Analysis of the Enron email data: NLP of Email Content

Aditya Krishn*
aditya.krishn@rwth-aachen.de
Rheinisch-Westfälische Technische Hochschule
Aachen, North Rhine-Westphalia, Germany

Mentor: Dr. Jürgen Lerner*
juergen.lerner@cssh.rwth-aachen.de
Rheinisch-Westfälische Technische Hochschule
Aachen, North Rhine-Westphalia, Germany

ABSTRACT

Archived organizational email data-sets have always been considered a valuable resource for different aspects of textual analysis such as topic modelling and sentiment analysis. But most of the experiments done are performed on synthetic data due to a lack of an real life and adequate benchmark. The Enron email data-set is boon for such research. In this report I examine the differences between the behavioural aspects of the employees through topic modelling and sentiment/emotion analysis as well as try to gauge the ethical aspect of the employees through detection of moral language.

KEYWORDS

Enron dataset, Sentiment analysis, Emotion analysis, Topic modelling, Moral language

1 INTRODUCTION

Enron Corporation was an American energy, commodities, and services company based in Houston, Texas. It was very attractive to investors and the fastest growing company making money at a rate no one has ever seen before(Figure 1). Fortune named Enron "America's Most Innovative Company" for six consecutive years. The company's share prices went up from $10/share to around $80/share from 1999–2000. Unfortunately, all the numbers were fabricated and were produced using certain manipulating accounting techniques as well as support of shell companies. On December 2, 2001 it filed for bankruptcy. During it's investigation the emails of the top 150 executives was packaged into a dataset and made public by the Federal Energy Regulatory Commission.

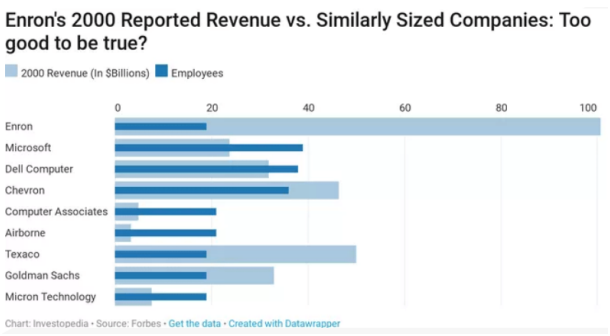

Figure 1: Enron Reported Revenue Compared

1.1 Dataset Background

The Federal Energy Regulatory Commission(FERC) subpoenaed all of Enron's email records as part of the ensuing investigation. Over the following two years, the commission released, unreleased, and rereleased the email corpus to the public after deleting emails that contained personal information like social security numbers. The Enron corpus contains emails whose subjects ranged from weekend vacation planning to political strategy talking points, and it remains the only large example of real world email datasets available for research. The dataset itself had a lot of integrity problems. It was later collected and prepared by Melinda Gervasio at SRI for the CALO (A Cognitive Assistant that Learns and Organizes) project; most of the integrity problems in the dataset had been resolved. It contains all kind of emails personal and official. Some of the emails have been deleted as part of the redaction effort due to requests from affected employees. William Cohen from CMU has put up the dataset on the web for researchers (http://www-2.cs.cmu.edu/ enron/).

1.2 Research Goal

There are mainly three goals:

- Sentiment Analysis: Inspect if there is any change in sentiment and emotion before and after the discovery of the scandal.
- Topic Modelling: Analyse what kind of topics come out of the email content of a level-A executive and low-level employee and whether the topics are similar or different.
- Moral Language Inspection: Inspect if email content before and around scandal have presence of moral language for an employee.

2 APPROACH

2.1 Dataset and Preprocessing

The original Enron email dataset, consisting of 92 percent of Enron's staff emails, i.e. 619,446 email messages in total, was posted to the web by the FERC) in May of 2002. A group of researchers at SRI International worked on these problems for their Cognitive Assistant that Learns and Organizes (CALO) project, and the resulting dataset was sent to and posted by Professor William W. Cohen at Carnegie Mellon University (CMU) [Cohen, 2004]. This dataset is called the March 2, 2004 Version, which is widely accepted by many researchers. In this version, the attachments are excluded, and some messages have been deleted upon the request of Enron employees. The resulting corpus contains 517,431 messages organized into 150 folders. The folder's name is given as the employee's last name, followed by a dash, followed by the initial letter of the employee's first name. For example, folder "allen-p" is named after Enron employee Phillip K. Allen.[6] Presumably we guess that each folder matches one employee, but this conjecture is not correct, which will be discussed in detail in the Data Cleaning Experiment section.

