Analyzing Patterns of Fraud from the Enron Email Corpus

Prashita Prathapan
Indiana University, Bloomington
pprathap@iu.edu

Boryana Borisova
Indiana University, Bloomington
bborisov@iu.edu

Abstract

The Enron email corpus is rife of research opportunities in the fields of data mining and machine learning. It is a dataset that consists of 500,000 emails of real employees from the now bankrupt Enron Corporation, over a span of 3.5 years. This report employs methodology based on email classification, Naïve Bayes, Support Vector Machine, Random Forest classifiers, and Logistic Regression to determine whether it would be reliable enough to detect patterns of fraud through the contents of an email. Specifically our research attempts to answer the following question: Can the proposed methodology of analyzing email content of the top six executives be used for future attempts at revealing patterns of corporate fraud? To our knowledge, this analysis has not been done in a concentrated manner, that is, specifically focusing on these six individuals.

Keywords—email classification, Naïve Bayes, Support Vector Machine, Random Forest, Logistic Regression, corporate fraud, corporate corruption, Enron Corporation, cross-validation

1. INTRODUCTION

Before the name "Enron" became synonymous with large-scale corporate fraud and corruption, Enron Corporation was an American multi-billion dollar energy-trading company. At Enron's peak in 2000, its shares were worth $90.75; when the firm declared bankruptcy on December 2, 2001, they were trading at $0.26. Its leadership managed to fool regulators vis-à-vis institutionalized and systematic accounting fraud with fake holdings and off-the-books accounting. Since then, their data has been made public from about 150 users, mostly senior management of Enron, and contains a total of about 500,000 emails.

The dataset features a lot of information on the communication, associations, resources, among other things, between individuals and groups at Enron. In order to explore and understand how these factors interplayed to impact the destiny of this corporation, it is our challenge to extract, mine, and analyze this information in an effective way. Emails in research studies have often been understood through the lense of workplace collaboration [1] or task management [2]. This report revolves around fraud analysis and if the findings can be applied to future attempts at revealing patterns of corporate fraud. The paper focuses on targeting parsing through the email content of relevant employees in the scandal, as well as stock prices, and see if these relationships proved fraudulent activity. Specifically our research attempts to answer the following question: Can the proposed methodology of analyzing email content of the top six executives be used for future attempts at revealing patterns of corporate fraud?

Previous work has addressed whether top-level Enron employees had incriminating evidence in their work emails or if any unusual patterns could be uncovered in the months leading up to the scandal through exploratory data analysis [23]. To our knowledge, this analysis has not been done in a concentrated manner, that is, specifically focusing on these six individuals: - Sally Beck (Chief Operating Officer), Darren Farmer (Logistics Manager), Vincent Kaminski (Head of Quantitative Modeling Group), Louise Kitchen (President of EnronOnline), Michelle Lokay (Administrative Assistant) and Richard sanders (Assistant General Counsel). These six individuals were chosen because their directories were significantly large. A large directory would mean more data, which would mean, a higher accuracy rate. Our research is motivated by the scholars [23], who suggested that a more

concentrated approach be taken in analyzing this dataset to detect patterns of fraud through the contents of an email. This would ensure that only the emails of the most relevant people to the scandal are examined. Results revealed that the accuracy score for Support Vector Machine performed relatively better than the other classifiers, displaying an accuracy of 60.87%.

