Enron Emails Data-Set

Sami Ben Hassen
sami.benhassen@epfl.ch
237898

Firas Kanoun
firas.kanoun@epfl.ch
250235

Ali Fessi
ali.fessi@epfl.ch
247570

Abstract

In 2000, Enron(5) was one of the largest companies in the United States. Its stock kept soaring higher and higher day by day. By 2002, it had collapsed into bankruptcy due to widespread corporate fraud. In the resulting Federal investigation, a significant amount of typically confidential information entered into the public record, including around 500.000 emails and detailed financial data for top executives. Getting access to an email database without violating user privacy is close to impossible. The fall of Enron gave us the opportunity to dive into the detailed analysis of emails and how people use them. During this project, we'll try to investigate this data-set and try to answer the research questions we came up with below.

1 Credits

This document is the result of our project for the class "Applied Data Analysis" also known by its code CS-401 at EPFL under the supervision of the Teaching Assistant for the class Nuno Miguel Mota Gonçalves.

The Data-Set that we are working on was collected and prepared first by the CALO Project(2). Later, it was purchased by Leslie Kaelbling(9) at MIT. And now it is being distributed by CMU(3) as a resource for researchers who are interested in improving current email tools, or understanding how email is currently used.

2 Introduction

During this project, we'll try to investigate this data-set and try to answer the research questions we came up with. In section 3 we will explore our data: we will discuss the time at which people tend to send emails, people who send the most number of emails at this particular company and finally we'll take a look at the subjects of the emails and their content. In section 4 we will analyze the data-set as a social network. After discussing the above we will dive into more interesting research: we will create machine learning models that will classify emails as spam or ham. This a great opportunity to train such algorithms since email data-sets are very hard to come by (for privacy reasons). We will then end this paper by showing the results of some supervised and unsupervised models capable of finding corrupted individuals inside a company given that we cannot mention Enron without the words fraud or manipulation coming up.

3 Data Exploration

The Data-Set is represented as a list of folders. Each folder contains the Mail-Box of an employee. Therefore each folder contains a number of sub-folders representing for example inbox, outbox... We managed to compile everything in a nice pandas data frame which helped us in our data exploration.

3.1 Time is money

In this part of the project we explored different times people tend to send emails inside the company and showed how valuable time in e-mails can be in finding information about the company.

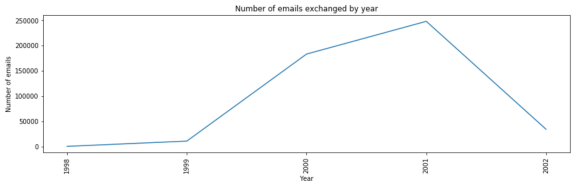

Figure 1: Number of emails exchanged by year.

From the yearly graph (Fig.1), we can see that the bankruptcy happened at the very beginning of the year of 2002 since the number of emails drastically decreased. If it had been towards the end we

would have had more emails exchanged that year. Checking with Wikipedia(6) we find that in "Dec. 2, 2001: Enron files for Chapter 11 bankruptcy protection".

A look at the number of emails sent by hour (Fig.2) and we get the exact schedule in the day of the employees. We can confidently say that they start their day at 8AM and finish at 6PM with a one hour break at noon. They are quite productive at 10AM and reach their highest productivity at 4PM before people start leaving. This doesn't apply to only ENRON employees though. Studies that have explored productivity in different countries around the world have shown similar productivity patterns(10).

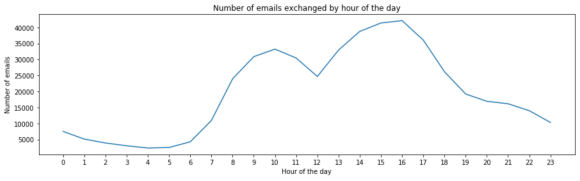

Figure 2: Number of emails exchanged by hour of the day.

