University of California-Berkeley

Statistics Undergraduate Research Project

# Exploratory Data Analysis of Enron Emails

*Author:*
Harish Kumar Palaniswamy

*Supervisor:*
Prof. David Aldous

May 15, 2015

**Abstract**

Enron Corporation was an American energy, commodities, and services company based in Houston, Texas. Before its bankruptcy on December 2, 2001, Enron employed approximately 20,000 staff and was one of the world's major electricity, natural gas, communications, and pulp and paper companies, with claimed revenues of nearly $111 billion during 2000. At the end of 2001, it was revealed that its reported financial condition was sustained substantially by an institutionalized, systematic, and creatively planned accounting fraud, known since as the Enron scandal. Enron has since become a well-known example of wilful corporate fraud and corruption. This report aims at answering whether top level Enron employees had incriminating evidence in their office emails or uncover any unusual patterns in the months leading up to the scandal through an exploratory data analysis.

# 1 Introduction

This dataset was collected and prepared by the CALO Project (A Cognitive Assistant that Learns and Organizes). It contains data from about 150 users, mostly senior management of Enron, organized into folders. The corpus contains a total of about 0.5 million messages. This data was originally made public, and posted to the web, by the Federal Energy Regulatory Commission during its investigation. The dataset consists of 517,431 messages that belong to 150 users, mostly senior management of the Enron Corp. Although the dataset is huge, topical folders of particular users are often quite sparse. For our purposes, we only look at sent emails and ignore the inboxes of all the employees. Through this approach, we can avoid accidentally analysing the spam emails that are among the received emails. Two main methods of analyses were employed, namely, topic modelling with Latent Dirichlet Allocation(LDA) and sentiment analysis.

## 2 Data Processing

Before beginning any sort of data analysis on the emails, it is important to pre-process all text in the emails first. Below is an example of raw email text.

```
Message-ID: <5525962.1075855679785.JavaMail.evans@thyme>
Date: Wed, 13 Dec 2000 07:04:00 -0800 (PST)
From: phillip.allen@enron.com
To: christi.nicolay@enron.com, james.steffes@enron.com, jeff.dasovich@enron.com,
joe.hartsoe@enron.com, mary.hain@enron.com, pallen@enron.com,
pkaufma@enron.com, richard.sanders@enron.com,
richard.shapiro@enron.com, stephanie.miller@enron.com,
steven.kean@enron.com, susan.mara@enron.com,
rebecca.cantrell@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: Christi L Nicolay, James D Steffes, Jeff Dasovich, Joe Hartsoe, Mary Hain, p
X-cc:
X-bcc:
X-Folder: \Phillip_Allen_Dec2000\Notes Folders\Sent
X-Origin: Allen-P
X-FileName: pallen.nsf

Attached  are two files that illustrate the following:

As prices rose, supply increased and demand decreased.  Now prices are
```

`beginning to fall in response these market responses.`

As can be seen, the emails in their raw form contain a lot of information that is unsuitable for topic modelling analysis. Ideally, we only want the actual content of the email, which in the case of this email, is only the last 3 non-empty lines. It is easy to extract the relevant text of each email because of the structured nature of them. Regex expressions were utilized to filter out the header information such as the Date, Content-Type etc. Each email was assigned a unique ID(UID) and written to a new file with name "Month-Year-UID.txt". The Month-Year were included in the filename since it makes it easier to distribution of topics over time by adopting such a naming convention. For example, the email above would be written to "12-200-UID.txt" after being processed by a custom Python script

# 3 Topic Modelling with LDA

Latent Dirichlet Allocation was performed on the dataset with the number of topics, $k = 4$. The following is a list of the top 10 terms for each of the 4 topics. Note that the words have been stemmed to save memory.

| Topic 1 | Topic 2 | Topic 3 | Topic 4 |
|---|---|---|---|
| "message" | "enron" | "market" | "thank" |
| "origin" | "deal" | "gas" | "call" |
| "pleas" | "agreement" | "price" | "time" |
| "email" | "chang" | "power" | "meet" |
| "thank" | "contract" | "compani" | "look" |
| "attach" | "corp" | "energi" | "week" |
| "file" | "fax" | "trade" | "day" |
| "copi" | "houston" | "busi" | "dont" |
| "inform" | "date" | "servic" | "vinc" |
| "receiv" | "america" | "manag" | "talk" |

