CS/STAT 5525: Final Project Report

Talha Ghaffar, Hamza Manzoor, Adithya Upadhya, Gabrielle Smith, Sihan Yu

December 9, 2017

1 Analysis Overview and Data Handling

Electronic security is an essential component to keeping a company running. Electronic security not only protects a company and its interests, but employees as well. Weak electronic security can leave a company vulnerable to company espionage, theft of sensitive employee information, and even monetary theft. In this report, we will discuss our discoveries in the given dataset that may indicate a series of **spam/suspicious** emails that may compromise company information, **suspicious activities of disgruntled employees**, **suspicious activity of employees after leaving the company**, a model that would **predict the chances of an employee leaving the company**.

Computationally intensive tasks such as **LDA topic modeling of emails** were carried out in the **ARC cluster**. URL and browsing history analysis was challenging due to the presence of 28 million records or 15 Gigabytes of data. To address these, the records were processed in **chunks of 1 million**. Scripts suffered high execution time owing to sequential execution and processing. We harnessed the power of **Parallel Computing** to improve the execution speed. Using Python's **multiprocessing module**, we spawned a pool of **4 processes** (on a quad core processor) which processed the records in Parallel and thus, the execution time of scripts were reduced significantly. The technologies utilized for Information retrieval and analysis include Python and Pandas and Data visualization libraries include Seaborn and Matplotlib.

Major analysis in this project include **Topic Modeling (content classifier)** and **Sentiment Analysis** of URLs and Emails. **Google Cloud Natural Language API** (online REST API) was used for Content Classification and Sentiment analysis of **6033 unique URLs** extracted from dataset. Email Topic modeling was performed using **LDA** which was run on **ARC cluster** owing to computational requirements. Email Sentiment analysis was performed using **Vader Sentiment** analysis library. Unfortunately, Google APIs imposed a **rate limit of 800,000 hits/day** and therefore, we were unable to utilize Google APIs for analysis of **2.6 million emails** and thus, Vader and LDA were used instead.

2 Spam/Malicious Emails in the System

We used Latent Dirichlet Analysis (LDA) to model topics in the email content. For LDA, determining the **optimal number of topics** is challenging. Initially we chose small sections of the email dataset (100,000 emails) and applied LDA with different number of topics. For each model based on different number of topics, we **measured coherence of the topics** and chose the model with most coherent topics. The visualization in Figure 1 shows a projection of identified topics from one such analysis.

By visualizing the emails clusters, we noticed that one cluster stood out that apparently looked like spam emails. Further analysis showed that employees have sent these emails to each other and outside the company lack any logical content. And since these emails are being sent from accounts of most of the employees and are spread through the entire data, we **suspect them to be malicious/infectious emails**. Our analysis indicates that the **first** of these emails were sent on 2^{nd} **January, 2010** from the account of the employee **HSB0196**. The **second** employee who sent such emails is **IRM0931**. However, we did not find any emails passed between these two employees.