3.2 Diving into the content

Taking a deeper look to examine the actual content of all these emails and removing some obvious stop words (Fig.3) we find out that the most used words within this company are usually part of the following categories:

- Meetings and time updates : report, date, hour, ahead, codesite, meeting, conference, start, request, call...
- The field the company is working in: gaz, power, energy...
- The years during which the company has been more active: 2000 and 2001.

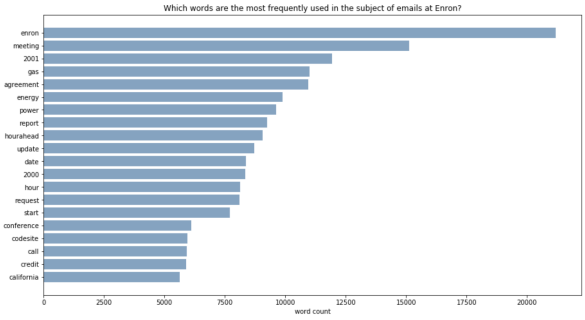

Figure 3: TOP Words used in the subject at ENRON

4 Network Analysis

For this part we will create a graph having as nodes as the name of employees and the connection between them as edges.

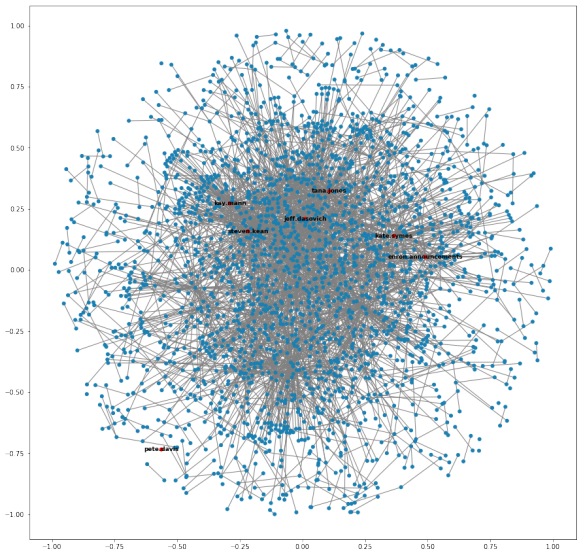

Figure 4: The Social Network at ENRON

From Fig.4 we can see that the most influential people are more or less in the center of this network, Nodes that are very far away from the center are likely to represent email addresses that are used as proxy like 'pete.davis' which was used for auto-generated emails(4). This is why we need to look at the centrality which will give us a more in depth look at the connections within the firm.

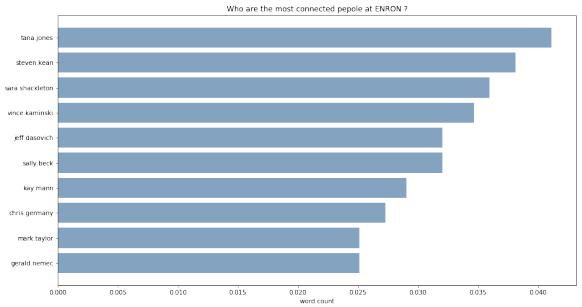

Figure 5: Most connected people at ENRON

Among the most central people we find : Steven Kean which was the Vice President and Chief of Staff and Vince Kaminski as the Risk Management Head (4) which explains the fact that they are one of the most connected employees.

5 SPAM Detection

5.1 Data Acquisition

To train a classifier for SPAM detection we had to find a source of SPAM emails we could use.

We have found "SpamAssassin"(11) and "HoneyPot" which are sources that provide these type of emails. Some of these emails have been mixed with our ENRON data-Set and were made available publicly(12). Again here the data is provided as text files inside folders and sub-folders so we did the same thing as in the data exploration part and we ended up with a balanced data frame that has 30.000 emails.

5.2 Train Machine Learning Models

After cleaning up the email contents, transforming the words into vectors and applying different machine learning models we get the results displayed in table 1:

	Accuracy
Linear SVM	**0.989948**
SGD Classifier	0.984681
Naive Bayesian Multinomial	0.986802
Logistic Regression	0.982977
Extra Trees	0.973425
Random Forest	0.973754
AdaBoost	0.972014
Extra Forest	0.944377
k-Nearest Neighbors	0.785367

Table 1: Different ML algorithms results.