2. RELATED WORK

2. 1 Email Classification

Emails remain one of the major communication tools, regardless of the rise of mobile applications and social networks. As a consequence, the steady growth of email users has attracted massive amounts of unsolicited emails. Handling and classifying this vast amount of spam emails remains a prominent challenge. In 2009, it was estimated that more than 97% of emails were classified as spam [8], which is perhaps why a lot of researchers have focused on spam detection. Email classification fits in the broader field of text categorization, which automatically sorts a set of documents into categories from a predefined set, having multifarious applications. Common explorations within the literature of email classification are attributed to spam detection [3, 4] or automated foldering [5, 6, 7]. The documents can be classified by three methods: unsupervised, supervised and semi-supervised. One of the main learning methods are machine learning based techniques, such as Logistic Regression, Naïve Bayes (NB) [11], Random Forest [12], and Support Vector Machine (SVM) [10]. The other common technique for the email classification is a rule learning based method. Random Forest, which is one of the methods that we employed, is another common method in email classification. Scholars [12] have relied on it for the classification of phishing emails. They [12] successfully developed an improved phishing email classifier with a prediction accuracy of 99.7% and fewer numbers of features.

Aside from identifying spam, email classification can be used in the context of priority-based filtering or assigning messages to user-created folders. A novel research paper by [16] automatically classifies emails into activities, which can help develop a useful and accurate (more than 80%) activity management system. Since the most reliable of the four methods employed were SVM and Naïve Bayes, only these two methods will be closely reviewed in this section, with the hopes that researchers in this field can gain a better understanding of the existing solutions in the major areas of email classification.

2.1.1 Naïve Bayes for Email Classification

The advantages of using a Naïve Bayes algorithm on email classification are widely known, such as the short computational time, small amount of training data, and improved classification performance by removing irrelevant features. Aside from mentioning this method is easy-to-use and performs decently, [17] outline the disadvantages: it requires a large number of records to obtain good results, and the threshold value is needed when doing multiclass classification. In a review of machine learning algorithms for text classification, [18] mention Naïve Bayes for email and spam categorization. They posit that this method works well on numeric and textual data and implementing it is easy. However, they do mention that real-word data violates conditional independence, it performs very poorly when features are highly correlated, and it does not take into consideration word frequency.

2.1.2 Support Vector Machine for Email Classification

SVM is commonly used in supervised machine learning techniques because it provides relatively more accurate results than other techniques. SVM can produce better results with all the available features in the master feature vector because it is not prone to over-fitting [14, 21] and it is robust in noise. [13] provides an excellent review of email classification techniques, dataset analysis, features set

analysis, and performance measure analysis. They identified supervised machine learning as the most widely used email classification technique, with SVM as the most frequently used and best performance, followed by decision trees and the Naïve Bayes technique. [15] propose an email classification method based on Support Vector Machine and reach an 89.9% average accuracy.

2.2 Fraud Detection in Enron Email Corpus

A combination of corruption, fraudulent accounting, and poor regulation led the darlings of Wall Street, Enron Corporation, to file for bankruptcy in December 2001. It would be of merit to explore what other scholars have contributed to the literature of fraud detection as related to the Enron email corpus. Very few studies have looked at deceptive cues in this email corpus. The work of [9] is an exception. Instead of learning a predictive model for deceptive emails, they have examined structural features of emails (i.e. message length, word usage, word frequency) and ranked the emails by how likely they are to be deceptive [9]. They applied deception theory with singular value decomposition to the Enron email dataset to create a tool that can be applied to criminal investigations of organizations, internal auditing, and regulatory compliance. [9] did not, however, test whether the above mentioned linguistic cues predicted deception. Other authors [18] have explored whether there is a relationship between fraudulent activities and linguistic cues of deception within a large corporate social network. By applying a model of interpersonal language to the Enron email dataset, they revealed that during times of fraud, emails were composed with higher degrees of abstractness. Finally, social network analysis has also been levered to detect suspicious financial activities in emails. [19] demonstrate that the suspicious activities of a few individuals affect the entire network and thus suggest removing certain individuals

2.3 Email Classification on Enron Email Corpus

As mentioned earlier, the two main subfields of research within email classification lays in spam detection and foldering. One paper [21] reviews In one of the earlier papers following the release of the Enron email dataset, [20] used Maximum Entropy, Naive Bayes, SVM, and Winnow to establish that classification accuracy in many cases is low and hence recommend further research in email foldering. Other scholars [22] investigate the Enron email corpus against recipient recommendation systems by applying a multi-class multi-label classification reranking scheme. Their reranking scheme outperformed the baselines and offer a sound contribution to email recipient prediction.