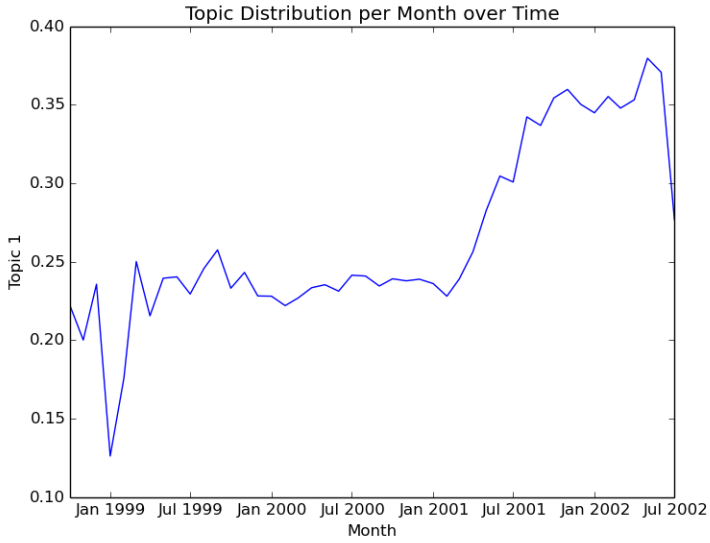
1. Topic 1 contains a lot of meeting related words, perhaps they are from emails that were sent as meeting notices.

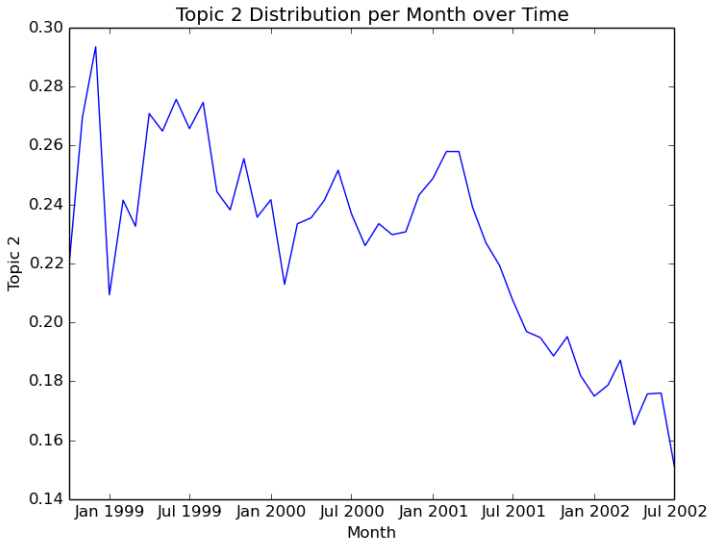
2. Topic 2 while related to business seems to be more about the process rather than the content of the core business. It has a lot of terms relevant to business legalities.

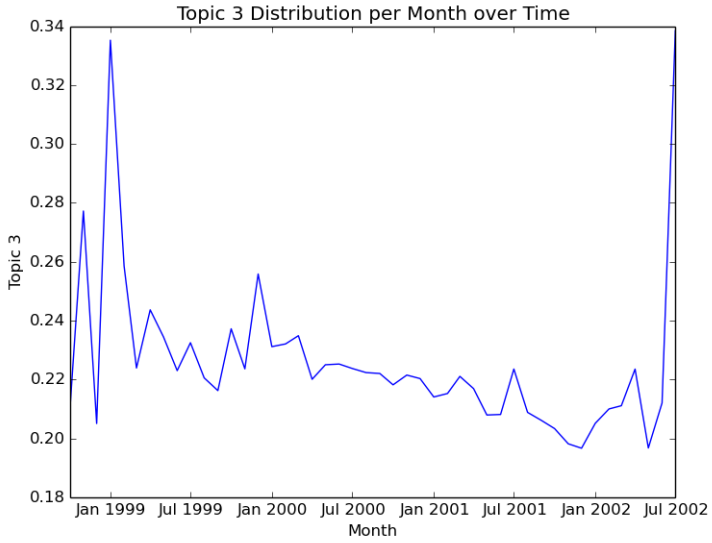
3. Topic 3 contains words that are directly related to the core business of Enron like "gas", "power" etc.

