Assessment 2.2 - Analysis & Summary Report

Data Acquisition and Wrangling

Daniel Thompson,   
1103822831

Bachelor of Data Analytics

Contents

Executive Summary

This report presents a comprehensive analysis of email data pertaining to the Enron scandal, aiming to illuminate the internal workings of the company during its downfall. The Enron scandal, one of the largest corporate fraud cases in history, exposed significant shortcomings in corporate governance and ethics, leading to regulatory reforms aimed at improving transparency and accountability in financial reporting.

The analysis focuses on two key questions:

1. Were Enron employees employed in 2001 involved in email exchanges containing specific keywords related to fraudulent activities, and were these individuals among the most active in email communication during that year?

2. Did the average word count and number of emails sent or received by employees of Enron decrease from 2001 to 2002?

To address these questions, approximately 500,000 emails were analysed using various data wrangling, cleansing, and analysis techniques.

Key Findings:

- A significant proportion of Enron employees employed in 2001 were engaged in email exchanges containing keywords associated with fraudulent activities. Specifically, 59% of employees utilised three or more of these keywords in their email conversations. Further analysis revealed that this group constituted 86% of the top 50 employees in terms of keyword usage frequency.

- Despite some outliers, the average word count per email remained relatively consistent over the years, with exceptions noted for certain individuals who significantly increased their word count between 2001 and 2002.

- However, there was a notable decline in the number of emails sent by employees from 2001 to 2002. Median email counts decreased from 1100 in 2001 to 233 in 2002, indicating a substantial reduction in email communication.

Recommendations for Future Research:

1. Further exploration of outliers in email usage to understand their communication patterns and potential impact on organisational performance.

2. Investigation into the relationship between employees' communication patterns and organisational performance.

3. In-depth analysis of keyword usage to examine the context and broader content of emails containing these keywords.

Overall, this analysis provides valuable insights into the communication patterns and activities of Enron employees leading up to its collapse, contributing to a better understanding of the scandal's underlying dynamics and potential avenues for further research and investigation.

Introduction

The Enron scandal marked a significant failure in corporate governance and ethics, resulting in one of the largest bankruptcies in history. Enron's collapse in 2001 was fuelled by deceptive accounting practices orchestrated by top executives like CEO Jeffrey Skilling and Chairman Kenneth Lay. They employed complex financial schemes, notably off-balance-sheet partnerships like Special Purpose Entities (SPEs), to conceal debts and inflate profits, misleading investors and regulators (\*\*REF\*\*).

As an analyst for the Federal Energy Regulatory Commission (FERC), I was tasked with investigating the collapse of Enron. Supplied with approximately 500,000 emails, I have utilised many data wrangling, cleansing and analysis techniques to answer two questions. By answering these questions inside this report, I hope to gain an insight of the internal dynamics of the company during its collapse.

The two questions I aim to answer in this report are the following:

1. Were Enron employees that were employed in 2001 involved in email exchanges containing any of the listed keywords (Special Purpose Entities, Market-to-market, Off-balance, Fraud, Insider Trading), and were these individuals also among the most active in email communication during that year?

1. Did the average word count and number of emails sent or received by employees of Enron drop from 2001 to 2002?

The report structure encompasses several key sections. Initially, the Data Acquisition and Profiling segment delineates the methodologies employed for acquiring and pre-processing the dataset. Subsequently, the Data Analysis and Results section provides insights gleaned from the analysis, focusing on patterns and trends observed within the data. Following this, Reflections and Recommendations offer considerations on ethical practices and propose future research directions. Finally, the Conclusion succinctly summarizes the overarching findings and their implications.

Data Acquisition and Profiling

The main data source was provided by the University of South Australia (2024), although it was originally bought by Andrew McCallum for $10,000 USD from the FERC (Markoff, 2011). From examining the schema used throughout the emails, it was determined that plain text files of data would be acquired utilising command-line text processing utilities such as *grep*, *wc* and *sed* to filter through the dataset and extract data that met certain criteria.

Problems were encountered when it was noted that the data extracted was not structured that same throughout and that certain python methods (strip(), for example) would need to be depended on to see that correct snippets of data were being utilised.