3. DATA

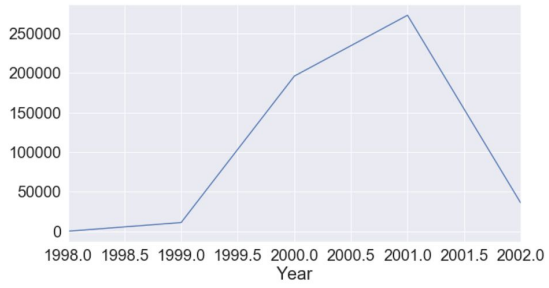
Shortly after the Enron Corporation met its demise amid revelations of accounting fraud in 2001, the Federal Energy Regulatory Commission obtained and released the emails of 150, mostly high-ranking employees, to the public. The cleaned-up version of the dataset that we used was published on May 7, 2015, which is published at https://www.cs.cmu.edu/~./enron/. It contains over 500,000 messages from 1997 to 2002. Duplicate emails were removed by removing duplicate folders, which was done and made available by [24] and [25]. This version removed any delivery-failure notices, redundancies, and automated messages from Listservs, among other rubbish. A sample email is pictured below:

```
Message-ID: <18782981.1075855378110.JavaMail.evans@thyme>
Date: Mon, 14 May 2001 16:39:00 -0700 (PDT)
From: phillip.allen@enron.com
To: tim.belden@enron.com
Subject:
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Phillip K Allen
X-To: Tim Belden <Tim Belden/Enron@EnronXGate>
X-cc:
X-bcc:
X-Folder: \Phillip_Allen_Jan2002_1\Allen, Phillip K.\'Sent Mail
X-Origin: Allen-P
X-FileName: pallen (Non-Privileged).pst

Here is our forecast
```

The corpus is particularly trailblazing because it is among the very few publicly available datasets of real emails, since typically this type of information is confined by numerous privacy and legal restrictions. This dataset has enabled and continues to allow for the study of interactions and processes within and among organizational entities. The main challenge with this corpus was trying to understand which emails are deceptive and which emails are not. Even though Enron Corporation has become synonymous with scandal, fraud, and deception, these attributes can't be uniquely ascribed to specific people or topics. Ths, it's best to identify deception by using time stamps with links to known fraudulent activities. This is why we specifically examined the January 2000 through December 2001. Our data includes labels that were manually provided by us and this is explored in further detail in the 'Methods' section.

Below you will find a graph that depicts the total number of emails annually from 1998 through 2002. We provide this information to show that the bulk of the dataset lays in the time span from 2000-2001, which is why we have chosen this period for our data analysis. These numbers correlate with the events that circulated Enron at the time. In September 2000, the media was increasingly reporting on the prevalence of market-to-market accounting in the energy industry, which was what Enron did. By February 2001, Chief Accounting Officer Rick Causey told budget managers, "From an accounting standpoint, this will be our easiest year ever. We've got 2001 in the bag," [26]. Soon enough, the company was faced with several serious concerns. By the end of August 2001, investors' confidence declined and the company's stock value continued to fall. On December 2, 2001, the company declared bankruptcy.



4. METHODS

4.1 Data Pre-Processing

Essential features of the dataset, namely the content of the email, sender email address, receiver email address, folder name, and the date, were extracted from each email using Python. The code can be found attached in the submitted zip file. In the next step, from the main corpus of emails, we filtered out the emails that were sent in the year 2000 and 2001. Further, we narrowed down our analysis to the emails sent by only the following individuals - Sally Beck (Chief Operating Officer), Darren Farmer (Logistics Manager), Vincent Kaminski (Head of Quantitative Modeling Group), Louise Kitchen (President of EnronOnline), Michelle Lokay (Administrative Assistant) and Richard sanders (Assistant General Counsel). These individuals had the largest directories, according to [23], in order to ensure that only the emails of the most relevant people to the scandal are examined. A random sample of 200 emails from the subset of top executive emails is used to carry out the email classification. We modeled this random sample number after the prominent work by [20], who took a similar sample size to train and then test the data.