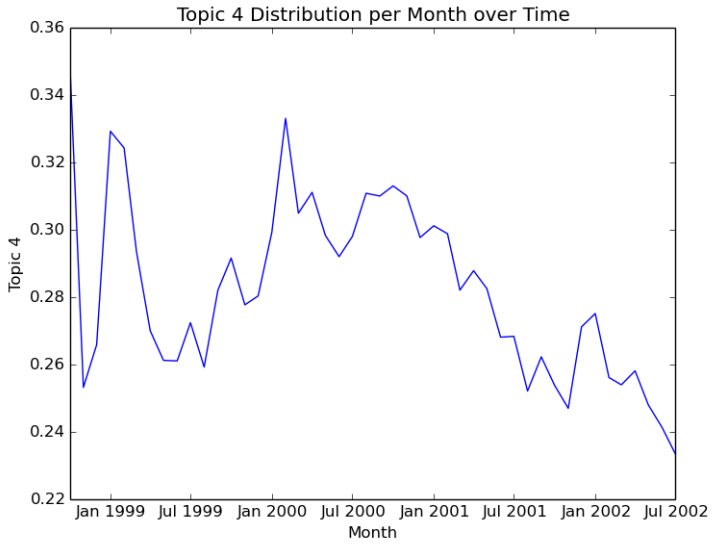
4. Topic 4 also seems to be meeting-related but in a more casual tone and setting.

Below are the plots of the distribution of each topic over time(Oct 1998 to September 2002) partitioned by month.


Topic Distribution per Month over Time
Topic 1
Month
0.40
0.35
0.30
0.25
0.20
0.15
0.10
Jan 1999
Jul 1999
Jan 2000
Jul 2000
Jan 2001
Jul 2001
Jan 2002
Jul 2002


Topic 2 Distribution per Month over Time
Topic 2
Month
0.30
0.28
0.26
0.24
0.22
0.20
0.18
0.16
0.14
Jan 1999
Jul 1999
Jan 2000
Jul 2000
Jan 2001
Jul 2001
Jan 2002
Jul 2002




From the above plots, Topic 1 has a huge dip in Jan 1999, but bounces back up quickly and stays more or less constant until Jan 2001, after which it keeps increasing. The dip in Jan 1999 might be due to edge cases where emails related to Topic 1 got pushed over to Dec 1998 and Feb 1999, which

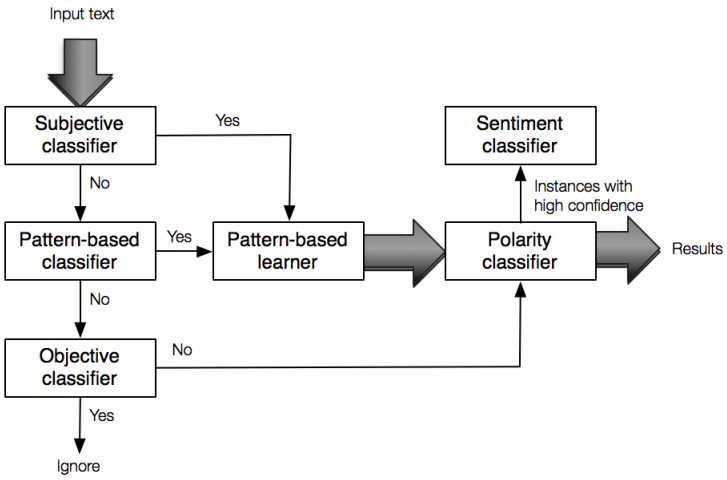
does seem to be the case by looking at the 2 spikes in the data at those times. The more interesting pattern is the steady increase in Topic 1 after Jan 2001, this is actually an interesting occurence. It is well known that Jeffrey Skilling and Kenneth Lay, then CEO and Chairman of Enron respectively at the time of the scandal, were holding regular meeting with their top executives in order to pressure them into finding new ways to hide Enron's debt. Since the scandal was only made known to the public in Oct 2001, it could well be the case that this abnormal rise in meetings was an indicator of Enron's executives trying to cover up their accounting fraud.

Topic 2 on the other hand has a general decrease throughout the years. Since Topic 2 contains words like "contract", "deal" and "agreement", this might well be an indicator of dwindling business activity throughout the years. The huge decrease after Jan 2001 makes sense given that it was only 8 months before the scandal.

Topic 3 which is related to core business terms is more or less constant throughout the time except for spikes at the start and end of the range of dates. Topic 4 which is an indicator of casual meetings seems to be on the decline in the days leading up to the scandal which makes sense. Although the topic model plots do reveal some interesting patterns over the time, it is still unclear as to whether the models reveal any incriminating evidence. However, investigators could look at the points in time where abnormal patterns started in order to look for email evidence.