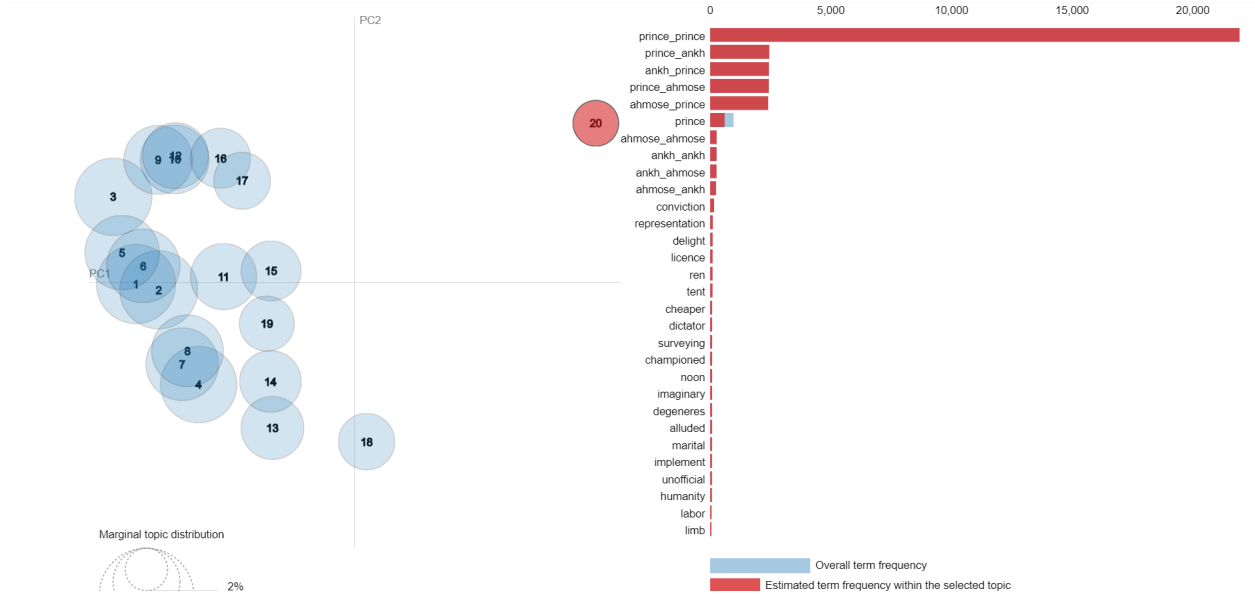

Figure 1: Topic visualization on a subset of emails. Topic 1 stands quite apart from others and these emails mostly comprises of three words **(prince, ahmose, ankh)** repeated multiple times. The content of these emails indicates them to be scam/suspicious emails. These emails comprise **2.3% of total emails**.

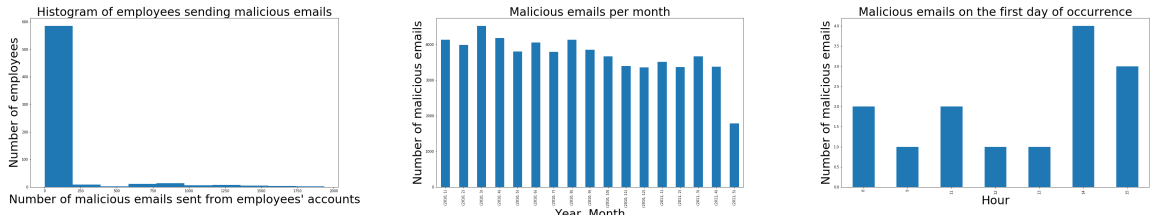

Figure 2: **(Left)** Histogram of employees sending suspicious emails. 647 employees have sent these emails at least once. 444 of these employees' accounts have sent these emails more than once. **(Middle)** suspicious/scam emails sent from employees' accounts in each month **(Right) Origin of malicious emails**. First email was sent from the account of HSB0196. Our analysis revealed that this employee was **infected after visiting a malicious URL** and one hour later, the same malicious content appeared in his PC to thumb-drive **file transfer info**

To find the origin of this content, we analyzed the websites visited by the employees and found that a total of **17 URLs contained this content** and 209 different employees visited those URLs. **The employees HSB0196 and IRM0931 were both infected by visiting such a URL**. Furthermore, we found **the presence of such Spam/suspicious content** in the **content being copied to their thumb-drives**. All these users visited such malicious URLs before copying such content to their thumb-drives which explains the source of infection. Furthermore, **such malicious emails were also sent to 2519 different people outside the company from DTAA employees' accounts (emails with dtaa domain name)**.

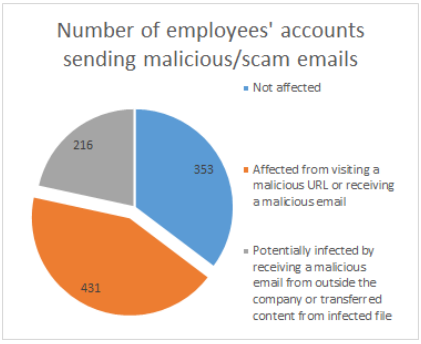

Figure 3

2.1 Conclusion

431 (of total 647) *different employees were affected either by visiting a malicious URL or by receiving a suspicious email for other DTAA employees.* To explain **how other 216 employees were infected**, we hypothesize two different scenarios: **1)** It's possible that some DTAA employees received such emails from **people outside the company's network**. since such emails were initially sent out to 2519 unique individuals outside the company **2)** It's also possible that the content was transferred from the **infected thumb-drives to the users' systems**. Since the data regarding file transfers from thumb-drives to PC are not available(only data regarding file transfer from PC to thumb-drives are available), we can only **speculate that their systems were infected by these infected thumb-drives**.