For data analysis, we used Jupyter Notebook Python and the following libraries: pandas, numpy, sklearn, and seaborn. The first three libraries were used for modeling and the last library was used for the visualization of the data.

4. 2 Email Labeling

The content of the 200 randomly sampled emails was read and manually labeled into three classes. The categories can be described as follows :

1. Class 1 - Business-specific emails
2. Class 2 - In-person meetings emails
3. Class 3 - Others (personal, recruitment related, advertisements, holidays)

Business-specific emails contained information on company logistics, transactions, and other project-specific deals. In-person meetings represent the emails that talked about explicitly meeting to discuss work-related topics. Other information, such as casual speech and non-business related talk were categorized in the third class.

4.3 Training and Testing

We split the filtered emails into a 4:1 ratio of training to testing data. We then used the training data as input for the four classifiers mentioned below in 'Classifiers.' We used the cross-validation technique to evaluate the test set accuracy by using five folds of the training set. After that, we calculate the accuracy rates and created boxplots to compare all four classifiers to one another.

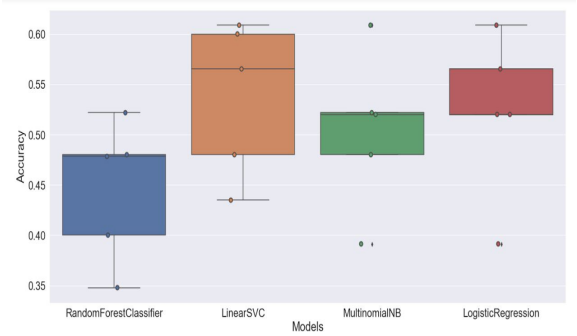
4.4 Classifiers

Initially, we had an unsupervised problem, as we had no labels for each email in the corpus. Hence, we manually labeled each email and converted this into a supervised learning problem. After this conversion from an unsupervised problem into a supervised one, we were then able to use different classifiers to carry out the email classification. The four classifiers that have been used were: Naïve Bayes, Support Vector Machine, Random Forest classifiers, and Logistic Regression.

5. EVALUATION

The evaluation metric that we used was the cross-validation accuracy for five folds. We plotted the accuracy score for each classifier and found that Support Vector Machine performed relatively better than the other classifiers. It had an accuracy of 60.87%. Naïve Bayes was taken as the baseline classifier and gave a maximum accuracy of 48% and it was compared with other classifiers. Random Forest performed the worst of the four methods which is why we discarded it. Logistic Regression performed similar to Naïve Bayes, but the latter is more widely used in text classification, hence why we used it as a baseline. In the boxplots below, you will see how each method performed related to one another.

Since this is a classification problem, the accuracy is simply the percentage of the correctly predicted labels over the total count of labels. We used the cross-validation method in order to use the entire training set, yielding higher accuracy rates.



6. DISCUSSION AND CONCLUSION

Machine learning involves two major phases: training and testing. The predictive accuracy of the classifier exclusively relies on the information gained during the training process. Therefore, if the information gained is low, the predictive accuracy is going to be low, but if the information gained is high, then the classifier's accuracy will also be high. In our research, the information gained was low, which led to a low accuracy rate. Since the accuracy

was so low, we were not able to predict the other emails and observe any trends related to corporate fraud. Recalling the results mentioned above, Support Vector Machine performed relatively better than the other methods, with a maximum accuracy of 60.87%. Naïve Bayes assumes that there are no conditional dependencies among the features and hence it is considered the baseline, or the lowest performing method. SVM, on the other hand is less over-fitting and robust to noise, making it a highly reliable linear classifier. However, we needed an accuracy of at least 70% to predict all the emails in the corpus and to observe the relationship between the classes like the dwindling business activities or suspiciously increasing in-person meetings.