Each employee folder contains subfolders, such as “inbox”, “sent”, “_sent_mail”, “discussion_threads”, “all_documents”, “deleted_items”, and subfolders created by the employee. A large number of duplicate emails exist in those folders. An Enron email message contains the following header fields(figure 2) in order (the header field in parenthesis is optional): “Message-ID”, “Date”, “From”, (“To”), “Subject”, (“Cc”), “Mime-Version”, “Content-Type”, “Content- Transfer-Encoding”, (“Bcc”), “X-From”, “X-To”, “X-cc”, “X-bcc”, “X-Folder”, “X-Origin”, and “X-FileName”. The email content is separated with the headers by a blank line. The signature and quotation are continued if they exist. The header field names initiated with an “X-” means that the field values are from the original email message. The field values of “From”, “To”, “Cc” and “Bcc” are converted from those of “X-From”, “X-To”, “X-Cc” and “X- Bcc” correspondingly by the SRI researchers based on their rules.
The May 7, 2015 version of the dataset was used for the analysis which contained the employee folder name as key and the emails as values.[1]

> Message-ID: ♥0965995.1075863688265.JavaMail.evans@thyme>
> Date: Thu, 31 Aug 2000 04:17:00 -0700 (PDT)
> From: phillip.allen@enron.com
> To: greg.piper@enron.com
> Subject: Re: Hello
> Mime-Version: 1.0
> Content-Type: text/plain; charset=us-ascii
> Content-Transfer-Encoding: 7bit
> X-From: Phillip K Allen
> X-To: Greg Piper
> X-cc:
> X-bcc:
> X-Folder: \Phillip_Allen_Dec2000\Notes Folders\'sent mail
> X-Origin: Allen-P
> X-FileName: pallen.nsf
>
> Greg,
>
> How about either next Tuesday or Thursday?
>
> Phillip

Figure 2: Email Structure and components

The dictionary structure was converted to a dataframe structure with all the fields in the values added as column. Furthermore, the fields ’file’, ’Mime-Version’, ’Content-Type’, ’Content-Transfer-Encoding’ were dropped. Certain preliminary steps such as conversion of date column to datetime format were performed.

2.2 Sentiment Analysis

VADER (Valence Aware Dictionary and sEntiment Reasoner)[5] is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.
The heuristics of VADER go beyond what would normally be captured in a typical bag-of-words model. They incorporate word-order sensitive relationships between terms.
VADER belongs to a kind of sentiment analysis that depends on lexicons of sentiment-related words. In this methodology, every one of the words in the vocabulary is appraised with respect to whether it is positive, neutral or negative. Additionally, it also quantifies how much of positive or negative emotion the text has and also the intensity of emotion. A main advantage of VADER because of which it was suitable for the problem statement was that it does not require any training data. Apart from the magnitude for each polarity it also provides a compound score. The compound score is the sum of positive, negative and neutral scores which is then normalized between -1(most extreme negative) and +1 (most extreme positive).
In figure 3 we can see that with the use of conjunction in a sentence, the positive & compound score has decreased.

```
print(sentiment.polarity_scores("This is an excellent car with great mileage"))
{'neg': 0.0, 'neu': 0.435, 'pos': 0.565, 'compound': 0.8316}
print(sentiment.polarity_scores("This is an excellent car with great mileage but it's power output could have been better"))
{'neg': 0.0, 'neu': 0.598, 'pos': 0.402, 'compound': 0.8294}
```

Figure 3: Example of VADER sentiment analysis

2.2.1 Correlation Analysis. As an additional exercise to get more insights correlation analysis between the compound score from VADER as well as the closing price of Enron stock traded was done to understand whether the sentiment of the company or a particular employee was reflected in the sentiment of shareholders. Pearson coefficient was used for the similarity calculation. Pearson correlation coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance.

2.3 Emotion Analysis

Emotion is a biological state associated with the nervous system brought on by neuro-physiological changes variously associated with thoughts, feelings, behavioural responses, and a degree of pleasure or displeasure. Human being can easily identify the emotions from text and experience it. But what about the machines, are they able to identify the emotions from text?
Apart from sentiment analysis to gauge the mood of the sender of the email with respect to the receiver detecting emotion was the natural next step. A python package text2emotion[4] which helps to classify the tone of text into five basic human emotions i.e Happy, Angry, Surprise, Fear and Sad was used. At its core is a huge corpus of words and even emoticons tagged to the five emotions. It works in the following three stages:

- Text Pre-Processing: Remove the unwanted textual part from the message such as stop-words to make the content suitable for emotion analysis.
- Emotion Investigation: Find the appropriate words that express emotions or feelings. Check the emotion category of each word. Store the count of emotions relevant to the words found.

