree	0.92	0.93	0.73	0.75
Multinomial Naïve Bayes	0.79	0.81	0.59	0.63
Logistic Regression Naïve Bayes	0.91	0.90	0.83	0.83

Table 4: Results comparison for supervised learning classifiers with both labelled datasets

6 Discussion of Results

Through running the experimentation in section 5 to detect insider threats, the following has been noted. The Regular Expression Pattern Matching classifier developed by the authors performed well. This is due to the fact each sentence in the email was checked against the word list and scored, and the dictionary was updated for each sentence (Brown, et al., 2013). Therefore, each occurrence of a word or phrase, from the word list in a given sentence, ensured that the count was updated. This allowed for better accuracy. Scores were also only applied when the value of the score was above the given threshold of 0.5, which ensured that there was more confidence that a label was being correctly assigned. This is done to minimize false positives. Due to the fact that it is a textual and qualitative labelling process, metrics such as false positives and accuracy cannot be checked. To check the results of this approach, the K-means algorithm also labelled the dataset by means of classifying emails into clusters. The labels assigned to the emails within the two labelled datasets, the dataset labelled by the K-means algorithm, and the dataset labelled by the Regular Expression Pattern Matching classifier were compared. There were only 3149 emails out of 10001 emails (31.49%) that were assigned the same label from both approaches. Both of the classifier approaches need refining such that a higher percentage can be yielded for this comparison.

For the K-means algorithm, the centroids had to be manually fabricated in order to include as many phrases from the word list. The similarity values showed that the algorithm did well in arranging similar emails in the clusters based on the fabricated centroids. This was limited to research conducted on behaviours of possible employees within the various categories. The labelled dataset from the K-means clustering process was used to train the machine learning algorithms. Therefore, if this labelled dataset is flawed and results are unreliable, the machine learning will incorrectly identify and classify the types of insider threats.

To ensure that better results are obtained, the word lists and centroids will need a lot of work. More phrases that originate from the Enron or any other similar dataset need to be included. Email datasets containing insider threats are not readily available and, as such, training data for word lists and centroids had to be manually created based on research of behaviours. Datasets containing emails that hold insider threats would need to be fabricated by an automated process to ensure that a more extensive dataset is created. The labelled dataset was manually inspected to check whether the labels assigned to the emails were correct. Further research would need to explore other approaches of inspections.

The Enron email dataset itself required a lot of normalizing and cleaning in order to obtain reliable results. This process can also be refined. For example, the spelling correction method could be refined in how the correct word is located, and the spellcheck dictionary could be extended to include languages other than English.

This process of labelling content in isolation could be made more effective if a combination of other attributes of the email data are included in the prototype, such that a confidence score can be computed for each user based on weighted email attributes. Furthermore, it is not feasible to manually check each label assigned to each email in a given dataset and therefore, there is room for error. Thus, features that include quantifiable attributes will be easier to measure and should be added to the current prototype. Header information would specifically add a rich layer of attributes that can enhance this prototype.

In summary the following can be gleaned based on the results from the research at hand:

- The main types of insider threats and associated human behaviours, malicious and non-malicious, were established based on past research, to identify certain phrases that would be found in corporate emails. These types were then mapped to these identified human behaviours. It was noted that this was a crucial step in obtaining quality classification results. To enhance this process, the wordlists would need to be greatly expanded to include a wider range of scenario-based phrases. This would ensure that a wider range of emails could be covered and this would improve accuracy. Smarter means of constructing these lists would need to be explored in future work.

- An insider threat classification prototype was created based on the phrases identified, using machine learning techniques. This included acquiring a large email corpus, applying data normalizing techniques, and running machine learning algorithms. It was noted that the normalization was a critical step, because a cleaner dataset will be labelled more accurately and faster than one that has bad data. The supervised machine learning algorithms performed very well for big data, but their results depended on the quality of the normalized training data. For future work, there is a need for a readily available large, labelled email dataset containing insider threats. This would be used as training data for the machine learning classifiers. This complete and correct labelled dataset would ensure that testing datasets could be

labelled accurately. Another enhancement for future work, would be adding other email related attributes, such as email header details, to the classification and scoring processes.

7 Conclusions and Future Work

A novel approach is presented in this paper that contributes to minimising the risk of insider threats via email systems. This is done by constructing malicious and non-malicious human behaviour categories that insiders possess. Phrases that would be used by those insiders and that appear in their email communications are identified for each category of behaviour. Machine learning algorithms are used to identify commonly occurring insider threats and to group the occurrences into insider threat classifications. It was found that the decision tree algorithm yielded the highest accuracy and precision, 0.92 and 0.93 respectively. A highlight of the research at hand is the construction of a prototype that shows how a tool can be developed, that assists in the automated detection of insider threats in email systems. The architecture of the proposed prototype includes text processing tasks such as tokenization, stemming, vectorization, classification and clustering. These tasks provide a way to label emails according to the types of insider threats within large email datasets.

In future work, the word lists and centroids which are used in the proposed process of labelling the dataset, will be refined and methods to automatically generate big data files of similar phrases, such that a larger set of possible cases of, for the types of insider threats can be identified and be used for the labelling process.

8 References

Ali, Z., 2018. *Insider Threats – 2018 Statistics.* [Online] Available at: https://www.uscybersecurity.net/insider-threats-2018-statistics/[Accessed 14 March 2019].

Alsmadi, I. & Alhami, I., 2015. Clustering and classification of email contents. *Journal of King Saud University - Computer and Information Sciences,* pp. 46-57.

Aski, A. S. & Sourati, N. K., 2016. Proposed efficient algorithm to filter spam using machine learning techniques. *Pacific Science Review A: Natural Science and Engineering 18,* pp. 145-149.