2.2 Recommendations/Action Items:

Based on the discovery of this malicious content in the system, we recommend DTAA to conduct a malware scan of these employees' PCs. Also, we suggest improving the security of the employees' machines by installing powerful **anti-virus software**.

3 Disgruntled Employees and their suspicious activity

Analysis of the content of the emails from employees to supervisors revealed **some employees** who were **discontent and disgruntled** with the company. These employees **felt unappreciated** and they communicated their unhappiness to their supervisors. We found **10 such employees, all IT Administrators, in the Electronic Security team**. All of these employees **left the company in the same month** in which they indicated job dissatisfaction. Furthermore, these employees **transferred key-loggers** (a program that copies and stores all the keystrokes made to a computer to their devices) **on their thumb-drives and that is their last recorded activity** which is **extremely suspicious**. These PCs with key-loggers installed have not been used by any other employee after these IT admins quit. Thus no employee has been affected by these Keyloggers.

9 of these employees have the **same supervisor 'Frances Alisa Wiggins'**. She supervised 20 different employees and **9 of them left the company with high job dissatisfaction**. This suggests **ineffective management** by Frances Alisa Wiggins.

3.1 Recommendations/Action Items:

The PCs used by these IT Admins should be cleaned and scanned for malware and keyloggers. The company should pursue legal action against these employees if they are found guilty. Also, the **company should review the performance of the supervisor Frances Alisa Wiggins** because under her management, several employees quit their jobs on the grounds of dissatisfaction with the job.

4 Potential Intellectual Property Theft

We found that of the **155 employees who left** the company, **2 of them logged on to the system after they had left the company**. Both of these employees also sent emails, using company email accounts, to their supervisors along with other people. Although there was no suspicious activity in the emails, one of these two employees sent an explanation email to his supervisor telling him that he had a rough year because of a divorce, through which we can say with somewhat confidence that this employee was fired. For the **other employee,** we found that he **transferred more than 20 files/documents onto his thumb-drive** after he had left the company. This employee was a **physicist** at the company in the **Research department**. This activity on his part is **extremely suspi-**

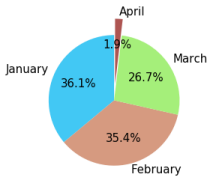

Figure 4: Percentage of files copied after leaving company. This employee logged on to the system in April and copied 25 files post his termination in March.

cious and suggests a **potential intellectual property (IP) theft**.

4.1 Recommendations/Action Items:

The company should investigate how these employees had access to their systems and email accounts even after they had left. Also, we recommend an investigation of the **files copied by the Physicist** which are identified above in order to concretely determine IP theft.

5 Predicting Employee Attrition

We attempted to create a model in order to predict the chances of an employee leaving the company. We analyzed data for different features that could predict an employee leaving the company. Here we describe the features we used and the rationale for them.

5.1 Feature Creation and Selection

Initially we created the following features from the dataset:

1. **Employees' Role/Team/Supervisors**: Employees with certain roles/supervisors have relatively higher attrition rates. One role/team we analyzed and discussed in section 2 is IT Administration where employees had high job dissatisfaction. In the current setup of the company, this feature indicates employee attrition. We also used employees' department, functional unit and supervisor as feature initially because the same criterion is reflective there as well. However, in our final model we only used employees' role because all these features are not linearly independent.