# 4 Sentiment Analysis

The main challenge in conducting sentiment analysis on the emails was the complete absence of training data i.e. emails with positive or negative labels. Therefore, unsupervised sentiment learning was performed in accordance with the pipeline shown in the image below.

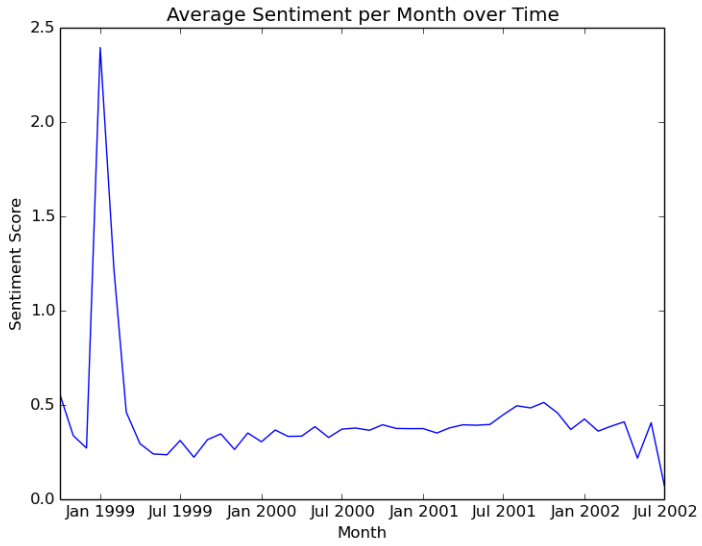



As the pipeline indicates, initially, the input text is split into sentences and each sentence is fed to a high precision subjectivity classifier. In case the sentence is classified as subjective then syntactic patterns are learned from this instance. On the other hand, if the sentence is classified as objective, then we can utilize it as a training example by feeding it to the pattern-based classifier.

The pattern-based classifier outputs the class of the sentence based on the learned patterns so far. If the instance is subjective then again more patterns are learned from it, otherwise it is fed to a high precision objectivity classifier. If the sentence is classified as objective, then it is ignored, otherwise it is fed

to the polarity classifier.

Finally, the polarity classifier estimates the numerical sentiment and normalized sentiment values and outputs the result. Optimally, we could take the instances with high confidence from the polarity classifier and feed it to the SVM classifier to train the SVM better and improve the classification performance but since this option would take up to 25 hours on a regular machine, this approach was not utilized.




Clearly, there is an abnormal spike in Jan 1999, followed by a steady increase till mid 2001. The abnormal spike can be explained Enron's exemplary performance in the year before, Fortune Magazine had just named Enron as the "Most Innovative American Company" 3 times in a row and the stock price of Enron was at an all-time high. The slight and steady increase in sentiment from Mar 1999 to Jul 2001, although counterintuitive at first does make logical sense when one examines the situation carefully. Since these sentiment scores are for sent emails, particularly by the most powerful people in the organization, it would be logical for them to send positive sounding

emails to their subordinates, especially when prominent board members such as Jeffrey Skilling, Kenneth Lay etc. were aggresively encouraging their own employees to buy Enron stock.

# 5 Conclusion & Further Research

In conclusion, although topic modelling and sentiment analysis do provide some useful insights into the data, it is still unclear as to whether they can be used as investigation tools. However, with that being said, there are some ways in which the analysis can be improved upon. For example, analysis can be conducted by focusing on users who directories are especially large, namely, 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), Richard Sanders (Assistant General Counsel) and Williams III (Senior Analyst). This would ensure that only the emails of most relevant people to the scandal are examined and this might reveal more interesting patterns.

# 6 References


1. A Rather Nosy Topic Model Analysis of the Enron Email Corpus. (2013, November 3). Retrieved May 15, 2015, from http://rforwork.info/2013/11/03/a-rather-nosy-topic-model-analysis-of-the-enron-email-corpus/

2. A Rather Nosy Topic Model Analysis of the Enron Email Corpus. (2013, November 3). Retrieved May 15, 2015, from http://rforwork.info/2013/11/03/a-

rather-nosy-topic-model-analysis-of-the-enron-email-corpus/

3. A Rather Nosy Topic Model Analysis of the Enron Email Corpus. (2013, November 3). Retrieved May 15, 2015, from http://rforwork.info/2013/11/03/a-rather-nosy-topic-model-analysis-of-the-enron-email-corpus/