Brown, C. R., Watkins, A. & Greitzer, F. L., 2013. Predicting insider threat Risks through Linguistic Analysis of Electronic Communication. *46th Hawaii Int. Conf. Syst. Sci,* pp. 1849-1858.

Butkovic, A., Mrdovic, S. & Mujacic, S., 2013. IP geolocation suspicious email messages. *21st Telecommunications forum TELFOR 2013,* pp. 881 - 884.

Cappelli, D., Moore, A. & Trzeciak, R., 2012. *The CERT Guide to Insider Threats.* Westford, Massachussetts: Pearson Education Inc.

Chi, H., Prodanoff, Z. G., Scarllet, C. & Hubbard, D., 2016. Determining Predisposition to Insider Threat Activities by using Text Analysis. *Future Technologies Conference,* pp. 985-990.

Claycomb, W. R. et al., 2013. Identifying Indicators of Insider Threats: Insider IT Sabotage. *IEEE.*

Cluley, G., 2016. *Citibank IT guy deliberately wiped routers, shut down 90% of firm's networks across America.* [Online] Available at: www.tripwire.com/state-of-security/featured/citibank-it-guy-deliberately-wiped-routers-shut-down-90-of-firms-networks-across-america [Accessed 14 February 2019].

Cukierski, W., 2015. *The Enron Email Dataset.* [Online] Available at: https://www.kaggle.com/wcukierski/enron-email-dataset [Accessed 18 January 2018].

Hussain, R. & Qamar, U., 2014. *An Approach to Detect Spam Emails by Using Majority Voting.* s.l., s.n.

IBM, 2019. *IBM.* [Online] Available at: https://www.ibm.com/za-en/ [Accessed 14 March 2019].

Kowalski, E., Cappelli, D. & Moore, A., 2008. US Secret Service and CERT/SEI Insider Threat Study: Illicit Cyber Activity in the Information Technology and Telecommunications Sector. *US Secret Service and CERT Program Software Engineering Institute.*

Leber, J., 2018. *The Immortal Life of the Enron Emails.* [Online] Available at: https://www.technologyreview.com/s/515801/the-immortal-life-of-the-enron-e-mails/ [Accessed 7 February 2018].

Lepinsky, R., 2013. *Analyzing Keywords in Enron's Email.* [Online] Available at: Rodger's Notes: https://rodgersnotes.wordpress.com/2013/11/24/analyzing-keywords-in-enrons-email/ [Accessed 18 January 2019].

Mayhew, M., Atighetchi, M., Adler, A. & Greenstadt, R., 2015. Use of Machine Learning in Big Data Analytics for Insider Threat Detection. *Milcom 2015 Track 3 - Cyber Security and Trusted Computing,* pp. 915-922.

Mujtaba, G. et al., 2017. Email Classification Research Trends:. *IEEE Access,* pp. 9044-9064.

Munshi, A., Dell, P. & Armstrong, H., 2012. Insider Threat Behavior Factors: A comparison of therory with reported incidents. *2012 45th Hawaii International Conference on System Sciences,* pp. 2401-2411.

Nizamani, S., Memon, N., Glasdam, M. & Nguyen, D. D., 2014. Detection of fraudulent emails by employing advanceed feature abundance. *Egyptian Informatics Journal,* Volume 15, pp. 169-174.

NLTK, 2019. *NLTK 3.4 documentation.* [Online] Available at: https://www.nltk.org/[Accessed 13 March 2019].

Okolica, J., Peterson, G. & Mills, R., 2007. Using Author Topic to detect insider threats from email traffic. *Digital Investigation 4,* pp. 158-164.

Pandas, 2019. *Python Data Analysis Library.* [Online] Available at: https://pandas.pydata.org/ [Accessed 12 March 2019].

Python, 2019. *Python.* [Online] Available at: https://www.python.org/ [Accessed 13 March 2019].

Rosen-Zvi, M., Griffiths, T., Steyvers, M. & Smyth, P., 2004. *The author-topic model for authors and documents. In: Proceedings of the 20th conference on uncertainty in artificial intelligence.* s.l., s.n.

Sashikanth, D., 2015. *Analysis of communication patterns with scammers in Enron corpus.* [Online] Available at: https://arxiv.org/abs/1509.00705 [Accessed 18 January 2018].

Spooner, D., Silowash, G., Costa, D. & Albrethsen, M., 2018. Navigating the Insider Threat Tool Landscape: Low Cost Technical Solutions to Jump Start an Insider Threat Program. *2018 IEEE Symposium on Security and Privacy Workshops,* pp. 247-257.

Tribolet, M., 2016. *Investigating Enron\u2019s email corpus: The trail of Tim Belden.* [Online] Available at: https://linkurio.us/blog/investigating-the-enron-email-dataset/ [Accessed 18 January 2018].

Van der Walt, E. & Eloff, J., 2018. Are Attributes on Social Media Platforms Usable for Assisting in the Automatic Detection of Identity Deception?. In: A. University, ed. *HAISA 2018.* Dundee: s.n., pp. 56-66.

White, J. & Panda, B., 2009. Implementing PII Honeytokens to Mitigate Against the Threat of Malicous Insiders. *IEEE,* p. 233.

Whitman, E., 2016. *Goldman Sachs Employee Email Surveillance: Which Terms Trigger Review Amid Concerns Over Losses And Insider Trading?.* [Online] Available at: https://www.ibtimes.com/goldman-sachs-employee-email-surveillance-which-terms-trigger-review-amid-concerns-2383065 [Accessed 16 February 2018].

Young, W. T., Memory, A., Goldberg, H. G. & Senator, T. E., 2014. Detecting Unknown Insider Threat Scenarios. *2014 IEEE Security and Privacy Workshops,* pp. 277-288.