- Emotion Analysis: Convert to a structure of dictionary with keys as emotion categories and values as emotion score. Summarise to see which is the dominant emotion and tag it with that particular option.

2.4 Topic Modelling

One of the primary applications of natural language processing is to automatically extract what topics people are discussing from large volumes of text. A topic modeling tool takes a single text (or corpus) and looks for patterns in the use of words; it is an attempt to inject semantic meaning into vocabulary. Topic modeling programs do not know anything about the meaning of the words in a text. Instead, they assume that any piece of text is composed (by an author) by selecting words from possible baskets of words where each basket corresponds to a topic(figure 4). If that is true, then it becomes possible to mathematically decompose a text into the probable baskets from whence the words first came. The tool goes through this process over and over again until it settles on the most likely distribution of words into baskets, which we call topics.[3]
One of the most commonly used topic modelling technique is called Latent Dirichlet Allocation(LDA). LDA is a bag-of-words algorithm that helps us to automatically discover topics that are contained within a set of documents. In LDA, Latent indicates the hidden topics present in the data then Dirichlet is a form of distribution. As indicated by Dirichlet, the Dirichlet distribution is assumed to govern the distribution of topics and word patterns in documents. Dirichlet distribution is different from the normal distribution. When ML algorithms are to be applied the data has to be normally distributed or follow Gaussian distribution. The normal distribution represents the data in real numbers format whereas the Dirichlet distribution represents the data such that the plotted data sums up to 1. "Allocation" here refers to the process of giving something, in this case, topics.

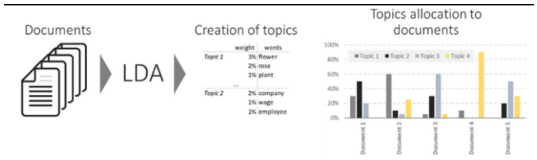

Figure 4: LDA Topic Modelling

There are basically three components to LDA:

- Dimensionality Reduction: where rather than representing a text T in its feature space as {Word_i: count(Word_i, T) for Word_i in Vocabulary}, you can represent it in a topic space as Topic_i: Weight(Topic_i, T) for Topic_i in Topics
- Unsupervised Learning: where can be compared to clustering, as in the case of clustering, the number of topics, like the number of clusters, is an output parameter. By doing topic modelling, we build clusters of words rather than clusters of texts. A text is thus a mixture of all the topics, each having a specific weight
- Tagging: abstract "topics" that occur in a collection of documents that best represents the information in them.

LDA has three important hyper-parameters(figure 5):

- alpha: which represents the document-topic density factor
- beta: which represents word density in a topic
- k: the number of components representing the number of topics you want the document to be clustered or divided into parts.

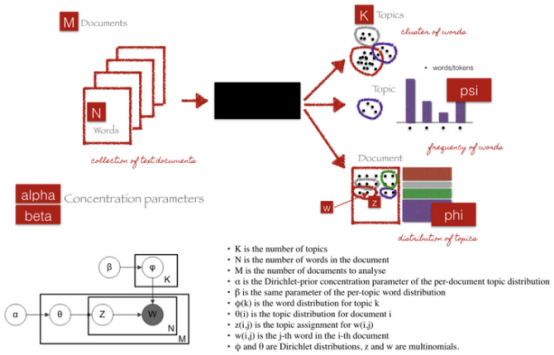

Figure 5: LDA Parameters

2.5 Morality Detection

Considering that Enron was one of the biggest scandal to have ever happened with the company paying its creditors more than $21.7 billion from 2004 to 2011, it is difficult to fathom how its leadership managed to fool regulators for so long with fake holdings and off-the-books accounting. Additionally, an interesting aspect to analyse might be that though most of the top executives especially A-class as well as some of the legal team knew about the scandal, were they morally aware or cognizant of the fact that they were doing something wrong?
Even upon searching for any related work done on measuring morality or judging the ethical aspect of people from text, no work was found. I decided upon a crude method to calculate the similarity and that was through detection of morality in email content using synsets.
Synset is a special kind of simple interface that is present in the python NLP library, NLTK to look up words in WordNet[2]. WordNet is the lexical database i.e. dictionary for the English language, specifically designed for natural language processing by Princeton University. Synset instances are the groupings of synonymous words that express the same concept. Some of the words have only one synset and some have several. Basically in the lexical database the distance between the tokens of each sentence in focus will be compared with the members of the synsets of morality. The smaller the distance the more chances of presence of morality in the communication between employees.
Since just a few synsets limited the scope of comparison I also used the hypernyms and hyponyms of the respective synsets. Hypernyms are more abstract terms while hyponyms are more specific terms. Both come into the picture as synsets are organized in a

structure similar to that of an inheritance tree. This tree can be traced all the way up to a root. Hypernyms provide a way to categorize and group words based on their similarity to each other.
In figure 6 we can see a small subgraph of the WordNet digraph which consists of different words as vertex by edges. Each vertex v is an integer that represents a synset, and each directed edge v→w represents that w is a hypernym of v. The WordNet digraph is a rooted DAG that is, it is acyclic and has one vertex as the root that is an ancestor of every other vertex. However, it is not necessarily a tree because a synset can have more than one hypernym.

