Nolan MacDonald

Dr. Jeffrey Borowitz

INTA 6450

September 24, 2024

Course Project Part I Proposal: Enron

Introduction

Following the collapse of Enron Corporation in December 2001, an investigation was conducted by the Federal Energy Regulatory Commission (FERC). In result, the corpus was originally generated from Enron email servers and released by FERC in 2003.The Enron Corpus database was originally generated from Enron email servers by the Federal Regulatory Commission (FERC) with more than 1.6 million emails that was originally released on March 26, 2003 (Leber). The original release of the Enron Corpus was later re-released due to concerns of commercially sensitive information, personal information or information with security concerns. The dataset was eventually purchased by Leslie Kaelbling at MIT and along with SRI International, the data was cleaned up and a number of integrity issues corrected to make the dataset available. The Enron Corpus database is constructed of over 600,000 emails from 158 employees, mostly senior management. The Enron email corpus remains one of the most valuable email databases for research to date and will be utilized to determine wrongdoing as discussed in this project proposal.

Wrongdoing

Wrongdoing is considered to be an act that is legal or immoral. In the case of the Enron accounting scandal, a general definition of wrongdoing refers to the act of violation of law by committing fraud, which can violate both civil and criminal laws. For the purpose of identifying wrongdoing at Enron, wrongdoing is defined as knowingly exploiting a mandatory process to ensure a desired outcome for an individual or company. In this case of Enron, wrongdoing is identified as ensuring a beneficial outcome for the company itself including financial gain, and neglects the possibility of individual gain.

Email Expectations and Signs

Several red flags indicating wrongdoing at Enron occurred both at the individual level and company level. Considering for this case that wrongdoing is defined as ensuring a desired outcome for the company, it is expected that emails will show several wrongdoings that benefitted Enron financially in each situation. It is expected that analyzing the email dataset will show that the company resulted in financial gains from accounting fraud and market manipulation. Accounting fraud occurred at Enron with executives hinting at efforts to hide losses and inflate profits through loopholes in accounting. When analyzing the dataset, it is expected there will discussions on gains and losses, special purpose entities and shell companies to assist in the company's appearance of having financial health. In addition, Enron employees were discovered discussing strategies to manipulate energy markets by creating power shortages and inflating energy prices, especially in California. To capitalize on energy markets and maximizing energy prices for profit, there is an expectation that discussions will be related to energy laws, power shortage, energy prices, and California in general.

Strategy

Considering that manually reading all available emails is not ideal, one strategy that could be implemented to find emails is a structured approach using the programming language Python. By taking a structured approach, the task of detecting wrongdoing in the Enron Corpus is broken down into steps to address crucial aspects of the investigation.

One approach to identifying potential wrongdoing would be using a structured approach with Python that incorporates text mining, Natural Language Processing (NLP) and network analysis. Text mining is performed by extracting the key information from the raw data provided in the Enron Corpus, which includes considerations of searching keywords and phrases as well as anomaly detection. Text mining would include looking for flags that indicate fraudulent behavior for maximizing profits on the financial reports. NLP applies machine learning and linguistics to provide an understanding of the language that is presented in the dataset. Network analysis allows for analyzing relationships and communication patterns between the Enron employees by mapping the social and communication networks. Using this approach will assist with identifying suspicious behavior or communications to detect potential wrongdoing. A conclusion can be developed in this structured approach by incrementally gaining insight to allow for examining the content of the emails as well as the interactions between individuals. To detect wrongdoing, the fields of text mining, NLP and network analysis involve multiple stages of development which will be discussed further with methodology.

Utilizing the approach described above, red flags or suspicious emails can be identified that can assist in determining wrongdoing for the financial benefit of Enron. Utilizing Python with several packages will assist in managing the data and developing models related to text mining, NLP and network analysis. Discussion of computational tools and packages will be discussed further in methodology. With several methods utilized in this investigation, it is expected that there will be several findings of potential wrongdoing. Only from a structured approach with multiple findings will a conclusion be clear. The goal of conducting a thorough analysis with multiple methods will allow for identifying emails from persons of interest and confirming whether wrongdoing occurred.

External Information

The primary external information necessary is the Enron Corpus. The Enron Corpus dataset that can be utilized has been collected and prepared by the Carnegie Mellon University CALO Project (A Cognitive Assistant that Learns and Organizes) and is available from the Enron emails online database (Cohen).