Another problem of note was when the word count was extracted for each email. Although attempts were made to extract word counts from only the bodies of the email. When these counts were spot-checked, it was determined that the word counts were generally not within 10% of the real count. Therefore, it was determined to count the metadata in the word counts as this was commonly the same length within 2 words throughout the emails, or less than 10% of the word count.

Data Analysis and Results

Question 1

Were Enron employees that were employed in 2001 involved in email exchanges containing any of the listed keywords (Special Purpose Entities, Market-to-market, Off-balance, Fraud, Insider Trading), and were these individuals also among the most active in email communication during that year?

The first question’s text files extracted each keyword utilising *grep.* A text file was also extracted which contained a list of emails sent or received in 2001 which was used to correlate the list of employees who were working under Enron’s employ at this time.

From this, I was able to count the number of times each of the keywords listed above were used in each person’s email account. As seen below, 59% of employees utilised more than 3 of these words in their email conversations.

A pie chart with numbers and numbers

Description automatically generated

Further calculations found the following per employee:

Mean: 2.69

Mode: 3

Median: 3

From here, the question raised was ‘Was this a coincidence that all of these employees utilised the keywords on just one occurrence, or did the employees utilise these keywords on multiple occasions?’

A graph of a number of employees

Description automatically generated

From the above chart, marked in red are the employees who utilised at least 3 of the 5 keywords marked against their total occurrences of using the keywords. Out of these top 50 employees, 86% of them made use of at least 3 keywords.

A graph of numbers and a number of people

Description automatically generated with medium confidence

Here is the same chart with 4 or 5 occurrences. The number of employees drops to 44%.

Across all 149 employees, the following statistics were found for the number of uses of the keywords:

Mean: 1929.75

Median: 791

Mode: 1386

A graph of a number of keyword occurrence

Description automatically generated

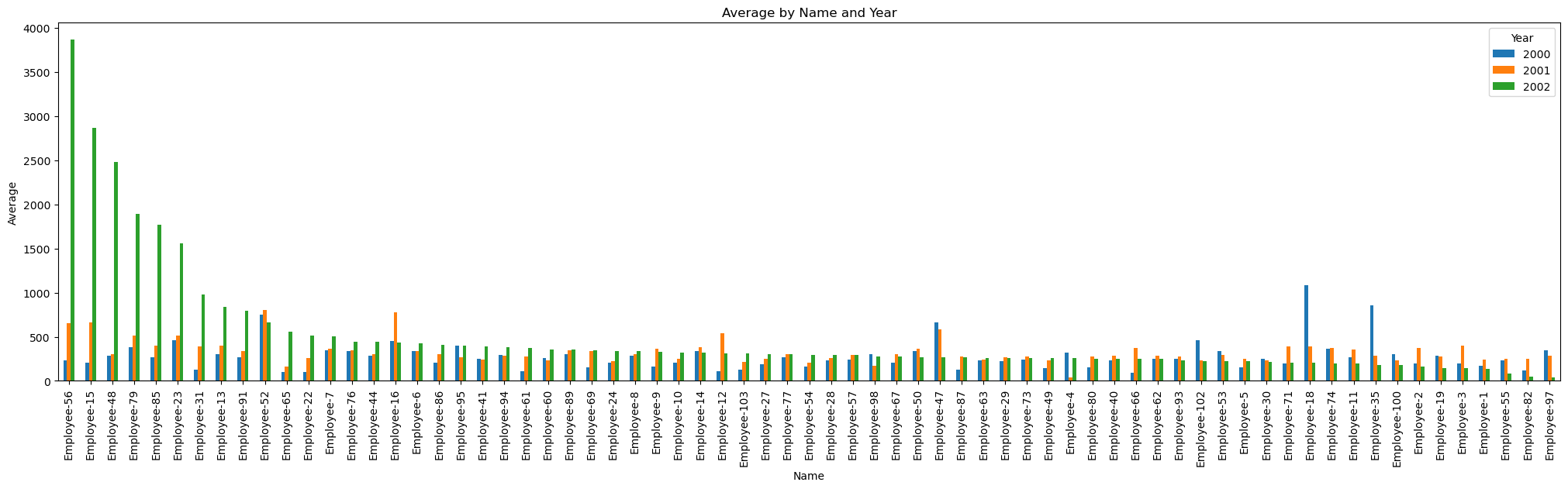
The above charts show that although there is a slightly positive skew, the presence of the mode being in between the mean and median suggests that there is a degree of clustering occurring at the lower end, possibly rendering the high-end entries as outliers when calculating statistics on the set as a whole.