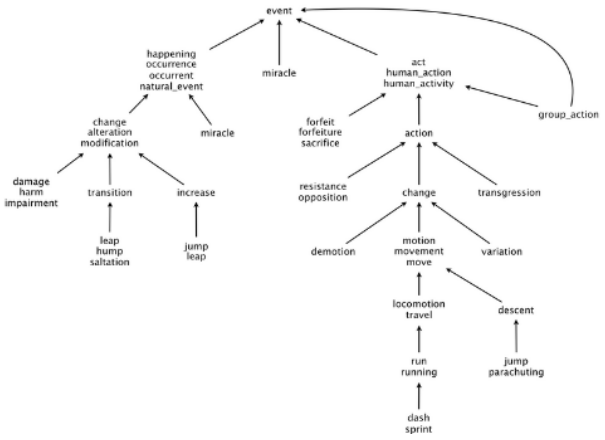

Figure 6: WordNet Digraph

Two kinds of similarity were calculated, between focus word and synset(wn.synset('morality.n.1')) and content, and then between content and focus word synset(wn.synset('morality.n.1'))

3 DISCUSSION AND RESULT

Our goal for this project is to explore the behavioural aspect of the employees with respect to the scandal as well as the ethical aspect through use of moral language in emails.

3.1 Exploratory Analysis

Since the nature of the data is of temporal nature along with communication between different persons certain elementary exploratory analysis such as trend analysis of email, distribution of the number of emails by each participant of the network, etc. was done.
Firstly, just to get an insight into when were the most emails sent the yearly email count was done. As expected most emails were sent during the period of the height of company activity centering on the revelation of the scandal. The period can be seen to be between 1999 and 2003 in figure 7.
Next the number of emails(Figure 8) belonging to a particular employee was observed. It was observed that the most emails belonged(sent and received) to V. Kaminski who was Enron's managing director for research. It should be noted that he was not found guilty and had instead warned superiors that the off-the-books partnerships and side deals engineered by Mr. Fastow were unethical and could bring down the company. The second person on

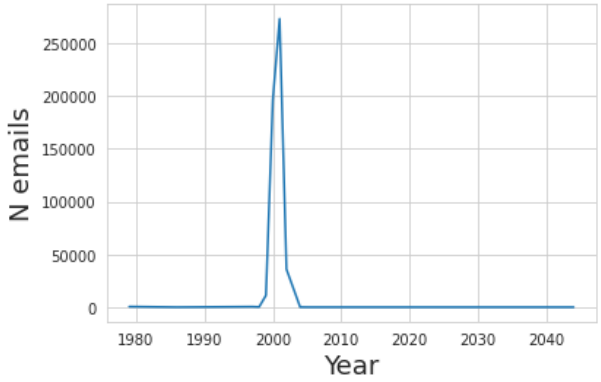

Figure 7: Number of emails sent each year

the list was J. Dasovich who was a former government relations executive for Enron. It is quite interesting that Jeff had one of the most content word count in the list. This is might be an indicator that communication with the government was mostly formal and contained more words.

	Content word count	N emails	Subject word count
user			
kaminski-v	255.906025	28465	4.217530
dasovich-j	603.393391	28234	5.237373
kean-s	490.837561	25351	4.959331
mann-k	207.195501	23381	4.211796
jones-t	185.462607	19950	4.748221

Figure 8: Email Count Analysis

It was really interesting who sends the most emails to whom, a basic social network analyses of email senders and recipients. I only look at emails sent to single email address, which may be more important personal communications. Apparently some people send a lot of emails to themselves. This maybe be very interesting to look at the differences between emails sent to selves and to others.(Figure 9)
Plotting the in and out degree of the email communication showed that the network largely resembled a scale-free degree distribution.(figure 10) The most notable characteristic in a scale-free network is the relative commonness of vertices with a degree that greatly exceeds the average. The highest-degree nodes are often called "hubs", and are thought to serve specific purposes in their networks, although this depends greatly on the domain.
Going over the word content of the subject(Figure 11) and body(Figure 12) of the emails one cannot find a lot of relevant insights. But a small highlight might be the mention of stock, sell, and price as separate words which could be investigated further.