Methodology

As previously discussed, investigating wrongdoing with a structured approach can be beneficial to form a complete analysis of email content and interactions between individuals. Using Python with text mining, NLP and network analysis divided into several stages of development can connect insights to create a coherent investigation determining if the act of wrongdoing has occurred.

The framework of investigating wrongdoing will be divided into six steps that will allow for conducting a thorough analysis. The first step is data preprocessing, which will convert the email data from the Enron Corpus into a format that can be utilized for analysis. The second step will involve a keyword and phrase search, where emails can be identified that contain language or phrases that indicate potential wrongdoing. The third step involves NLP and four subtasks that use different methods to reveal patterns, themes and sentiments within the emails. The first NLP subtask is topic modeling, which will group emails into common topics or themes. Next, using NLP a sentiment analysis should be performed to evaluate the tone of the emails as this could indicate conflict or stressful situations that hint at wrongdoing. Additionally, using NLP, a named entity recognition (NER) analysis should be performed. NER extracts important people, companies and locations that are mentioned in the emails to track key entities that may be involved in suspicious activities. Finally, using NLP for text classification will train models to classify emails as suspicious or normal based on content, automating the process of identifying high-risk emails if patterns are established. After NLP, the fourth step is performing a network analysis. Network analysis will allow for understanding communication patterns between employees. This includes building the communication network where the communication between employees can be visualized by constructing a network graph. In addition, performing community detection for network analysis will help identify common communication between groups and abnormal communication that may indicate suspicion. The fifth step is anomaly detection, which will assist in detecting unusual behavior in email patterns and content. Using outlier detection will determine outlier emails based on frequency or content. The final step is to cross-reference the analysis with external documents such as financial records. This allows for external validation to correlate the emails and determine occurrences such as financial irregularities. In summary, the investigation of wrongdoing involves a step-by-step framework to conduct a thorough analysis; (1) Data preprocessing, (2) Keyword and phrase search, (3) NLP methods including topic modeling, sentiment analysis, named entity recognition, and text classification, (4) Network analysis considering communication networks, (5) Anomaly detection, and (6) Cross-referencing external documentation. The methodology for each step will be expanded upon further below.

Data preprocessing is the first step and the foundation of the analysis as it necessary to be able to perform all subsequent steps. Prior to preprocessing, the database consisting of Enron emails needs to be obtained from the Enron emails online database (Cohen). Data preprocessing involves converting the raw email data into a format that is usable for text mining, NLP and network analysis. Data preprocessing is beneficial considering the Enron Corpus consists of about half a million email files that all contain metadata such as sender, recipient, and subject, as well as the email body text. Preprocessing will extract the key information and clean the text for further analysis in the additional steps.

Preprocessing data considers a few actions that will need to be performed to parse the unnecessary information, standardize the data, and prepare the data for analysis. Preprocessing data should consider text extraction, normalizing data, tokenization, removing stop words and stemming. Text extraction removes email headers, footers and text that are not particularly useful. An example of text extraction would be removing email signatures or automated responses. Normalizing data is performed by converting all text to lowercase and removing the unnecessary characters to ensure text data is clean and consistent. This creates a standardized dataset, which is helpful for reducing noise and improving accuracy in text mining and NLP. For example, emails may contain the words “CEO”, “ceo” and “C.E.O.” which should all be treated analogously. Tokenization is beneficial for simplifying the complexity of the analysis by splitting the body of the email into words or tokens. The body of each email also needs stop words removal, where common words are removed. Email bodies typically have common words like “the” or “and” which do not have significance to the analysis and can be removed to reduce noise in the data. Finally, stemming should be performed, which preprocesses text by reducing words to the root form to be used in text mining and NLP so that words are standardized to focus on the core meaning. For example, if the text is referring to sending emails, there can be multiple forms such as “send”, “sending” and “sent” that can be reduced to a single representation of the word. In summary, data preprocessing is advantageous to prepare the data for text analysis and network analysis. For instance, tokenized and cleaned email bodies will be used for keyword searches in step two, sentiment analysis and topic modeling for NLP in step three, while sender and recipient data are essential for building the email communication network in step four.