We can clearly see that Linear SVM, with a great accuracy of **98.99%**, is the best model among the ones we have tested.

6 Fraud Detection

6.1 Building the financial data

This part of the project shows how hard it could be to fetch and compile clean usable data.

Since the scandal, people have actually managed to collect a lot of the company's and employees' financial information that was publicly available. There is a pdf of this information available on findlaw(7).

After scrapping these pdf pages and organizing them in a data frame, we tried to find matches between employees in both data-sets.

Below is a list of the interesting features we have added from the e-mails data set, the whole list is available in the notebook:

- The number of emails sent to POIs
- The number of emails received from POIs.
- The Stress Score, representing how many words denoting fear employees used in their emails.

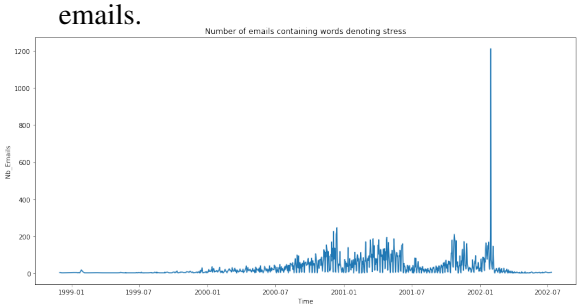

Figure 6: The stress score at ENRON

Interestingly, looking at Fig.6 we notice that the "Stress Score" is starting to rise on July 2000, while the stock has started to decrease on January 2001. It's amazing to see that the content of e-mails gave us such an important insight 6 months before the public started to suspect a strange behaviour from the management, which was the idea behind adding this feature to our models.

Finally, we used an article from usatoday(13) that provides a list of all the people involved in the scandal. We extracted the names of these employees that either reached a settlement with the government, pleaded guilty or testified in exchange of persecution immunity. We labeled them as POIs (People Of Interest) and used this label as the target of our models.

6.2 Supervised Learning Models

We have noticed that our data-set is heavily unbalanced, containing a ratio of only 0.14 of POIs which makes it quite hard for some machine learning models to achieve good results. After looking up different methods(1) to overcome this problem, below are the best that have worked for us:

6.2.1 Over-Sampling

Over-Sampling increases the number of instances in the minority class by replicating them in order to present a higher representation of the minority class in the sample. We therefore duplicated 3 times the rows of POIs in our training set. Table.2 shows the results.

6.2.2 Boosting-Based Approach and Decision Trees

Boosting is an ensemble technique to combine weak learners to create a strong learner that can make accurate predictions. Boosting starts out with a base classifier / weak classifier that is prepared on the training data. After each iteration, the

	Precision	Recall	F_Score
KNN	**0.82**	**0.82**	**0.82**
Multi.NB	0.500	0.277	0.357
SVM	0.500	0.277	0.357
Linear SVM	0.500	0.277	0.357
RBF SVM	0.500	0.277	0.357
L.Regression	0.500	0.277	0.357
SGD Class.	0.48	0.471	0.45

Table 2: Different ML algorithms results after Over-Sampling.

new classifier places more weight to those cases which were incorrectly classified in the last round. We also know that decision trees are not affected too much by unbalanced data so we thought it was a good idea to use them.

Table.3 shows the results of some decision trees and boosting ML algorithms.

Fig.7 shows the confusion matrix of the applied model on the test set.