Recalling our research question, (Can the proposed methodology of analyzing email content of the top six executives be used for future attempts at revealing patterns of corporate fraud?) we comment on our limitations. Our work had significant shortcomings due to human error in the labeling of the emails. It became evident that there was a lot of overlap in the three classes that we came up with, making the labels not as exclusive and as effective as we had anticipated. Perhaps we could have expanded the timeline of the data that we took, instead of 2000-2001, we could have taken 1998 through 2001. However, that analysis has to be normalized since the employees in the company over these years could have highly fluctuated and thus leading to erroneous results.

Future directions to improve this research could be done by taking multiple random samples of 200-400 emails from the top executives and using them to train the model, thereby improving the accuracy. A highly accurate model will allow for the prediction of labels of the rest of the emails in the corpus, without actually sifting through their contents manually. Out of these labels, trends of fraud would become more observable. Another limitation was that we were not pedagogically fluent in the energy sector, thus a future recommendation could be to have experts in the energy trading industry label the emails, in order to reduce human error. Also, with more computation power we can crunch the entire dataset to get more accurate results.

REFERENCES

1. Bellotti, Victoria, Nicolas Ducheneaut, Mark Howard, and Ian Smith. 2003. Taking Email to Task: The Design and Evaluation of a Task Management Centered Email Tool. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '03). Association for Computing Machinery, 345–352, New York, NY. DOI: https://doi.org/10.1145/642611.642672
2. Balakrishan, Aruna D., Tara Matthews, and Thomas P. Moran. 2010. Fitting an Activity-Centric System into an Ecology of Workplace Tools. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 787–790.
3. Sahami, Mehran, Susan Dumais, David Heckerman, and Eric Horvitz. 1998. A Bayesian Approach to Filtering Junk E-mail. In *Learning for Text Categorization: Papers from the 1998 Workshop*, Madison, Wisconsin. AAAI Technical Report WS-98-05.
4. Segal, R., J. Crawford, J. Kephart, and B. Leiba. 2004. SpamGuru: An Enterprise Anti-Spam Filtering System. In *Proceedings of the First Conference on Email and Anti-Spam*.
5. Segal, R. and J. Kephart. Incremental Learning in SwiftFile. 2000. In *ICML '00: Proceedings of the 17th International Conference on Machine Learning*, 863–870, San Francisco, CA.
6. Aery, Manu, and Sharma Chakravarthy. 2004. eMailSift: Mining-Based Approaches to Email Classification. In *SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Information Retrieval*, 580–581. ACM Press.
7. Kiritchenko, Svetlana, and Stan Matwin. Email Classification with Co-Training. In *CASCON '01: Proceedings of the 2001 Conference of the*

Centre for Advanced Studies on Collaborative Research, 192–201. IBM Press.