With data preprocessing complete, the second step is using text mining to conduct a keyword and phrase search. Keyword and phrase searches search for key terms that are associated with wrongdoing such as fraud or insider trading. Using this text mining technique is helpful to narrow down the scope of the investigation by identifying a subset of suspicious emails. To develop keyword searching, a comprehensive list of terms will need to be generated that relate to wrongdoing and behavior that indicates knowingly exploiting a mandatory process to ensure a desired outcome for an individual or company. The list of keywords can involve many terms across different categories. In the case of Enron, these terms should include accounting and financial terms such as salary, stock options, bonuses, payments, fees, and income. As discussed, this is a comprehensive list, so terms should include anything related as well. For example, total stock value, exercised stock options, restricted stock, total payments, and deferral payments. For phrase matching, a list should be constructed beyond individual keywords that include phrases or context-specific combinations of words to perform an additional search of body text. In summary, using text mining to perform a keyword and phrase search will be useful not only to identify emails of concern, but also the context of how words are used, frequency, and who is involved in the discussions to determine potential wrongdoing.

The third step involves NLP and four subtasks to conduct a complete analysis. Using multiple techniques will expand analyzing the dataset beyond key words and phrases to gain insight on patterns, themes, and sentiments within the emails. For simplicity, these four subtasks will be referred to as 3A through 3D. Subtask 3A will use the NLP technique topic modeling, which helps group similar emails into common topics. Topic modeling should utilize the Latent Dirichlet Allocation (LDA) generative model, which will output all possible outcomes for a given phenomenon (Beysolow). This could reveal clusters of emails focused on market manipulation or offshore accounts that were part of the wrongdoing performed at Enron. To develop and perform this analysis, the Python scikit package can be utilized for LDA (Cournapeau). Subtask 3B should be developed to perform the NLP technique sentiment analysis. Sentiment analysis allows for evaluating the tone of emails that could indicate conflict or hint at wrongdoing. Using sentiment analysis could discover an email with negative sentiment when discussing financial situations to bring up a red flag for fraud. Using the Python package nltk, natural language processing with sentiment analysis can be performed (Bird). Subtask 3C involves using the NLP technique named entity recognition, or NER. This method will automatically extract important entities mentioned in email bodies and track key employees or companies involved in suspicious or fraudulent behavior. An example of NER would be recognizing several emails referring to shell-companies or entities that may indicate use of secret accounts for fraud. Subtask 3D is the final NLP subtask, which uses the technique text classification. Text classification models are trained to classify emails as suspicious or normal based on content and labeled data. Text classification is helpful to automate the process of identifying high-risk emails, which can be classified based on discussing financial data, legal strategy or other topics. To categorize an email as suspicious, thresholds will have to be defined and met.

Step four involves using network analysis to understand communication patterns between employees. Building an email network is helpful with network analysis to visualize communication between individuals by creating a graph with representation for each employee and email communication. When developing a network analysis measurements of network distribution will be important to identify to determine how nodes and edges are distributed in a network. This includes the important theory of betweenness centrality, which can show which nodes are likely pathways of information and what employees would act as bridges to facilitate communication for wrongdoing acts (Kadry). With use of the Python package, NetworkX, creation and analysis can be performed for complex networks (Swart). Email networks can help uncover key individuals, groups and relationships.

The fifth step is anomaly detection, helpful for detecting email communication patterns that may have abnormal behavior. One method of anomaly detection is using outlier detection, which determines if emails are outliers based on the frequency or content being not of the norm. This could be particularly useful in a case such as high volume of emails at a certain time from an individual, indicating a high stress situation or crisis. To perform anomaly detection, the scikit-learn package can be used with Python, considering an isolation forest or one-class support vector machines (SVM) (Pedregosa).

The sixth and final step is to use cross-referencing with external documents. Cross-referencing emails that are considered red flags or suspicious with financial records at the time can indicate actions of wrongdoing such as fraud occurred. One consideration would be to cross-reference when there are spikes in suspicious communications and compare with historical data published on Enron. In particular, financial statements during the time period of the suspicious emails discovered will need to be obtained.

Method Issues and Remedies

To reiterate the process investigating Enron emails, the determination of wrongdoing involves a step-by-step framework to conduct a thorough analysis; (1) Data preprocessing, (2) Keyword and phrase search, (3) NLP methods including topic modeling, sentiment analysis, named entity recognition, and text classification, (4) Network analysis considering communication networks, (5) Anomaly detection, and (6) Cross-referencing external documentation. Considering the several methods discussed, there are issues that need to be considered for each step and potential remedies or solutions to address each issue.