	From	To	count
17908	pete.davis@enron.com	pete.davis@enron.com	9141
38033	vince.kaminski@enron.com	vkaminski@aol.com	4308
28920	enron.announcements@enron.com	all.worldwide@enron.com	2206
28935	enron.announcements@enron.com	all.houston@enron.com	1701
26510	kay.mann@enron.com	suzanne.adams@enron.com	1528
38031	vince.kaminski@enron.com	shirley.crenshaw@enron.com	1190
14564	steven.kean@enron.com	maureen.mcvicker@enron.com	1014
26309	kay.mann@enron.com	nmann@erac.com	980
18926	kate.symes@enron.com	evelyn.metoyer@enron.com	915
18930	kate.symes@enron.com	kerri.thompson@enron.com	859

Figure 9: Email Sender and Recipient Analysis

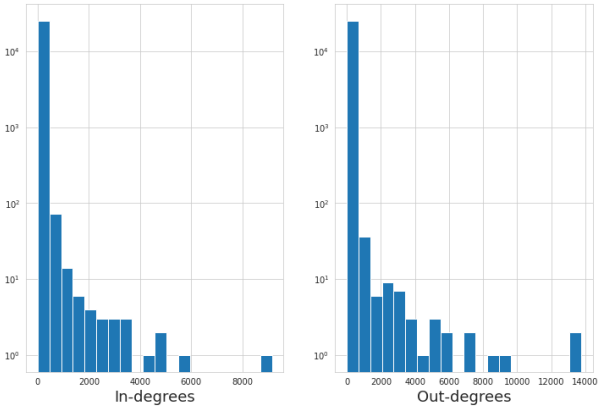

Figure 10: Graph type

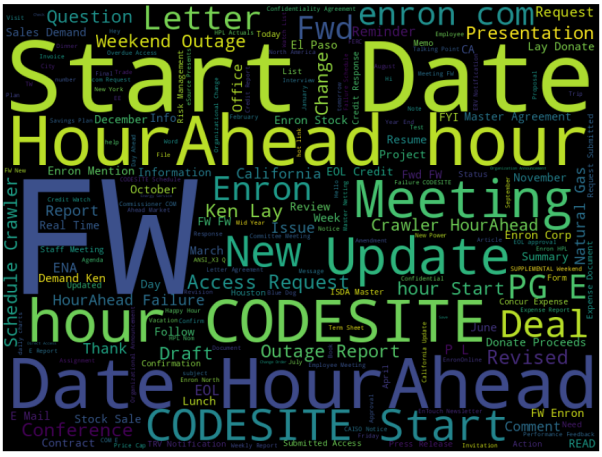

Figure 11: Email Subject Word Cloud

3.2 Sentiment Analysis

To first inspect whether VADER was the right tool to measure the sentiment considering it works without training on labelled data I took data of Kay Mann(head of legal) and selected 800 random

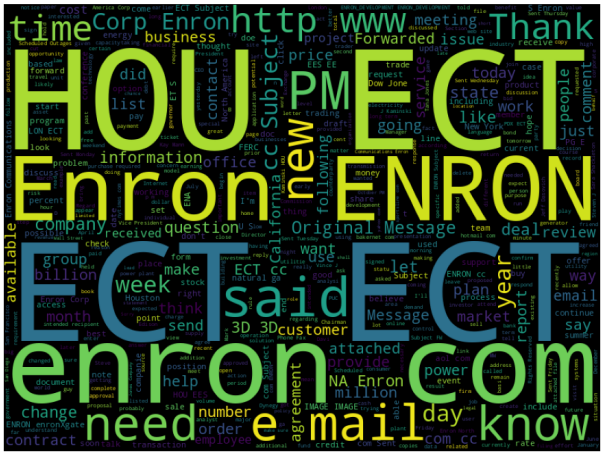

Figure 12: Email Body word cloud

rows/emails which I hand-labelled and then tested against the labels provided by VADER. Basically, the emails were either labelled as positive or negative and then checked corresponding to VADER labels. The results were significant enough to proceed further with it for the task of sentiment analysis. The evaluation metrics(figure 13) were:

(1) Precision: 0.833
(2) Recall: 0.928
(3) Accuracy: 0.884
(4) f1-score: 0.877

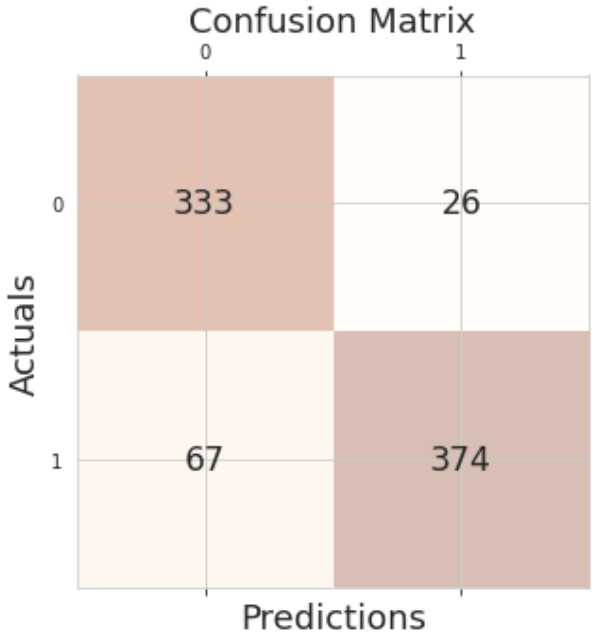

Figure 13: Confusion Matrix for VADER evaluation

As an additional exercise, the sentiment values were checked for correlation with the Enron stock price(figure 14) to see if the employee sentiment reflected the stock trend. No major correlation was seen and most of the employees with a significant number of emails were the top executives such as Pete Davis, Kay Mann and Kenneth Lay. But there was one anomaly, Suzanne Adams who was working closely with the core legal team but was found to be innocent. With

From	Correlation	Count
{{'chairman.enron@enron.com'}}	0.44	285
{{'suzanne.adams@enron.com'}}	0.40	263
{{'b..sanders@enron.com'}}	0.38	302
{{'robert.cotten@enron.com'}}	0.30	320
{{'feedback@intcx.com'}}	0.29	607
{{'legal <.taylor@enron.com>'}}	0.27	430
{{'newsletter@rigzone.com'}}	0.26	277
{{'louise.kitchen@enron.com'}}	0.26	1274
{{'pete.davis@enron.com'}}	0.26	9148
{{'public.relations@enron.com'}}	0.25	412
{{'kaminski@enron.com'}}	0.25	250
{{'kay.mann@enron.com'}}	0.25	13693
{{'kevin.hyatt@enron.com'}}	0.23	566

Email	Correlation	Count
{{'jkeffer@kslaw.com'}}	0.41	303
{{'controllers.dl-ets@enron.com'}}	0.29	293
{{'kay.chapman@enron.com'}}	0.28	340
{{'all.downtown@enron.com'}}	0.26	289
{{'judy.hernandez@enron.com'}}	0.26	344
{{'gregg.penman@enron.com'}}	0.25	386
{{'bill.williams@enron.com'}}	0.24	448
{{'suzanne.adams@enron.com'}}	0.23	1800
{{'jennifer.fraser@enron.com'}}	0.23	328
{{'bruce.mills@enron.com'}}	0.23	294
{{'dl-ga-all_enron_worldwide2@enron.com'}}	0.22	400
{{'pete.davis@enron.com'}}	0.22	9155
{{'kenneth.lay@enron.com'}}	0.21	1397

Figure 14: Correlation of Employee Sentiment with Stock Price

respect to changes observed in the sentiment score of employees before and after announcement of the SEC probe and the scandal there was no major pattern. Some employees showed a drop in their sentiment score around the announcement but the trend didn't follow the developments of the scandal. One pattern observed(Figure 15) was for Vince Kaminksi(vince.kaminski@enron.com) who as mentioned before was Enron's managing director for research and who had repeatedly tried to bring awareness internally to the inflation of revenue and systematic hiding of the losses. His sentiment score showed a similar pattern to the stock price as well as a limited pattern of his sentiment declining in terms of polarity with respect to the development of the scandal.

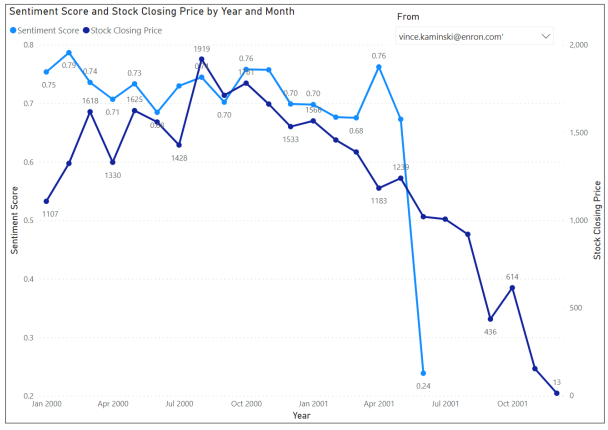

Figure 15: Sentiment score vs Enron Stock Price(Vince Kaminski)

3.3 Emotion Analysis

As discussed in section 2.3, apart from sentiment analysis an added layer for detection of the mood of the employees, emotion analysis was done. The same pre-processed dataset used for sentiment analysis was used here as well. The output was the five emotions namely: happiness, anger, sadness, surprise and fear. The emotions of two employees was generated: Kay Mann who was the head of legal and aware of and heavily involved in the scandal and Suzanne Adams who was the legal secretary. Additionally, the closing price of Enron stock was also taken for reference to check if there is any kind of correlation.