	Precision	Recall	F_Score
Extra Forest	**0.833**	**0.952**	**0.875**
Extra Trees	**0.833**	**0.952**	**0.875**
Random Forest	0.5	0.38	0.431
AdaBoost	0.754	0.73	0.74

Table 3: Different Decision Trees and Boosting-Based ML algorithms results without oversampling

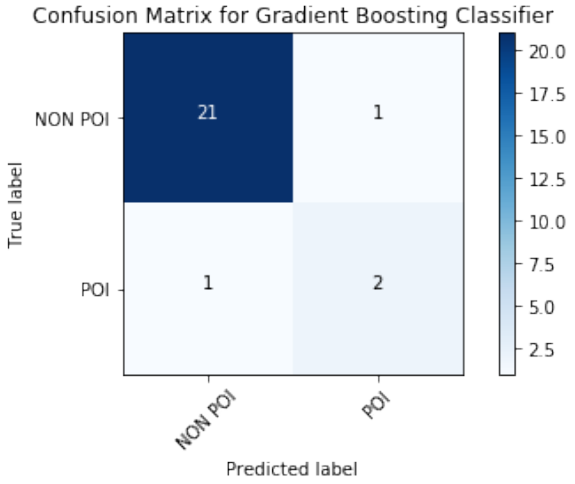

Figure 7: Confusion Matrix for Extra Forest Classifier.

6.3 Unsupervised Learning Models

For this part we decided not to use any prior knowledge about employees inside Enron. We thought that executives sharing comparable salaries, bonuses and e-mails content share somehow similar responsibilities. We therefore used K-MEANS (8) algorithm to perform this unsupervised task and try to classify people as corrupted or not. We started by building two clusters as we want to resolve a classification problem. We then ran our algorithm in two dimensions by selecting two features which gave us some encouraging results that can still be improved. We then selected the 16-most useful features and were positively surprised to see that without any knowledge of the scandal and despite the very low amount of data we had, our model was able to give astonishing prediction results. It was able to accuse of fraud the two main responsible of the scandal Lay Kenneth : founder, CEO and Chairman of Enron Corporation for most of its existence and Jeffrey Skilling : CEO of Enron Corporation during the case.

Table.4 shows the results

Accuracy	Precision	Recall	F_Score
0.875	0.885	0.984	0.932

Table 4: Results of the Unsupervised Learning Model

7 Conclusion

The Enron scandal gave us an opportunity to conduct studies on a real life email data-set that would have been hard to come by because of a lot of regulations and privacy reasons. From data exploration to SPAM/HAM classification and Fraud detection, we discussed what kind of analysis one can do on this type of data. We have had hands on experience on how to collect data from every source imaginable and how hard it could be to find clean data related to a specific subject. We experimented with a lot of machine learning algorithms and learned how to correctly choose a model that outperforms every other one. We never thought we could achieve such impressive results. As we presented in the first read me, our aim from the beginning was to build a tool for regulators to help them discover new scandals and shed the light on the main protagonists. We are deeply confident that in the future the work of legal departments and regulator institutions will need more and more expertise in Data Analysis in order to reach new levels in the war against financial criminals.

References

[1] How to handle unbalanced classification problems in machine learning? https://bit.ly/2CfJNB4.

[2] CALO Project cognitive assistant that learns and organizes. http://www.ai.sri.com/project/CALO. Accessed: 2018-12-13.

[3] CMU carnegie mellon university enron dataset. https://www.cs.cmu.edu/~./enron/. Accessed: 2018-12-13.

[4] Enron employees. http://www.inf.ed.ac.uk/teaching/courses/tts/assessed/roles.txt.

[5] Enron. https://en.wikipedia.org/wiki/Enron, 2018.

[6] Enron scandal. https://en.wikipedia.org/wiki/Enron_scandal, 2018.

[7] Findlaw : Legal information. https://www.findlaw.com/.

[8] k-means clustering. https://en.wikipedia.org/wiki/K-means_clustering.

[9] Leslie Pack Kaelbling panasonic professor of computer science and engineering. https://people.csail.mit.edu/lpk/. Accessed: 2018-12-13.

[10] Sales productivity around the world. https://www.pipedrive.com/en/blog/sales-productivity-around-the-world, 2018.

[11] Spam assassin. https://spamassassin.apache.org/.

[12] Spam ham enron emails. http://www2.aueb.gr/users/ion/data/enron-spam/.

[13] usatoday : A look at those involved in the enron scandal. http://usatoday30.usatoday.com/money/industries/energy/2005-12-28-enron-participants_x.htm.