In summary, while 59% of employees utilised 3 or more of the keywords, this group made up 86% of the top 50 employees when measured against how many times they used the keywords. Although the outliers heavily skew the statistics on use of these keywords, the over-use of these keywords has been noticed.

 Did the average word count and number of emails sent or received by employees of Enron drop from 2001 to 2002?

Data to answer this question was scraped utilising a combination of *grep* for the years, and *wc* for the word count for each email. In doing this, I chose to scrape for the year 2000 also, in case there was an anomaly not expected between the 2001 and 2002 data sets.

From taking the average word count of each employee, the following visualisation was created:



From this chart, you and see that apart from a select few, who drastically increased their word count between 2001 and 2002, most employees did not make any drastic alterations.

A graph with numbers and lines

Description automatically generated

The following chart presents the same statistics but visualizes them in a slightly different manner to provide a clearer indication of how a select few data points can significantly skew the statistics in a positive direction. This skewness is most evident in the difference between the values at the 75th and 100th percentiles. For instance, in the year 2001, the value at the 75th percentile is 376, whereas at the 100th percentile it jumps to 799. In contrast, for the year 2002, the corresponding values are 420 and 3863, respectively. This comparison underscores how a few extreme values can disproportionately influence the overall statistical summary.

The second part of this question asks the number of emails sent by employees during these years. Much like part 1, I also scraped 2000 for consistency.

A graph with blue and red lines

Description automatically generated

Highlighted in red, are the average emails sent per employee in 2002. This is compared to what was sent in 2000 and 2001 in blue. The difference is obvious, by understated.

A graph with numbers and lines

Description automatically generated

This chart highlights how great this change is. The median number of emails sent in 2001 by employees was 1100. This is compared to 2002 when only 233 were sent per employee. Averages were 1832 for 2000 and 348 for 2001.

Through these visualisations, it is determined that most employees did not alter their average word count, however they did drastically drop the number of emails sent from 2001 to 2002.

Reflections and Recommendations

As outlined in the proposal, analysts must make ethical considerations when analysing private use data. This includes creating safeguards to avoid identification of the people involved in the dataset. This has been accomplished in this report by replacing employees’ names, with a number to provide no tie to the person.

The overall process set out to answer the questions posed in the proposal was successful. It was established that most employees were using the keywords proposed, and the people that used at least 3 of those keywords were involved in the most emails with those words. Additionally, it has been observed that average word counts did not change through the years, with exception to a few outliers. However, employees dramatically decreased their email use in 2002 compared to previous years.

From this, future research suggestions are as follows:

1. Further examine the outliers in the email usage. This could be achieved by examining whether these individuals have a long history of large email usage and further investigation into their communication patterns.
2. While looking further into communication history, do the employees communication patterns play any influence on organisational performance?
3. Further exploration of keyword usage. Examining the context and broader content of these emails.

Conclusion

In conclusion, the analysis of email data pertaining to the Enron scandal has shed light on the internal dynamics of the company during its downfall. The findings reveal a significant involvement of employees in email exchanges containing keywords associated with fraudulent activities, underscoring the extent of deception within the organisation. While average word counts per email remained relatively stable, a substantial decrease in the number of emails sent from 2001 to 2002 suggests a notable shift in communication patterns, possibly indicative of the company's deteriorating situation. The identification of outliers in email usage and the exploration of keyword contexts present promising avenues for future research, offering opportunities to deepen our understanding of organisational behaviour and its implications for corporate governance and ethics.

* 1. References - List of sources cited throughout the report