It can be clearly observed that the sadness score(figure 16) for both Kay and Suzanne peaked after the stock price of Enron started going down drastically(which happened when Enron reported their first major loss to the tune of $137 million because of which most analysts dropped their ratings for Enron's stock). Additionally, this might be extrapolation on my part but when the closing price of the stock peaked Suzanne's sadness went significantly down while Kay who knew about the misdoings didn't show that kind of indication. Though the sadness score of both going down again seems to be an anomaly but it reached a local peak when the stock price has reached its bottom.

Figure 16: Sadness Scores vs. Stock Closing Price

3.4 Topic Modelling

Topic modelling as described in section 2.4 was accomplished using the LDA technique. Eight subsets of the data were independently used to create a Bag-of-words equivalent corpus. Basic preprocessing steps such as stop word removal and lemmatization was done to minimise noise and get more coherent results. Though there are multiple libraries suitable for LDA, Gensim was used and the number of topics to be extracted was given as 10.

Additionally, the eight subsets belonged to: the emails of the official chairman email, the official Enron announcement email, email of V. Kaminski Enron's managing director for research who had tried multiple times to warn superiors that the off-the-books partnerships and side deals engineered by Mr. Fastow were unethical and could bring down the company, email of Kenneth Lay who was the founder, chief executive officer and chairman of Enron and heavily involved in the scandal, email of Suzanne(legal secretary), a combination of all emails before and after '2001-10-21' basically the date when Enron announced it's facing a SEC probe following

which shares feel down to record low.
Insights from the respective LDA models:

- Total corpus before SEC probe announcement: Most of the terms are just general words specific to corporations plus specific to Enron's main sector(energy) such as power, energy, gas, communication deal, etc. One major thing to be noted is that almost all of the topic collections are really close to each other which is generally an unlikely case. This shows that the terms were not that differentiable.
- Total corpus after SEC probe announcement: One major difference is that there is significant increase of difference of the topic collection compared to before the SEC probe. There is addition of new terms such as database and credit as well which can be said to be closer to terms related to scandal.
- Chairman email ID: This email corpus can be said to be majorly focused on scandal related terms such as creditors, bankruptcy, reorganisation, creditors and financial. Upon closer inspection it was seen that this email was majorly used after the scandal. This might be due to the fact there was a change in CEO and other executives in early parts of 2001, the year scandal came to light so a specific email with respect to the position might have been created.(figure 17)
- Enron Company Announcement: General terms such as impact, system, outages, operation and business indicating that nothing related to the scandal was discussed or announced company wide even after its revelation.
- Vince Kaminski: Terms relevant to his position are seen. Apart from that risk is term observed in all collections which might make sense considering he repeatedly tried to warn the higher management about the scandal and wrongful book-keeping.
- Kenneth Lay: A lot of scandal related terms are again observed such as violation, creditors, stock, layoff, bankruptcy, declared. Also observed was the term kslaw which is actually an international legal firm that handles bankruptcy.
- Suzanne Adams: The terms observed were ones that are expected to be used by a normal employee such as outlook, mailbox, legal, migration, meeting, etc. No terms related to scandal were observed.

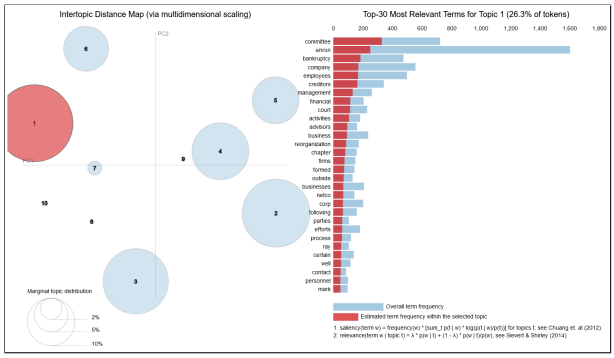

Figure 17: Topic Modelling using LDA for "chairman.enron@enron.com"

3.5 Morality Detection

AS described before in section 2.5 Morality Detection I used the synsets for words related to morality from the Wordnet lexicon adding both the hypernyms and hyponyms of the synsets to the collection. This resulted in a synset collection of 208 words. Next the tokens of the sentence in focus were tagged using POS tagger of NLTK so that the similarity of the sentence with the synset collection could be done. For each token of the synset collection the similarity value of the most similar word in the focus sentence was calculated after which the similarity score was normalised to get a single value.
The similarity calculation was done for multiple employees but the most interesting pattern was found for Suzanne Adams who was the secretary and paralegal in the legal team. We can see that there is a peak of similarity value(Figure 18) around June-July 2001, around the same time as when Enron reported their first major loss to the tune of $137 million because of which most analysts dropped their ratings for Enron's stock. But on closer inspection of the email content nothing really significant was observed. A few emails about general concern for the company's future and some problems with the financial aspect especially with respect to payments to their vendors/suppliers.