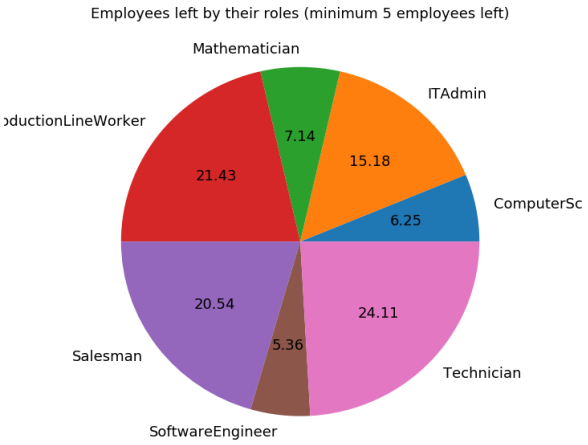

Figure 5: Employees left for each role (minimum 5 employees left). Some roles/teams have high attrition rates. We consider this as a feature that can contribute to predicting whether an employee will leave.

2. **Employees visiting job postings:** Applying topic modeling (Google content classifier) on visited URLs data, we observed that **521 employees visited URLs related to career and job postings**. We consider this to be reflective of employees' interest in finding a job and created a new feature number of **URL hits for job/career postings** by an employee.
3. **Employees sending out resumes:** Topic modeling of email data of employees who left the company revealed that some employees are applying for jobs via emails. We identified top keywords from the topics and identified all the employees who sent job related emails with attachments (people who apply for jobs send resume as attachment) and used this information as a feature.

4. **Employees' job search activities:** We created a feature that combined the information of job related URL visits and resumes/job applications sent out by employees to capture employees' job search activities.
5. **Sentiment of emails sent to supervisor**: Section 2 identified some employees that left because of job dissatisfaction and sent very negative emails to their supervisor before leaving. Therefore, we selected "sentiment of emails sent to supervisor" as a feature for our models.
6. **Positive and Negative URL hits**: Using sentiment analysis on URL, we created features with number of positive, negative and neutral URL hits for the employees.
7. **Mean** of sentiment of positive, negative and neutral URLs.

5.2 Model creation:

We initially used **Logistic Regression** as a baseline, but because of non linear relations in features, it didn't perform well. **SVM** with a radial basis function and polynomial kernels also didn't give better prediction results. Therefore, we used tree ensembles **Random forest** and **Gradient Boosting**. With proper tuning of both models, they both gave relatively good and almost similar predictions.

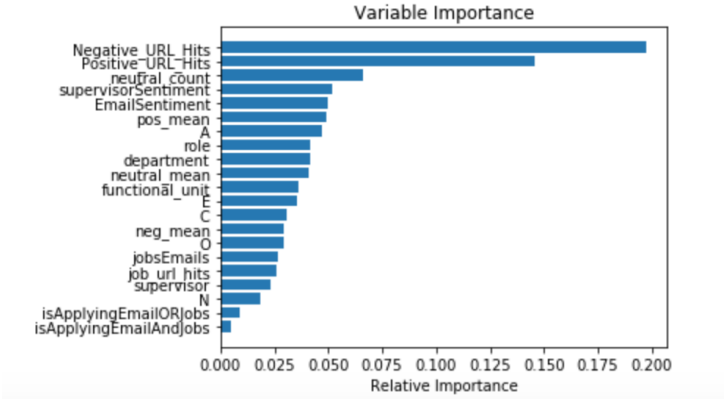

Figure 6: Feature Importance before tuning. Some of the features we mined from the dataset can be seen as relatively important for predicting employee attrition

We removed less important features during model tuning and report our results with the following features: Number of positive URL hits, Number of Negative URL hits, Is the employee applying for jobs, Number of URL visits to job postings, Role of an employee, Sentiment of Emails sent to supervisor.

5.3 Model evaluation:

With the features we selected from the data, we worked with following different models to try and predict whether an employee would leave the company.

Model	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**
Random Forest	95.6	98.93	76.12	95.76
SVM	83.90	99.17	0.6	84.47
Gradient Boosting	94.50	98.93	70.32	94.78
Logistic Regression	84.50	100	0	84.5

Table 1: Models experimented with in order to predict whether an employee would leave the company. Random Forest and Gradient Boosting with our selected features are able to give reasonably good predictions. However, SVMs and Logistic Regression were not able to predict well.