8. Alsmadi, Izzat, and Ikdam Alhami. 2015. Clustering and Classification of Email Contents. *Journal of King Saud University - Computer and Information Sciences*, 27(1): 46–57, https://doi.org/10.1016/j.jksuci.2014.03.014.
9. Keila, P.S., and D. B. Skillicorn. 2005. Detecting Unusual Email Communication. In Proceedings of the 2005 Conference of the Centre for Advanced Studies on Collaborative Research, 117–125. IBM Press, Toronto, Canada.
10. Xu, Ke, Cui Wen, Qiong Yuan, Xiangzhu He, and Jun Tie. 2014. A MapReduce based Parallel SVM for Email Classification. *Journal of Networks* 9(6): 1640–1647. https://core.ac.uk/download/pdf/25793084.pdf#page=284.
11. Rusland, Nurul Fitriah, Norfaradilla Wahid, Shahreen Kasim, and Hanayanti Hafit. 2017. Analysis of Naïve Bayes Algorithm for Email Spam Filtering Across Multiple Datasets. In *IOP Conference Series: Materials Science and Engineering*, 226(1): 012091. IOP Publishing. doi:10.1088/1757-899X/226/1/012091.
12. Akinyelu, Andronicus A., and Aderemi O. Adewumi. 2014. Classification of Phishing Email Using Random Forest Machine Learning Technique. *Journal of Applied Mathematics*, Article ID 425731. https://doi.org/10.1155/2014/425731.
13. Mujtaba, Ghulam, Liyana Shuib, Ram Gopal Raj, Nahdia Majeed, and Mohammed Ali Al-Garadi. 2017. Email Classification Research Trends: Review and Open Issues. In *IEEE Access* 5: 9044–9064. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7921698.
14. Oveis-Gharan, M. A., and K. Raahemifar. 2014. 'Multiple Classifications for Detecting Spam Email by Novel Consultation Algorithm. In Proceedings of the IEEE 27th Canadan Conference on Electrical and Computer Engineering, New York, NY, 1–5.
15. Shi T. 2012. Research on the Application of E-Mail Classification Based on Support Vector Machine. In *Frontiers in Computer Education. Advances in Intelligent and Soft Computing,* 133, edited by S. Sambath, and E. Zhu. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-27552-4_129
16. Dredze, Mark, and Tessa Lau. 2006. Automatically Classifying Emails into Activities. In *Proceedings of the 11th International Conference on Intelligent User Interfaces*, 70–77. ACM Press, Sydney, Australia. http://doi.acm.org/10.1145/1111449.1111471.
17. Khan, Aurangzeb, Baharum Baharudin, Lam Hong Lee, and Khairullah Khan. 2010. A Review of Machine Learning Algorithms for text-Documents Classification. In *journal of Advances in information technology,* 1(1): 4–20. DOI: 10.4304/jait.1.1.4-20.
18. Louwerse, Max, King-Ip Lin, Amanda Drescher, and Gün Semin. 2010. Linguistic Cues Predict Fraudulent Events in a Corporate Social Network. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 32(32).
19. Tang L., G. Barbier, H. Liu, J. Zhang. 2010. A Social Network Analysis Approach to Detecting Suspicious Online Financial Activities. In *Advances in Social Computing,* edited by S.K. Chai, J. J. Salerno, and P. L. Mabry P.L. SBP Lecture Notes in Computer Science, vol 6007. Springer, Berlin, Heidelberg.
20. Bekkerman, Ron, Andrew McCallum, and Gary Huang. 2004. Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora. In *Computer Science Department Faculty Publication Series*, 218. https://scholarworks.umass.edu/cs_faculty_pubs/218.
21. Metzger, J., M. Schillo, and K. Fischer. 2003. A Multiagent-Based Peer-to-Peer Network in Java for Distributed Spam Filtering. In the *3rd CEEMAS*, Czech Republic.
22. Carvalho, Vitor R., and William W. Cohen. 2007. Recommending Recipients in the Enron Email Corpus. In *Machine Learning.* http://www.cs.cmu.edu/~vitor/papers/old_CCtechreport.pdf.

23. Palaniswamy, Harish Kumar. 2015. Exploratory Data Analysis of Enron Emails. https://www.stat.berkeley.edu/~aldous/Research/Ugrad/HarishKumarReport.pdf.
24. Klimt, B., and Y. Yang. 2004. Introducing the Enron Corpus. *Conference on Email and Anti-Spam*, Mountain View, CA.
25. Shetty, Jitesh, and Jafar Adibi. 2004. The Enron Email Dataset Database Schema and Brief Statistical Report. Technical report, Information Sciences Institute.
26. Gibney, Alex., Jason Kliot, Susan Motamed, Peter Coyote, Bethany McLean, and Peter Elkind. 2005. Enron: The Smartest Guys in the Room. Los Angeles, CA: Magnolia Home Entertainment.