For data preprocessing, the task needs to be done with detail and precisely or it can lead to incomplete preprocessing. Incomplete preprocessing would result in the dataset containing noise that impacts the accuracy of the results if not parsed correctly. In addition, preprocessing applications such as stemming could be a caveat as they can strip away important meaning. Techniques in Python can be implemented in Python to prevent these issues, such as regular expression of cleaning rules to handle email information that is not useful.

Keyword and phrase searches have their benefits, but it requires a vast list of options to search for. If one of the keywords or phrases is not included, it could be missed in the analysis. In addition, searching for keywords does not determine hidden implications of wrongdoing or understand subtle language that should be flagged as a suspicious email. The resolution to these concerns is the use of the NLP method topic modeling, which should be developed in the third step.

The third step involves using NLP techniques, which have concerns over being vague in both sentiment and topic modeling. For example, emails in a professional environment typically have formal language that do not easily reveal underlying intent and make it difficult for sentiment models to determine. In addition, topic models can have difficulties capturing subtle topics and detailed topics. Sentiment modeling can be resolved with contextual sentiment models that are trained for business datasets to handle the subtleties. Considering topic modeling, the approach discussed was to use LDA models, which are useful to assist in topic detection refinement.

Using network analysis in the fourth step may struggle building with an incomplete email network, which is a possibility in this case considering the dataset does not include the Enron emails in their entirety. By focusing on sub-networks such as the key executives involved in the case, this will allow for noise reduction from an incomplete network.

The fifth step involves anomaly detection which can look at email patterns such as frequency. In this case, a high volume of emails may not necessarily indicate wrongdoing such as during a quarterly deadline. To resolve this issue, anomaly detection could be combined with content of emails such that certain keywords are detected during high volume email periods.

Finally, cross-referencing may involve issues of determining what needs to be obtained externally to check with emails that indicate suspicious behavior. In addition, there may be issues obtaining the external references that cannot be resolved. A resolution is to prioritize the references that need to be obtained depending on a combination of high-risk emails, employee behavior and email networks that are analyzed and determined to be of utmost importance.

In summary, there are concerns for each method and step that is involved in investigating wrongdoing with potential remedies to consider. The most difficult part of the six step process also is forming the extensive analysis as it involves several models to be developed and confirmed to be satisfactory to come to a determination that wrongdoing has occurred or not.

Conclusion

In conclusion, the definition of wrongdoing and relation to the Enron Corpus are discussed. An investigation of wrongdoing utilizing the public database containing Enron emails is proposed that involves text mining, natural language processing and network analysis. By utilizing multiple methods to conduct an analysis, a comprehensive understanding of potential wrongdoing can be formed to identify employees who exploited a mandatory process to ensure a beneficial outcome for the company Enron. The analysis is to be conducted by using a step-by-step framework as discussed that includes six key steps; (1) Data preprocessing, (2) Keyword and phrase search, (3) NLP methods including topic modeling, sentiment analysis, named entity recognition, and text classification, (4) Network analysis considering communication networks, (5) Anomaly detection, and (6) Cross-referencing external documentation.

Works Cited

Beysolow, Taweh. *Applied Natural Language Processing with Python: Implementing Machine Learning and Deep Learning Algorithms for Natural Language Processing.* 1st ed., 2018.

Bird, Steven; Loper, Edward; Klein, Ewan. *Natural Language Processing with Python.* O'Reilly Media Inc, 2009.

Cohen, William W. "Enron Email Dataset." May 7, 2015 ed., Carnegie Mellon University, 2015.

Cournapeau, David. "Latentdirichletallocation." scikit-learn https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html.

Kadry, Mohammed Zuhair Al-Taie; Seifedine. *Python for Graph and Network Analysis.* 1st ed., Springer, 2017. *Advanced Information and Knowledge Processing*.

Leber, Jessica. "The Immortal Life of the Enron E-Mails." MIT Technology Review https://www.technologyreview.com/2013/07/02/177506/the-immortal-life-of-the-enron-e-mails/.

Pedregosa, Fabian. "Scikit-Learn: Machine Learning in Python." *Journal of machine learning research*, 2011.

Swart, Aric A. Hagberg; Daniel A. Schult; Pieter J. "Exploring Network Structure, Dynamics, and Function Using Networkx." *Proceedings of the 7th Python in Science Conference (SciPy2008)*, 2008, pp. 11-15.