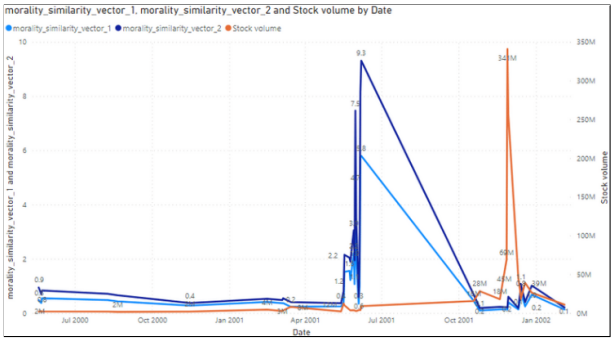

Figure 18: Morality Similarity with Volume of Stock Traded

4 CONCLUSION

A number of techniques were used to capture the behavioural aspect of the employees and their response to the scandal. The accuracy of VADER to measure sentiment was successfully appraised following which we saw that the sentiment of the employees didn't shift completely to negative after the discovery of the scandal. Certain employees who acted as whistle-blowers internally showed a limited pattern of negative sentiment on announcement of the series of losses and probe by the regulatory body into Enron's financial. With respect to emotions of the employees the emotions followed the development of the scandal in a finite manner.
The topic collection extracted from the emails show that the high-level executive knew about the scandal unlike the low and middle-level employees. Even in company announcements to all the employees mention of any scandal related term wasn't seen pointing to the fact that the whole matter was hidden from common knowledge of the employees showcasing criminal behaviour.

Since a proxy for measurement of the ethical nature of the employees was used reliable results cannot be stated. Additionally, even indication of moral language with respect to the approach used didn't reflect in the corresponding email content.

4.1 Future Work

The following things are planned to be worked upon to improve and enhance the project:

- Parallelization: Though most of the tasks are already parallelized to use the full potential of CPU, explore frameworks and libraries that help deploy the task code on GPUs.
- Targetted Sentiment Analysis: Use name-entity recognition based approaches for better result. Considering the fact that more than one employee is mentioned in a lot of emails, it would be great to see what the sender's sentiment is with respect to different entities.
- Improvement in moral language detection: Develop better ways for moral language detection using existing social science(sociology and psychology) based approaches.
- Transformer-based models: Use transformer-based models like BERT in different tasks such as topic modelling for better accuracy and results.
- Enhancement in topic modelling pipeline: Use evaluation strategies for the model such as perplexity and coherence for better results.

5 CODE

Git: https://git.rwth-aachen.de/adityakb95/masterproject_enronanalysis
The link to the dataset apart from being mentioned here is also present in README section of the repository

6 ACKNOWLEDGMENTS

A big thank you to Dr. Jürgen Lerner for their help and guidance in the project. Not only did they assisted in better formulating the approach but also in making sure that the final result is insightful and filled with takeaways relevant to the problem statement.

REFERENCES

[1] Will Cukierski. 2015. The Enron Email Dataset:500,000+ emails from 150 employees of the Enron Corporation. https://www.kaggle.com/datasets/wcukierski/enron-email-dataset/.
[2] Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.
[3] Shawn Graham, Scott Weingart, and Ian Milligan. 2012. Getting Started with Topic Modeling and MALLET. *Programming Historian* 1 (Sept. 2012). https://doi.org/10.46430/phen0017
[4] Band Gupta et al. 2020. text2emotion. https://pypi.org/project/text2emotion/.
[5] Clayton J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text.. In *ICWSM*, Eytan Adar, Paul Resnick, Munmun De Choudhury, Bernie Hogan, and Alice H. Oh (Eds.). The AAAI Press. http://dblp.uni-trier.de/db/conf/icwsm/icwsm2014.html#HuttoG14
[6] Yingjie Zhou, Mark Goldberg, Malik Magdon-Ismail, and Al Wallace. 2007. Strategies for cleaning organizational emails with an application to enron email dataset. In *5th Conf. of North American Association for Computational Social and Organizational Science*.