

Arindam Samanta

Context Aware Text Analysis

BUDSC-680

Bellevue University

­­

TABLE OF CONTENTS

[Business Problem 3](#_Toc98877911)

[Background/history 3](#_Toc98877912)

[Data explanation 4](#_Toc98877913)

[(Data Prep / Data Dictionary / etc) 4](#_Toc98877914)

[methods 5](#_Toc98877915)

[Key Findings #1 5](#_Toc98877916)

[Key Findings #2 5](#_Toc98877917)

[Key Findings #3 5](#_Toc98877918)

[Analysis 6](#_Toc98877919)

[Conclusion 7](#_Toc98877920)

[Assumptions 7](#_Toc98877921)

[Limitations 7](#_Toc98877922)

[Challenges 7](#_Toc98877923)

[Future Uses / Additional Applications 7](#_Toc98877924)

[Recommendations 7](#_Toc98877925)

[Implementation Plan 7](#_Toc98877926)

[Ethical Assessment 7](#_Toc98877927)



# Business Problem

During Enron’s rise to the top, they were intertwined with multiple counts of fraudulent activity that could have been detected years before Enron’s fall if investigators had the right tools. I am trying to solve the problem by proposing a text analytics solution to it. In order to understand the motives of the fraudsters we have to analyze unstructured data, such as emails and memos. Finding the relevant information quickly from the huge pile of textual data is a daunting task for the auditors.

One way to tackle the problem is to use Text Analytics using NLP and pattern matching to facilitate timely fact extraction and data organization.

# Background/history

Enron Corporation was an American energy, commodities and services company based out of Houston, Texas. In 2001, they filed for bankruptcy. Before their Dec. 2, 2001 bankruptcy filing, Enron employed 20,000 staff. They were one of the world’s leading electricity, natural gas, communications and pulp and paper companies, with claimed revenues of nearly $101 billion in 2000. Later it was revealed that its reported financial condition was sustained substantially by an institutionalized, systematic, and creatively planned accounting fraud, known since as Enron scandal.

# Data explanation

## (Data Prep / Data Dictionary / etc)

The 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 This data was originally made public and posted to the web, by the Federal Energy Regulatory Commission during its investigation.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.

Although the dataset is huge but folders of particular users are often quite sparse. I decided to look into the sent emails folder. Through this approach I could avoid analyzing spam email or junk email folders.

Data Source: https://www.cs.cmu.edu/~enron/

# MetHods

Two main methods of analyses that I employed were namely,

1. Topic modelling with Latent Dirichlet Allocation (LDA)
2. Sentiment Analysis

## Key Findings #1



## Key Findings #2



## Key Findings #3



# Analysis

Insert any charts/graphs/ data

# Overhead view of hands shaking over a business papersConclusion

Time to wrap it up. What is your conclusion? How would you synthesize all the information into something even the busiest CEO wants to read? What are the key takeaways? How does your product/service/methodology uniquely address the issues raised by your study?

## Assumptions

* Takeaway #1
* Takeaway #2
* Takeaway #3

## Limitations

## Challenges

## Future Uses / Additional Applications

## Recommendations

## Implementation Plan

## Ethical Assessment